1. *Introduction*

1.1 Integrated Circuit Trends

The increased complexity available Irom integrated circuits with passing time has a significant influence on the design of digital computer Systems, Integrated circuits (known as chips), where an entire circuit is manufactured in a single piece of silicon, first appeared around1960. At that time, the chip size and transistor dimensions were such- that only a few simple gates offering primitive logic functions such as not, nand, nor etc. could be accommodated; this level of integration is called small scale integration (SSI).

Improvements in the processing techniques in subsequent years have resulted in a steadily increasing chip area and a -progressively reducing future size. This has allowed a complexity increase of approximately one hundred every ten years. Thus by 1970, medium scale integrated (MSI) circuits with about a thousand transistors appeared, while by 1980 large scale integrated (LSI) circuits of approximately one hundred thousand devices were possible. At this irate, chips capable of containing around ten million components should be available to the designer by 1990. There is every confidence that this projection of several million transistors per chip is realistic. This level of integration is very large scale integration (VLSI).

The capability of integrated circuits has reached the point where an entire system can be integrated, rather than just some small portion of it. As the *chip* content becomes more complex, the problem of producing a correct design at the first attempt within an acceptable time scale becomes increasingly difficult. This is highlighted by Barron's corollary to Moore's second low which suggests that the design of a million transistors will take thirty men ten years! With such integrated systems, the complexity and management of -the design is dominating all other problem areas.

1.2 Choice of Technology

Although other materials are available for manufacturing integrated circuits, such as silicon-on-saprphire, gallium arsenide etc., there is a considerable cost penalty involved in their use. Thus silicon remains and is likely to stay) the most economically effective way of implementing VLSI.

Two distinct types technology are fabricated- in silicon based upon the bipolar junction transistor and the metal-oxide-semiconductor (MOS) transistor. Since the processing for these technologies is very different. it is not practical to mix them within a chip or in a wafer of chips.

MOS logic occupies a much smaller area of silicon than the euavalant bipolar logic. This is partly due to a smaller device size and partly due to the fact that MOS structures require fewer components. Thus MOS technologies has a much higher potential packing density.

An MOS logic circuit requires appreciably less current and hence less-power than its bipolar counterpart. However, bipolar circuit operate faster than MOS circuits. Even so, the speed-power product for MOS logic compare favorably with that for bipolar logic; this product is used as a figure of merit to compare logic families, since greater speed can. usually be obtained at the expense of increased current consumption and therefore increased power dissipation.

Thus, for example. The popular commercially available dipolar logic family, low-power Schottky transistor- transistor logic (LS-TTL), has a typical seed (propagation delay) of 10 ns and a quiescent (static) power dissipation of 2 mW for a two-input nand gate, giving a speed-power product of 20 pJ. This compares with the some function in the commercially available MOS logic family, complementary metal-oxide-semiconductor (COMS), which has a typical propagation delay of 40 as and a static power dissipation of 10 nW. giving 0.0004pJ as the speed-power product.

The structure of an MOS transistor Is much simpler than that for bipolar devices and this makes its manufacturing process easier. This in turn should result in fewer faults occurring in fabrication and hence increase the number of working chips compared with the number obtain in a similar area from bipolar technology. The greater yield of good chips offered by MOS technology is of importance, since a higher proportion of Chips do not function correctly owing to manufacturing defects.

MOS technology also offer the advantage of being Able to implement dynamic logic, where stases are stored temporarily on capacitance inherent in the circuit structure; this leads to further reductions in area and power, and such circuits are obviously important in the cotext of VLSI, It is not possible to implement dynamic logic in bipolar technology and thus MOS offers a greater choice of design implementations.

Thus in terms of area, power dissipated, yield and flexibility, MOS technology is superior to bipolar technology. Furthermore, of the two technologies only MOS is capable of realising VLSI. It is therefore the technology chosen for use in this text.

Although at present bipolar transistors are faster that MOS devices, this situation is likely to charge in the future. The speed of an MOS transistor is depended upon its size. This speed in increasing with time, owing to fabrication advances which continue to reduce the surface future size. The speed of a bipolar device is dependent upon a vertical dimension (defining the base width)

which is near its (quantam) limit for the fastest transistors. As a result, it is likely that the speed improvement of biploar technology in the future will be small compared with that achievable in MOS. so by 1990, MOS speeds are forecast to match those of the fastest biploar technology, and gate delays for simple function should be better that 500ps.

Within MOS technology, them are two main logic families available to the designer and both are described in this book. The NMOS logic family is based upon n-channel MOS transistors while COMS requires both n-channel and p-channel MOS transistors. The NMOS process is simpler than that for CMOS since only one transistor type is involved. In addition, NMOS logic structures require fewer devices and occupy (about 60 pet cent) less area than the equivalent COMS circuit. Despite this, CMOS is likely to be eventually the design medium for VLSI as it requires much less power than NMOS and its circuit speed is superior so that of NMOS. The additional area required by CMOS is not thought to be a limitation of VLSI realisation. This is because it is likely that most of the chip area will be required for interconnections and therefore circuits will occupy a relatively small area.

1.3 Design Approaches

There are three approaches to implementing digital design. The first is to design with chips which are available 'off the shelf' from manufacturers. Although there is a wide range of SSI, MSI and LSI devices available in silicon and other technologies, the designer is limited to the integrated circuits on offer. Often, there is a trade-off between using a few MSI and LSI chips which do not Include some required features and using many SSI and MSI devices which exactly perform the task required. In both cases, the availability of specific functions influences the resulting design.

The most efficient Implementation of a design from the viewpoint of functionally, space and power is to Integrate it. Here, the designer has total control over the chip function including the specification of the content of each layer manufactured in silicon. Now, the only limitation of the chip content, in principle, is the designer's imagination! this second design approach is refereed to as full custom design and. for LSI/VLSI designs, MOS technology is used.

Full custom design has associated overheads of the design time plus development and manufacturing costs; these are considerable. These overheads are largely avoided by the third approach which is semi-custom design in the form of the uncommitted logic array (ULA). Here the silicon is preformed as a set of uncommitted logic cells, each of which can be configured to perform a variety of simple logic functions. The designer thus only needs to specify the cell inter- connections and the cell configuration to provide a user-specified function. Thus the only significant time and costs involved are- those associated with the inter- connection layer(s).

Although this approach clearly dose not yield the same efficiency or flexibility as full custom, it is useful as relatively fast and cheap method of costuming. ULAs are often used in Preference to incorporating a group of standard SSI and MSI chips in designs; this is particularly cost-effective if many such parts are required. Although bipolar ULAs are available, most arrays use CMOS Technology.

It is full custom design which Offers the potential to realise a VLSI circuit that is totally defined by the designer. It is therefore the different design phases involved in committing a full custom MOS circuit on to silicon which forms the subject matter of this book. It should be noted that since MOS semi-custom circuits, with their fixed logic types and placement, can be regarded as a sub-set of full custom design. The text is also applicable to this form of design.

1.4 The Design Process

The design methodology adopted for all digital design is that of a top-down, hierarchical approach. Here the design is divided into a number of distinct levels where each level is derived from the information in the level above it. In this way, a design progresses from the initial system specification to its actual implementation. Thus, at each level down in the hierarchy the design becomes a progressively more detailed description of how the system specification is to be implemented.

The different levels in the hierarchy for a full custom design me shown figure 1.1 System design is performed at the highest level. This takes the system specification and translates it into a block diagram of the architecture. The diagram shows the system's functional blocks, such as cache memory, register arithmetic blocks, logic blocks etc., and their interconnecting data paths.

In a complex system, the architecture cannot be directly derived from the system specification and the system design stage itself requires a top-down design approach. This begins by defining the units and communicating data paths which comprise the system. Each unit is then taken in turn partition into communicating sub-units, which in their turn are expanded. In this way. large system is progressively partitioned and defined until the design is sufficiently detailed for the architecture to be drawn.

An important feature of a design is the ability to be able to test it, in order to verify that it operates correctly providing such a test capability is an integral part of any chip design. Its inclusion may well involve additional circuitry and influence the logical design. For these reasons, the test strategy should be determined at the highest level in the design hierarchy. Testing is put of the style, design and should be appended to the design at some lower in the-hierarchy as an afterthought!

At the system design level, the architecture is checked aping the system specification to ensure that all required hardware features and data paths have

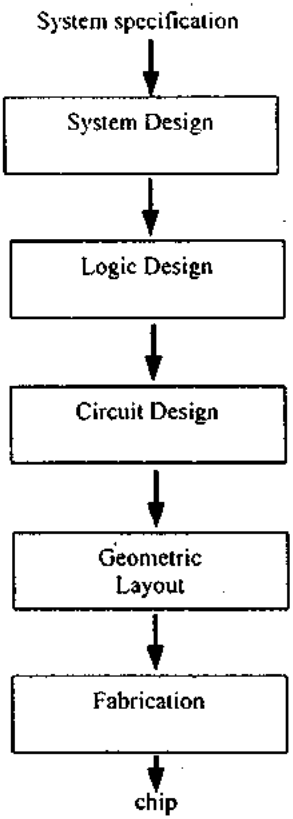


Figure 1.1 Top-down design hierarchy for full custom design

been included. It is also usual to check that the required number of input and output pins can be accommodated around the chip periphery. However, other feasibility checks of this nature are difficult since the silicon area occupied by architectural blocks, is not known at this stage. Nevertheless, it is usual to perform a tough block placement. This is checked for feasibility later when the areas can be estimated from design work at the circuit level.

At the next level down in the hierarchy, the architectural blocks we expanded into logic diagrams. Here each item drawn represents a particular logic function, such as a gate. Control and timing logic is also included at this level. Logic simulation can be used to verify that the logic functions correctly and performs the tasks required by the system specification. These test routines can also be used at a later date for performing functional tests on the fabricated chip.

The next level down is the circuit design. Here the logic is translated into circuits with dimensions assigned to the transistors. The circuit diagrams are

often drawn as stick diagram rather than with conventional transistor symbols. A stick diagram is a pictorial representation of the circuit in terms of the lines and connections required at each layer of the -silicon. It is thus more detailed than a transistor circuit and it assists the translation process to the next level down in the hierarchy.

At the circuit design level, the logic is implemented with simple, regular transistor sutures wherever possible. Circuit simulation can verify the design at this level and provide an indication of the power dissipation and speed. By combining the stick diagram with the-fabrication rules for the geometric layout, the size of each circuit can be estimated. This allows the designer to check that the design can be accommodated on the chip and that there is sufficient area left for circuit interconnections. Unfortunately, it is not until this level that changes to the system and/or logic may become necessary to meet a speed or sin (or power) criterion. Clearly, the design process has to be repeated from the highest level that is modified.

Once the circuit design is correct, translation to the next level down in the hierarchy can proceed. The circuits we allocated to positions on the silicon and geometric shapes are generated for each silicon layer corresponding to the circuits end their interconnections. If stick diagram have been used then a geometric layout can be derived from than by just 'fleshing out' their lines. The layout, is also checked against the circuit design to confirm that circuit details have been correctly translated to the lower level. The layout k also checked for violations of the fabrication process layout rules.

After the layout stage. the design normally passes out of the designer's hands. The data representation of the geometric layout is normally used to produce a (Photographic) mask of each silicon layer. The masks are then used at the different production stage of the fabrication process to produce the specified chip.

1.5 Organization and Notation

This book treats the design of a silicon chip in term of the hierarchical structure of figure 1.1 and aim to give the reader an understanding of the principles involved at each level of the design hierarchy. However, in -order to approach the design of the upper levels of the hierarchy, it is necessary to appreciate the limitations and design considerations of the lower levels. Thus the organization of this text is from the bottom upwards rather than the top-down approach described k the previous section.

Consequently, this book commences at the circuit design Level with a description in chapter 2 of the MOS transistor's operation and its use in simple logic circuit. The next level down in the hierarchy is considered in chapter 3 where the design rules for do geometric layout of circuits are preceded by a description of the fabrication process.

This enables the higher levels of the hierarchy to be discussed. Chapter 4 describes the implementation of logic and storage elements using. Where possible, simple repetitive circuits. Chapter 5 discusses the system considerations that are associated with the highest level in the design hierarchy.

Finally, a design example is described in chapter 6 using a top-down design approach. It is hoped that this will give the reader a greater insight into the design principles involved at each Level in the hierarchy and of the interaction between levels.

Throughout the text, the discussion of logic circuits assumes. Positive logic that is, the voltage level for a logic '1' is greater than the voltage level for a logic 'O'. Thus a high input/output is a ' 1', while a low input/output is a 'O'. Similarly, an active or applied input is a '1'. While an inactive input or -an input which is removed is a '0'.

1.6 Further reading

D. J. Kinniment, 'Component technology - the next 10 years', State of the Art Report - Supercomputer Systems Technology, Series 10. No. 6. pp. 317-33, Pergamon, 1982.

W. W. Lattin, J. A. Bayliss, D. L. Budde, J. R. Rattner and W. S. Richardson, 'A methodology for VLSI chip design'. *Lambda* (now *VLSI Design),* No. 2 (1981)pp.34-44.

2 MOS Devices and Basic Circuits

The purpose of this chapter is to describe the operation of the different types of MOS devices available and then to develop the characteristic equations which describe their behaviour. This enables the simple circuits which form the basis of MOS digital circuit design to be presented.

2.1 The MOS Structure

An insight into the behaviour of an MOS device can be gained by considering its structure, and figure 2.1 shows the structure of an n-channel or NMOS transistor.

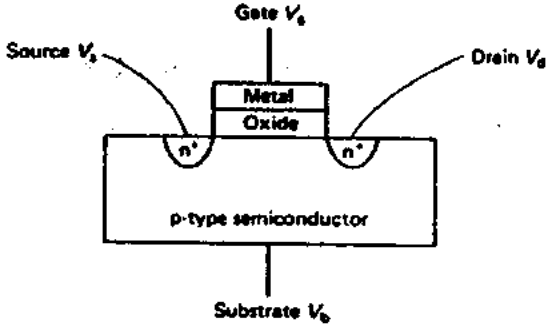


Figure 2.1 NMOS transistor structure

The basis of the transistor is a metal–oxide–semiconductor structure, hence the name MOS. The device input is called the 'gate'. Originally this was a metal plate, although nowadays it is usually made of polycrystalline silicon (commonly called poly or polysilicon). The oxide is very pure silicon dioxide and it separates the gate from the semiconductor material; it acts as an insulator.

The semiconductor is pure silicon which has been doped with relatively small amounts of an impurity such that the impurity atoms can easily replace silicon atoms in the regular, fixed crystal structure. Pure silicon is not a good conductor at room temperature as there are very few electrons of negative charge which

acquire a high enough energy to break away from the silicon atoms. The electrons which escape are free to move about the material and are referred to as 'free charge carriers'. They leave a vacancy or hole in the parent atom which now has a net positive charge. Other electrons of lower energy move to fill these vacancies and this movement of free carriers is equivalent to the movement of holes of positive charge.

An n-type semiconductor is obtained by doping pure silicon with an impurity possessing one more electron (in the outermost orbit) than silicon. This electron is only loosely bound to the impurity atom and can easily acquire enough energy to break away, leaving the fixed impurity atom positively charged. The freed electrons form the majority of free charge in the material, significantly reducing its resistivity below that of pure silicon. A few free holes still exist and these are referred to as the 'minority charge carrier'. This is depicted in figure 2.2a.

A p-type semiconductor arises when the impurity atoms have one fewer electron in the outer orbit than does silicon. There is thus a vacancy for an extra electron and electrons move to fill these vacancies, causing the fixed impurity atoms to become negatively charged. This electron extraction is equivalent to the injection of positive holes. Thus the majority of free charge carriers are holes and the minority charge carriers are electrons (see figure 2.2b).

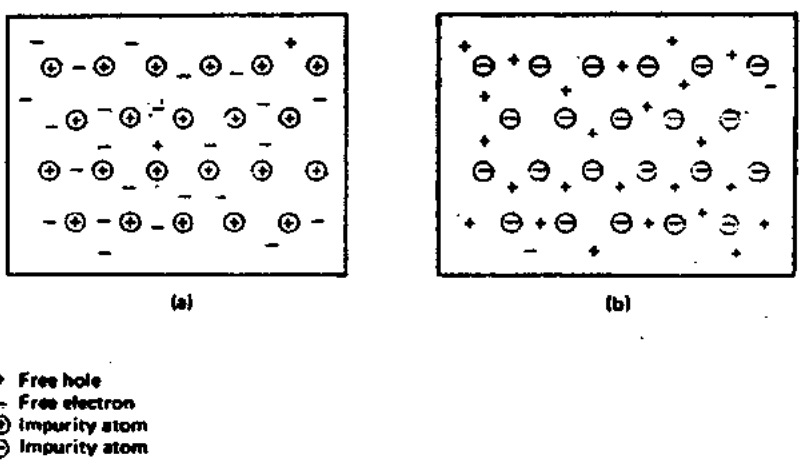


Figure 2.2 Charge within doped semiconductor: (a) n-type, (b) p-type

The semiconductor material of figure 2.1 consists of a lightly doped p-type substrate and heavily doped n-type regions, denoted n^+ and called the source and the drain, which are located at each end of the gate. Conventionally, the drain is the device output terminal and the source is the terminal that is common

to both the input and output circuits. It is therefore usual to specify the device's input voltage as V_{gs}, meaning $V_g - V_s$, and its output voltage as V_{ds}, meaning $V_d - V_s$.

A terminal is connected to the bulk substrate and in NMOS this is always connected to the most negative voltage available. This is so that the diodes formed by the substrate–source and substrate–drain pn junctions are always reverse-biased and hence never conduct.

2.2 Conduction

To make the transistor conduct, appropriate voltages have to be applied to the terminals. Figure 2.3 shows the effect of applying a positive bias to the gate and drain with the source and bulk substrate tied to 0 V.

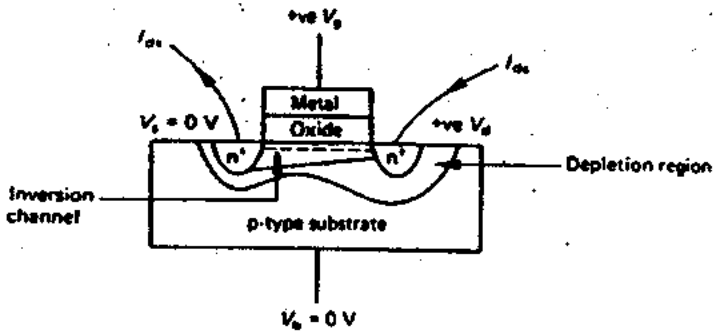


Figure 2.3 Conduction in an NMOS transistor

Although the conductivity of the semiconductor is less than that of the metal, it can nevertheless be considered to be a conducting material. Thus in figure 2.3, the oxide acts as an insulator between two conductors, so the structure resembles that of a capacitor.

The application of a positive gate bias with respect to the source causes a positive charge to accumulate on the metal and an equal negative charge, supplied by the source and the drain, to be induced in the semiconductor surface just beneath the oxide. This charge is in addition to the existing charge. If the induced charge is small then it only causes the surface layer to become less p-type than the bulk. If, however, the charge induced is large enough then the surface layer inverts from p-type to n-type, as shown in figure 2.3. There is now a continuous electron channel from the drain to the source, and current flows if there is a bias between them. In figure 2.3, current flows from the drain to the source as the drain potential is higher than that of the source.

A region depleted of free charge carriers separates all p-type regions from n-type and prevents conduction of reverse-biased pn junctions. Hence the device conduction path is isolated from the bulk substrate by a depletion region and this effect is also used to provide isolation between devices. Note that since the depletion width is dependent upon the junction reverse voltage, the depletion region is wider around the drain than around the source.

The device conduction path is also isolated from the gate by the oxide. Thus it follows that all the current flowing into the drain flows out of the source; this current is referred to as I_{ds}.

2.3 Threshold Voltage

The input potential V_{gs} at which the surface just becomes inverted is called the threshold voltage V_t. Below the threshold voltage an NMOS transistor is off and no current flows, while above the threshold the inversion channel is established and the device conducts.

A detailed analysis of the magnitude of the threshold voltage is complex and beyond the scope of this book. However, the factors determining V_t can be appreciated from a consideration of the excess charge within an NMOS device when the inversion layer is established. This is depicted in figure 2.4.

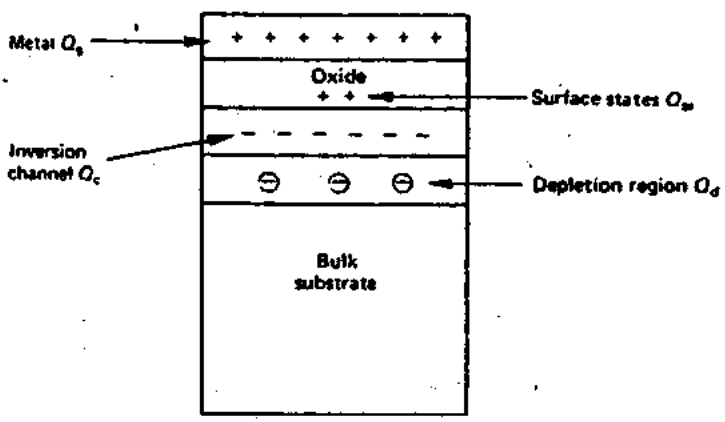


Figure 2.4 Excess charge within conducting NMOS transistor

The surface state charge Q_{ss} represents positive charge trapped at the oxide-semiconductor interface as a result of imperfections at this surface. This, plus the gate charge Q_g, must equal the induced charge in the inversion channel Q_c, plus the impurity atom charge Q_d in the depletion region. Thus

$$Q_m + Q_g = Q_c + Q_d$$

When the surface is just at the point of inversion, $Q_c = 0$ and

$$V_t = \frac{Q_g}{C_g} = \frac{Q_d - Q_m}{C_g}$$

where C_g is the capacitance across the insulator.

In practice, V_t has to overcome some in-built potential differences before the transistor is brought to the edge of conduction. As a result, another two voltage terms have to be included in the above expression for V_t. V_{dif} represents the voltage arising as a result of the difference between the gate and the semiconductor material. Since silicon gates are used nowadays, V_{dif} is small. The other voltage term, V_r, is the voltage across the depletion region just at the point of inversion; it is dependent upon the impurity concentration and is usually less than 1 volt. Thus

$$V_t = \frac{Q_d}{C_g} + V_r - \frac{Q_m}{C_g} + V_{dif} \qquad (2.1)$$

The last three terms can be regarded as constant potentials, while Q_d is dependent upon transistor parameters and the applied voltages. In particular, Q_d is dependent upon the impurity concentration of the semiconductor material beneath the oxide. This provides a mechanism for adjusting the threshold voltage.

For an NMOS device, the first two terms of equation (2.1) are positive and the last two negative, allowing the threshold to be made either positive or negative by suitable doping. The threshold is positive (usually 1 V) if the semiconductor surface between the source and the drain is heavily doped with a p-type impurity; such a device is off when the gate-source voltage is 0 V and is referred to as an 'enhancement mode NMOS transistor'.

The threshold can be made negative (typically -4 V) by doping the semiconductor surface with an n-type impurity. This has the effect of making the surface n-type, although the bulk semiconductor remains p-type. Thus even with $V_{gs} = 0$ V, the channel region is inverted and the device is on; an NMOS device which is on when V_{gs} is 0 V is called an 'NMOS depletion mode transistor'. Here, it is necessary to apply a negative gate-source voltage in order to repel electrons from the surface and turn the device off.

As well as NMOS devices, there are p-channel or PMOS transistors where the substrate is n-type and the drain and the source are heavily doped p-type regions. The substrate is connected to the most positive voltage available so that the drain–substrate and source–substrate pn junctions are always reverse-biased. A negative gate–source voltage causes holes to be attracted to and electrons to be

repelled from the semiconductor surface just beneath the oxide. This surface inverts to p-type if a sufficiently negative V_{gs} is applied.

Consideration of the excess charge in a PMOS device results in an expression similar to equation (2.1), except that all four terms are negative. Again, the threshold voltage can be adjusted by altering the impurity doping level in the surface beneath the oxide. However, logic circuits require only PMOS enhancement devices. These are fabricated with a negative threshold (normally -1 V) and are off when $V_{gs} = 0$ V.

Thresholds are quoted for transistors assuming a source–substrate voltage of 0 V. Changing the substrate voltage causes the threshold to change, and this effect is known as the 'body effect'. In NMOS devices, taking the substrate voltage negative of the source causes the depletion region surrounding the conduction path to widen and thus increases Q_d. As a result, V_t has to be increased to bring the transistor to the edge of conduction. If V_{t0} is defined as the threshold when the source–substrate voltage V_{sb} is 0 V, then the modified threshold V_t to take into account the body effect is

$$V_t = V_{t0} + \gamma(V_{sb})^{1/2}$$

γ is a constant dependent upon transistor parameters and tends to lie between 0.3 and 0.7 for MOS transistors. It is usual to use a value of 0.5 in calculations.

Similar reasoning for a PMOS transistor shows that increasing the substrate voltage above the source potential causes the threshold to become more negative.

2.4 I_{ds} versus V_{ds} Characteristic for NMOS Devices

Figure 2.5 shows the symbols used in this book to denote MOS transistors. It is assumed that the substrate of NMOS devices is tied to the most negative voltage available and a PMOS substrate to the most positive available voltage. Thus this connection is usually omitted from circuit diagrams.

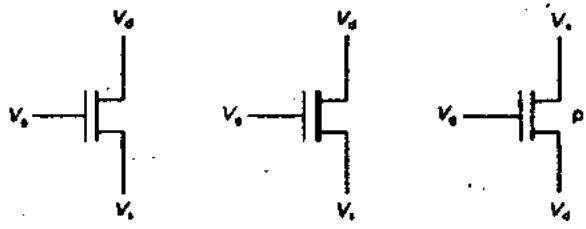


Figure 2.5 Symbols for MOS transistors: (a) enhancement mode NMOS, (b) depletion mode NMOS, (c) enhancement mode PMOS

Consider an NMOS device. When $V_{gs} < V_t$ the transistor is off, regardless of the drain voltage; the device does not conduct and no current flows. The device conducts when $V_{gs} > V_t$, and if a constant V_{gs} is applied then the resulting I_{ds} versus V_{ds} curve can be split into two regions.

(a) Resistive Region

Here, $V_{ds} < V_{gs} - V_t$. The voltage across the insulator at the source is V_{gs} and at the drain is V_{gd} (meaning $V_g - V_d$). Although the voltage across the insulator is not constant, a voltage in excess of V_t exists at all points across the oxide, causing the formation of a continuous inversion channel between the drain and the source. It will be assumed that the increase in voltage along the channel from the drain to the source is linear with distance.

The device structure therefore resembles an infinite number of capacitances between the drain and source, each one having a different voltage across it and therefore a different charge from its neighbours. The total charge induced in the channel is the sum of the charge induced on each of these capacitances.

Consider one of these capacitances of length dx situated at a distance x metres from the drain, as shown in figure 2.6.

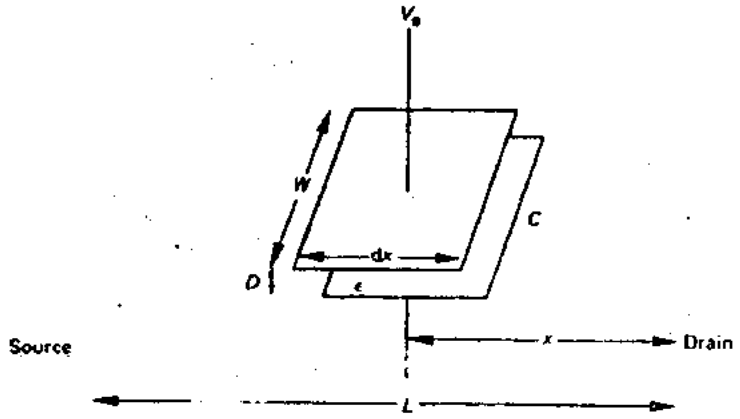


Figure 2.6 Elemental capacitor between source and drain

The channel width and length are W and L metres respectively. Thus the capacitance, C farads, of the structure shown in figure 2.6 is

$$C = \frac{We\,dx}{D}$$

where ϵ is the permittivity of the insulator in farads/metre and D is the thickness of the oxide in metres. The voltage v in excess of V_t across this capacitor is

$$v = (V_{gd} + \frac{x}{L}V_{ds} - V_t) = (V_{gs} - V_{ds} + \frac{x}{L}V_{ds} - V_t)$$

Thus the charge, q coulombs, induced on this capacitor is

$$q = C_v = \frac{We\,dx}{D}(V_{gs} - V_{ds} + \frac{x}{L}V_{ds} - V_t)$$

The total charge Q induced in the channel is

$$Q = \int_0^L \frac{\epsilon W}{D}(V_{gs} - V_{ds} + \frac{x}{L}V_{ds} - V_t)\,dx$$

$$= \frac{\epsilon WL}{D}\left[(V_{gs} - V_t) - \frac{V_{ds}}{2}\right]$$

Now $Q = tI_{ds}$, where t is the time in seconds for an electron to move across the channel and

$$t = \frac{\text{channel length } L}{\text{electron velocity}}$$

μ_n is the electron velocity per unit electric field (measured in metre^2/volt-second) and is called the 'electron mobility'. Thus the electron velocity in the channel is $\mu_n V_{ds}/L$. Hence the current I_{ds} in amps is

$$I_{ds} = \frac{Q}{t} = \frac{\epsilon W \mu_n}{LD}\left[(V_{gs} - V_t)V_{ds} - \frac{V_{ds}^2}{2}\right] \tag{2.2}$$

For a constant V_{gs}, I_{ds} increases with an increase in V_{ds}, as shown in figure 2.7. It should be noted that at drain-source voltages which are very small compared with $V_{gs} - V_t$, the equation for the channel current reduces to

$$I_{ds} = \frac{\epsilon W \mu_n}{LD}(V_{gs} - V_t)V_{ds}$$

giving a linear relationship between I_{ds} and V_{ds} for a constant V_{gs}.

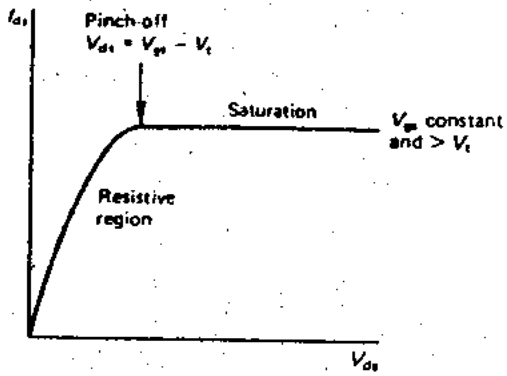


Figure 2.7 I_{ds} versus V_{ds} for an NMOS transistor

(b) Saturation

In this region $V_{ds} \geq V_{gs} - V_t$. As the drain voltage rises, the voltage across the insulator at the drain drops, and at $V_{ds} = V_{gs} - V_t$ it is V_t. This is the voltage necessary to just support inversion, and this point on the I_{ds} versus V_{ds} characteristic is called 'pinch-off'. At this point, the inversion channel ends just at the drain. As V_{ds} increases beyond pinch-off, the point at which inversion ceases moves away from the drain as shown in figure 2.8.

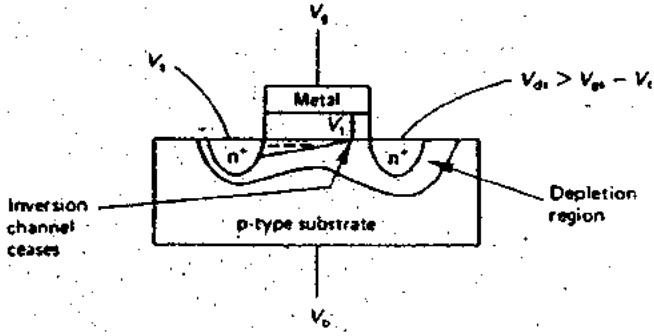


Figure 2.8 Inversion channel beyond pinch-off

The voltage difference along the inversion channel from the source to where it ceases is $V_{gs} - V_t$, and the excess potential $V_{ds} - V_{gs} + V_t$ is dropped between the end of the inversion channel and the drain. This creates a high electric field across this very short distance and the electrons from the inversion channel are quickly swept across this area to the drain.

At and above pinch-off the voltage between the source and the end of the inversion channel is constant at $V_{gs} - V_t$. The channel length can also be considered to be constant (equal to L). Thus the current flow is constant and the device is said to be 'saturated'. Replacing V_{ds} by $V_{gs} - V_t$ in equation (2.2), I_{ds} becomes

$$I_{ds} = \frac{\epsilon W \mu_n}{2LD}(V_{gs} - V_t)^2 \tag{2.3}$$

The I_{ds} versus V_{ds} characteristic for a constant V_{gs} is shown in figure 2.7. In practice, there is a slight increase in current with increasing V_{ds} above pinch-off, as a result of the reduction in the channel length.

2.5 Characteristic Equation for PMOS Devices

A PMOS enhancement device is off when V_{gs} is 0 V. Here it is necessary to take the gate voltage negative of the source in order to exceed the negative threshold and turn the transistor on. It is thus normal to connect the source terminal to a positive potential and operate the gate and the drain at voltages equal to or negative of this potential. Consequently current, I_{sd}, flows from the source to the drain in a PMOS device.

A similar analysis to that performed for NMOS devices allows the equations for I_{sd} versus V_{sd} to be obtained for the resistive and saturated regions. If the threshold voltage V_t for PMOS devices (only) is redefined to be the positive source-gate voltage at which the transistor just turns on, then equations similar to those for NMOS are obtained. In the resistive region where $V_{sd} < V_{sg} - V_t$

$$I_{sd} = \frac{\epsilon \mu_p W}{DL}\left[(V_{sg} - V_t)V_{sd} - \frac{V_{sd}^2}{2}\right] \tag{2.4}$$

Pinch-off occurs when $V_{sd} = V_{sg} - V_t$ and for $V_{sd} > V_{sg} - V_t$, the device is saturated. At and above pinch-off

$$I_{sd} = \frac{\epsilon \mu_p W}{2DL} (V_{gs} - V_t)^2 \qquad (2.5)$$

μ_p is the hole velocity per unit electric field and is known as the 'hole mobility'. It is two to three times less than μ_n. Since the current flow is proportional to the carrier mobility, an NMOS transistor will conduct more current than a PMOS device of similar size. This can be seen in figure 2.9 which shows characteristic curves for PMOS and NMOS transistors if typical values are assumed for the parameters.

The speed of a circuit is dependent upon the rate at which circuit capacitances can be charged and discharged. This in turn is dependent upon the current available from devices within the design. The current capability of PMOS devices can be made equal to that for NMOS devices by increasing their size to compensate for the difference between hole and electron mobility. However, these larger devices require a greater silicon area and have an increased circuit capacitance. For this reason, NMOS devices are used in preference to PMOS in circuit design.

2.6 Principles of Inverters

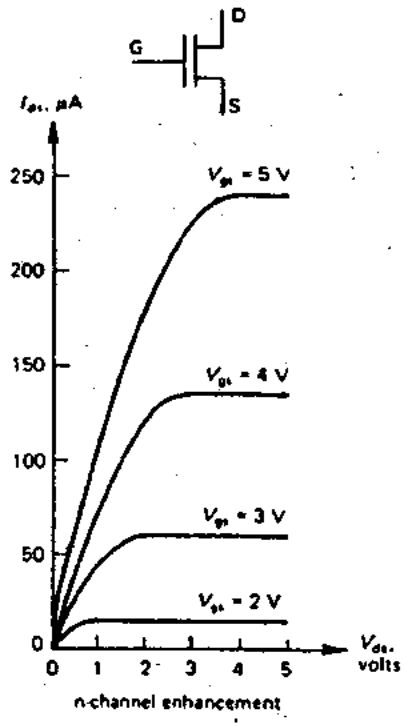
Now that the characteristics of MOS devices have been established, their use in some basic circuits can be discussed. The inverter will be considered first as, although it is the simplest logic function that can be implemented, it often forms the basis of more complex circuits.

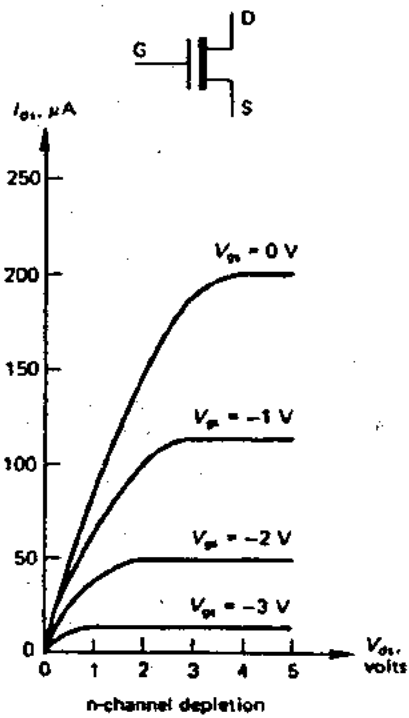
Figure 2.10 illustrates the principles involved in designing an inverter. It consists of a digital switch S which is closed if a high input voltage is applied and is open for a low input voltage. The switch output is connected via a load to the power rail. Thus a high input causes the output to be 0 V (low output) and current flows through the load and the switch. A low input leaves the switch open and no current flows in the circuit. As a result, no voltage is dropped across the load and the output is V_D (high output).

In MOS inverter design, an n-channel enhancement transistor is used as the digital switch. This is convenient as it has a positive threshold and a low input voltage below the device threshold causes the transistor to be off, while a high input voltage above the threshold causes it to be on.

The load can be implemented in a number of ways and in particular by the use of an MOS device. This results in a range of inverter circuits, each of which illustrates different design principles.

Many designers learn about bipolar junction transistors before being introduced to MOS devices and in that technology the most common implementation of an inverter uses a resistor as a load. Thus a convenient starting point for a description of MOS inverters is to consider a circuit which uses a resistor as the load.





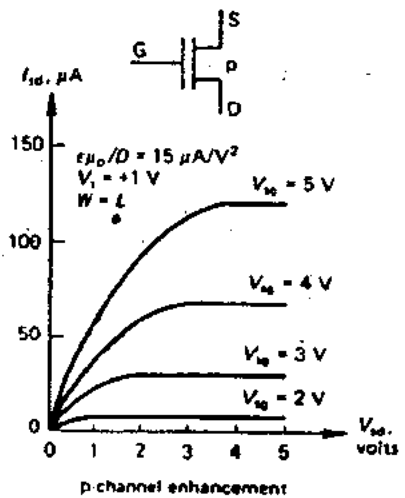


Figure 2.9 MOS transistors – typical characteristics

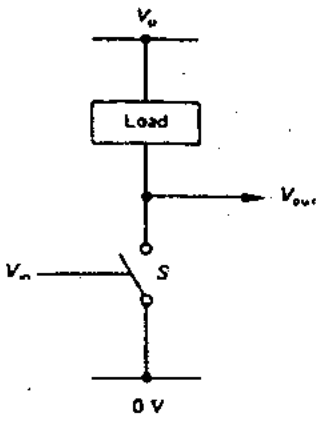


Figure 2.10 Inverter principle

2.7 NMOS Inverter with a Resistor Load

Consider applying a low input of 0 V to the circuit of figure 2.11. This is below the transistor's threshold voltage, V_{te}, and the transistor is off. No current flows and $V_{out} = V_p$. When a high input voltage of V_p is applied to the circuit, this is above the threshold so the transistor is on and current I_{ds} flows. By choosing the value of the resistance correctly, V_{out} can be made much less than V_{te} so that the output is low and can drive a succeeding stage.

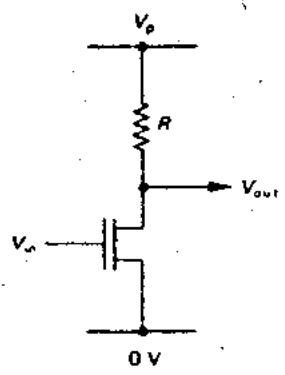


Figure 2.11 NMOS inverter with a resistor load

In order to see what value of R is required, it is necessary to take typical values for the parameters. V_p will be taken as 5 V since most MOS and bipolar logic families operate at this voltage. $\epsilon\mu_n/D$ will be taken as 30 μA/V^2 and V_{te} as 1 V. An equal gate width and length are assumed, so the width-to-length ratio of the transistor, known as the 'aspect ratio', is 1/1.

R is calculated to give a low V_{out} of 0.3 V_{te}, since this gives a reasonable noise margin between V_{out} and the gate threshold. Thus V_{ds} of the transistor is 0.3 V and, since V_{gs} is 5 V, $V_{ds} < V_{gs} - V_{te}$ and the device is operating in the resistive region. Equation (2.2) can now be used to find that

$$I_{ds} = 30\left(\frac{1}{1}\right)\left[(5-1)0.3 - \frac{0.3^2}{2}\right] = 34.7\ \mu\text{A}$$

This current also flows through the resistor. Hence

$$R = \frac{V_p - V_{out}}{I_{ds}} = \frac{5 - 0.3\ \text{V}}{34.7\ \mu\text{A}} = 135.4\ \text{k}\Omega$$

The silicon area required to implement this resistor is far larger than that for the transistor (typically by a factor of 300). Thus it is not practical to use a resistor as a load and so an MOS device is used instead.

2.8 NMOS Inverter with an NMOS Enhancement Transistor Load

This section considers the use of the same transistor type for both the load and the digital switch, see figure 2.12. T1 acts as the switch and is also referred to as the 'driver'. T2 is the load and its gate is connected to V_p in order to maximise its V_{gs}. Both transistors have the same threshold V_{te}.

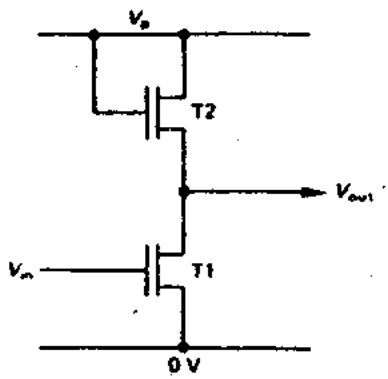


Figure 2.12 NMOS inverter with NMOS enhancement load

When V_{in} is low, it is less than V_{te} so transistor T1 is off. However, some very small leakage current flows through T1 and this is supplied by T2. Thus T2 is just brought to the edge of conduction and $V_{gs} - V_{te}$ for T2 is approximately 0 V, so

$$V_{out} = V_p - V_{te}$$

Again, taking V_p as 5 V and V_{te} as 1 V, the high level output voltage is 4 V if the body effect of T2 is neglected. However, T2's source-substrate voltage is significant at this magnitude of output voltage and the effect on the threshold voltage cannot be ignored. Using the expression for modifying the threshold voltage (given in section 2.3) with a substrate voltage of 0 V

$$V_{out} = V_p - V_{te} = V_p - \{V_{te0} + \gamma(V_{sb})^{1/2}\} = 5 - [1 + 0.5(V_{out})^{1/2}]$$

where the body effect constant γ has been taken as 0.5. Rearranging yields

$$(V_{out})^{1/2} = 8 - 2V_{out}$$

By squaring each side and solving the resulting quadratic equation, the high level V_{out} is found to be only 3.12 V.

The output from a gate is normally used to drive the input of other gates. Hence the high level input voltage for a gate is 3.12 V. When V_{in} is high at this level, it exceeds T1's threshold so T1 is on and conducts current. This current is supplied by T2 which is also on. The circuit output is required to be low for a high input and this can be obtained by selecting suitable widths and lengths for the gates of the driver and the load transistors. Again, it is appropriate to choose the gate sizes so that the low level output is 0.3 V_{te}.

Consider T1 with a V_{in} of 3.12 V and a V_{out} of 0.3 V. Since T1's source is at 0 V, its V_{gs} is 3.12 V and its V_{ds} is 0.3 V. Thus $V_{ds} < V_{gs} - V_{te}$ and so T1 operates in the resistive region. Applying equation (2.2) with an $\epsilon\mu_n/D$ of 30 μA/V^2

$$I_{ds} \text{ of T1} = 30\left(\frac{W_1}{L_1}\right)\left[(3.12 - 1)0.3 - \frac{0.3^2}{2}\right] = 17.73\left(\frac{W_1}{L_1}\right)\mu\text{A}$$

where (W_1/L_1) is the aspect ratio of transistor T1.

T2 is on and $V_{ds} = V_{gs}$. Hence the inequality $V_{ds} > V_{gs} - V_{te}$ is true and T2 is saturated. The gate-source voltage of T2 is $V_p - V_{out}$ and since V_{out} is low, the body effect can be neglected. Applying equation (2.3) with a V_{gs} of 4.7 V yields

$$I_{ds} \text{ of T2} = \frac{30}{2}\left(\frac{W_2}{L_2}\right)(4.7 - 1)^2 = 205.4\left(\frac{W_2}{L_2}\right)\mu\text{A}$$

where (W_2/L_2) is the aspect ratio of T2. I_{ds} of T1 and T2 are equated to obtain an inverter ratio k for the circuit

$$k = \frac{(W_1/L_1)}{(W_2/L_2)} = 11.6$$

It is usual to split the inverter ratio obtained between transistors T1 and T2 in order to prevent T1 from occupying a much larger area than T2. Thus a ratio of 11.6 would be taken as 12 for convenience and could be split as $(W_1/L_1) = 3/1$ and $(W_2/L_2) = 1/4$.

The main drawback of an NMOS enhancement load, apart from the loss of voltage for high outputs, is the speed of the rising edge of the gate. This can be appreciated by considering that there is some capacitance C_{out} on V_{out}, and V_{out} is at 0.3 V corresponding to a V_{in} of 3.12 V. If V_{in} now switches to 0.3 V, T1 turns off and V_{out} starts to rise. Initially, V_{gs} of T2 is 4.7 V but as V_{out} rises, V_{gs} of T2 decreases and T2's threshold rises (because of the body effect) so that less current flows in T2. This progressively slows down the rate of charging up C_{out}, and as V_{out} approaches 3.12 V, T2 tends to turn off and very little current flows to finish the charging of C_{out}. For this reason, n-channel depletion devices are preferred as a load.

2.9 NMOS Inverter with an NMOS Depletion Transistor Load

Referring to figure 2.13, the NMOS depletion transistor T2 forms the circuit load. The depletion threshold V_{td} of T2 is negative and since the gate-source voltage of T2 is zero, T2 is always on.

When V_{in} is low, it is less than the enhancement threshold V_{te} of T1, and T1 is off. T2 is on and supplies the very small leakage current of T1. Since I_{ds} of T2 is negligible, the drain-source voltage of T2 is very small, leading to a V_{out} of V_p. Thus there is no high level output voltage loss as with the NMOS enhancement load.

The high level input voltage to a gate is therefore V_p. This exceeds T1's threshold voltage and hence T1 and T2 are both on. Again it is necessary to choose the gate dimensions of the transistors in order to obtain the desired low level V_{out}. Taking V_{out} as $0.3V_{te}$ and V_p as 5 V again, T1's V_{gs} is 5 V and its V_{ds} is 0.3 V. Thus $V_{ds} < V_{gs} - V_{te}$ and T1 is operating in the resistive region. Equation (2.2) is applied to express I_{ds} of T1 in terms of its aspect ra

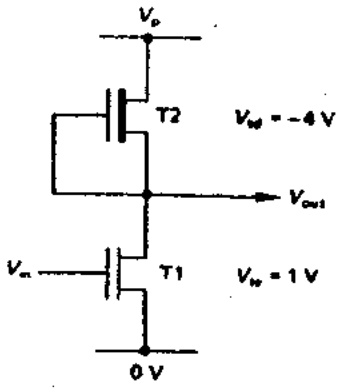


Figure 2.13 NMOS inverter with NMOS depletion load

$$I_{ds} \text{ of T1} = 30\left(\frac{W_1}{L_1}\right)\left[(5-1)\,0.3 - \frac{0.3^2}{2}\right] = 34.65\left(\frac{W_1}{L_1}\right)\mu\text{A}$$

T2 is on with a V_{gs} of 0 V and a V_{ds} of $V_p - V_{out} = 4.7$ V. Thus $V_{ds} > V_{gs} - V_{td}$ and T2 is saturated. Equation (2.3) is applied to find T2's I_{ds}

$$I_{ds} \text{ of T2} = \frac{25}{2}\left(\frac{W_2}{L_2}\right)[0-(-4)]^2 = 200\left(\frac{W_2}{L_2}\right)\mu\text{A}$$

Note that $\epsilon\mu_n/D$ for a depletion transistor is typically 25 μA/V^2 compared with 30 μA/V^2 for an NMOS enhancement device (owing to the fact that a depletion transistor requires additional impurity doping to define a threshold of -4 V). Equating the current in T1 and T2 yields

$$k = \frac{(W_1/L_1)}{(W_2/L_2)} = 5.8$$

For convenience, this inverter ratio would be taken as 6 and split as $(W_1/L_1) = 3/1$ and $(W_2/L_2) = 1/2$.

It is possible to reduce the inverter ratio and thus the silicon area occupied by T1 and T2 at the expense of a reduced noise margin. The lower limit for k is generally accepted to be 4 and this allows 0.5 V between the low level V_{out} and the enhancement threshold. However, the author recommends that a value of $k = 6$ or greater be adopted for practical designs.

2.10 Edge Times for NMOS Inverter with a Depletion Load

The circuit of figure 2.13 represents a practical inverter that can be implemented in NMOS technology. It is thus worth examining the edge time response of the gate, as this will indicate the maximum speed of operation and the factors upon which the speed is dependent.

In the circuit of figure 2.13 the maximum current capability of T1 and T2 is different. This can be seen by considering the saturation currents of the devices, and it has already been shown that this current for T2 is $200(W_2/L_2)\,\mu A$. The maximum current through T1 occurs when the gate-source voltage is a maximum (5 V) and the device is saturated. Using equation (2.3)

$$\text{maximum } I_{ds} \text{ of T1} = \frac{30}{2}\left(\frac{W_1}{L_1}\right)(5-1)^2 = 240\left(\frac{W_1}{L_1}\right)\mu A$$

Thus

$$\frac{\text{maximum } I_{ds} \text{ of T1}}{\text{maximum } I_{ds} \text{ of T2}} = \frac{6(W_1/L_1)}{5(W_2/L_2)} = \frac{6k}{5}$$

Let the total capacitance on the output be C_{out}, as shown in figure 2.14. C_{out} arises because of the capacitance of the gate of T2, the source of T2 and the drain of T1. In addition, C_{out} includes the capacitance of the connections between V_{out} and these transistor terminals. Thus C_{out} is proportional to the silicon area necessary to implement these features. If the minimum length or width allowable is 6 μm, then a typical value of C_{out} for this circuit is 0.1 pF.

Consider that the input is high and that the output has settled to its low level value of 0.3 V. If the input now changes instantaneously from high to low, then T1 turns off. T2 remains on and supplies current to charge the load capacitance (see figure 2.14). This causes the output to gradually rise to 5 V. The current I_c in the capacitor is related to the change in voltage dV_{out} across it with respect to time dt by

$$I_c = C_{out}\frac{dV_{out}}{dt}$$

The rise time is dependent upon the rate at which the capacitance can be charged during this period. This in turn depends upon the current available from T2.

When the input changes instantaneously from low to high, T1 turns on. Both T1 and T2 are on and during the fall time, T1 accepts current from T2 and from C_{out} as it discharges (see figure 2.14). Since the current in T2 is small compared with that in T1, the current T1 can accept determines the rate at which the capacitor discharges and hence the fall time.

It can therefore be expected that the falling and rising edge times will differ by a factor equal to the ratio of the current capability of T1 and T2. On this basis, the rising edge time will be 6k/5 times that of the falling edge time. This is illustrated in figure 2.15.

Consider in detail the rising edge at V_{out}. The current supplied by T2 causes the output to rise from 0.3 V to 5 V, as shown in figure 2.16a. It is usual to calculate edge times from the 10 per cent to 90 per cent points on the output waveform, as many edges are exponential in shape and take a disproportionately long time to settle to their final value. Applying this to the depletion load inverter, the rise time is the time for the output to rise from 0.8 V to 4.5 V.

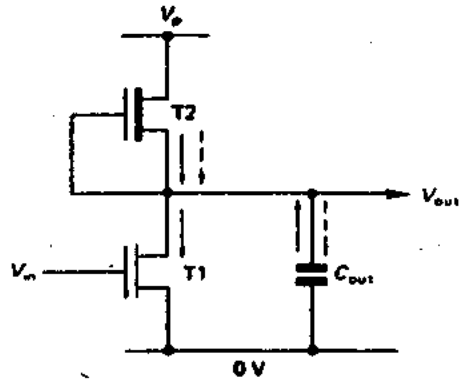


Figure 2.14 Current flow during edge times: ↑ rising edge current flow, ↓ current flow during falling edge

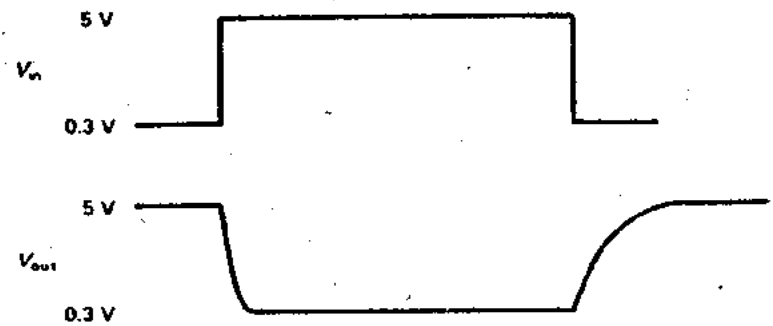


Figure 2.15 Output response to an input step

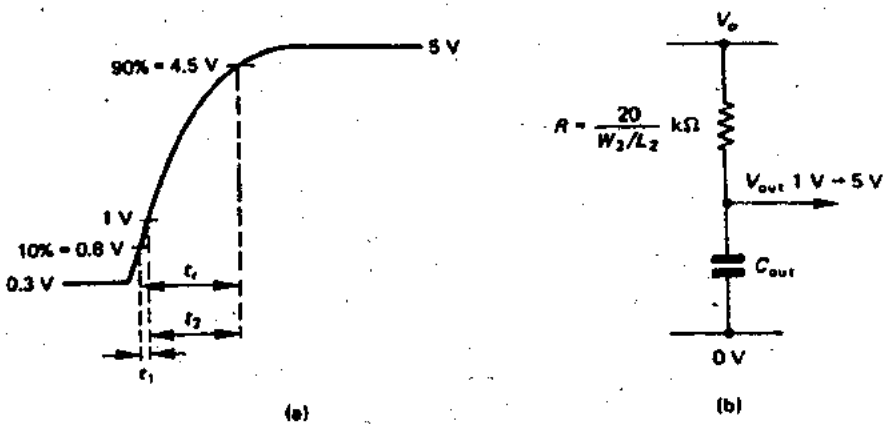


Figure 2.16 Rise time behaviour: (a) rising edge; (b) circuit approximation

During the output rise, T2 is saturated until the output reaches 1 V and thereafter is in the resistive mode. Thus the rise time can be split into two regions: t_1 is the time for the output to rise from 0.8 V to 1 V where T2 is in the saturated mode, and t_2 is the time for the output to rise from 1 V to 4.5 V where T2 operates in the resistive mode. It will be appreciated, from figure 2.16a, that since the initial part of the rise at V_{out} is fast in comparison with the latter parts and since the output voltage change during t_1 is small in comparison with the voltage change during t_2, the rise time is effectively determined by t_2 and t_1 can be neglected.

The calculation of t_2 is most easily approached by replacing T2 with a constant resistance R which is used to represent the transistor's behaviour in the resistive region of the characteristic. The resistance's characteristic is shown as a dotted line on figure 2.17. It has been obtained by drawing a straight line between the origin and the pinch-off point. Using this approximation to transistor behaviour, a value of R (applicable to any NMOS transistor) can be obtained.

$$R = \frac{\text{pinch-off voltage}}{\text{pinch-off current}} = \frac{(V_{gs} - V_t)}{\frac{\epsilon\mu_n}{2D}\left(\frac{W}{L}\right)(V_{gs} - V_t)^2}$$

$$= \frac{1}{\frac{\epsilon\mu_n}{2D}\left(\frac{W}{L}\right)(V_{gs} - V_t)}$$

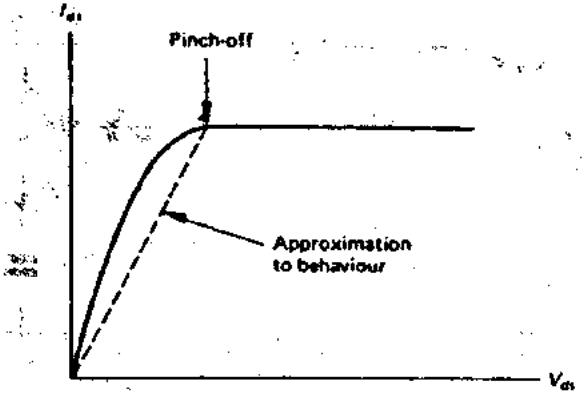


Figure 2.17 Characteristic approximation

Considering transistor T2 and substituting for V_{gs}, V_{td} and $\epsilon\mu_n/D$, the resistance can be expressed in terms of T2's aspect ratio

$$R = \frac{1}{\frac{25}{2}(W_2/L_2)[0-(-4)]}\ \text{M}\Omega = \frac{20}{(W_2/L_2)}\ \text{k}\Omega$$

Thus the equivalent circuit during time t_2 simplifies to an RC network, as shown in figure 2.16b. The output voltage V_{out} across such a network, where the initial voltage across the capacitor is V_i and the final voltage is V_p, is given by

$$V_{out} = V_p - (V_p - V_i)\exp[-t/(RC_{out})]$$

The time for the output to reach 4.5 V with an initial output of 1 V is

$$4.5 = 5 - (5-1)\exp[-t_2/(RC_{out})]$$

giving

$$t_2 = 2.08RC_{out}$$

If C_{out} is in picofarads and R in kilohms, then t_2 is given in nanoseconds. Substituting for T2's value of R, the rise time is

$$t_r = t_2 = \frac{42C_{out}}{(W_2/L_2)} \text{ ns}$$

From previous considerations concerning the current drive capability of T1 relative to T2, the fall time t_f is

$$t_f = \frac{5t_r}{6k} = \frac{5(W_2/L_2)42C_{out}}{6(W_1/L_1)(W_2/L_2)}$$

$$t_f = \frac{35C_{out}}{(W_1/L_1)} \text{ ns}$$

The expression for the fall time agrees well with results obtained by simulating the inverter circuit. However, the expression for the rise time is not so accurate as the body effect is significant for T2. As the output rises, the threshold of T2 rises, reducing the current in T2. This increases the time necessary to charge the output capacitance. Simulation results for the inverter show that

$$t_r = \frac{60C_{out}}{(W_2/L_2)} \text{ ns}$$

is more appropriate for the rise time.

In a fabrication process where the minimum allowable gate width or length is 6 μm, an inverter having T1 and T2 aspect ratios of 3/1 and 1/2 respectively will have a typical input capacitance of 0.05 pF and an output capacitance of 0.1 pF. Thus the circuit exhibits a rise time of 12 ns and a fall time of 1.2 ns.

The speed at which the circuit operates is dominated by the load capacitance. In practice, the output of a circuit drives other inputs and each input causes an increase in the load capacitance and hence the edge times. The absolute maximum frequency (in hertz) at which a circuit can operate is $1/(t_r + t_f)$ so, taking the capacitance figures of the last example, an inverter driving a similar inverter has a total load capacitance of 0.15 pF and an upper frequency limit of 50 MHz.

2.11 Ratioed and Ratioless Design

It is clear that for NMOS inverters with an NMOS depletion or enhancement load, it is necessary to choose suitable gate geometry ratios for both the load and drive transistors in order to obtain the desired low level output. Such designs are therefore called 'ratioed' circuits.

It should be noted that the width-to-length ratio of a transistor is an important design characteristic of MOS circuits. This ratio determines the circuit speed, as

edge times are inversely proportional to aspect ratios. The power dissipation is also determined by the gate geometry, since the current flow through a device is proportional to its aspect ratio.

Ratioless designs are obtained by implementing the inverter load of figure 2.10 with a digital switch where the load switch and the driver switch are never simultaneously closed. Consider that the load and driver switches operate in antiphase, as shown in figure 2.18. A high input causes the driver switch to close and the load switch to remain open, V_{out} is connected to 0 V. A low input closes the load switch and opens the driver switch, V_{out} is connected to V_p.

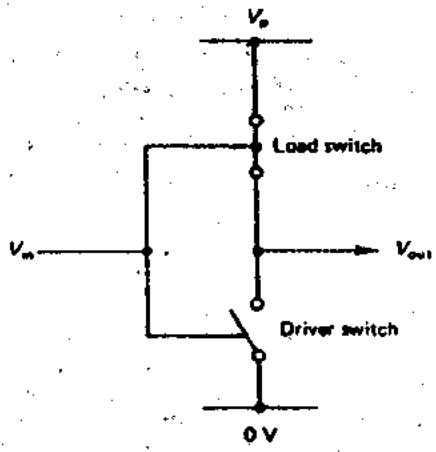


Figure 2.18 Ratioless inverter design

Since only one switch is normally closed at a time, no current flows between V_p and 0 V. Thus there is no static power dissipation. However, note that when the circuit changes state, both switches may become closed, causing a transient current to flow.

The switches are implemented with MOS transistors and, because the load and driver transistors are never on simultaneously, the correct output levels are obtained regardless of the transistor aspect ratios. For this reason, such circuits are described as 'ratioless'.

2.12 The CMOS Inverter

The complementary-channel MOS or CMOS inverter is shown in figure 2.19. It has an n-channel enhancement transistor as the driver switch and a p-channel enhancement device as the load switch. It is assumed that the substrate of T1 is

connected to 0 V and T2's substrate to V_p. The threshold of T1, V_{te}, is the minimum gate-source voltage at which T1 conducts, while T2's threshold, V_{tep}, is the minimum source–gate voltage at which it turns on.

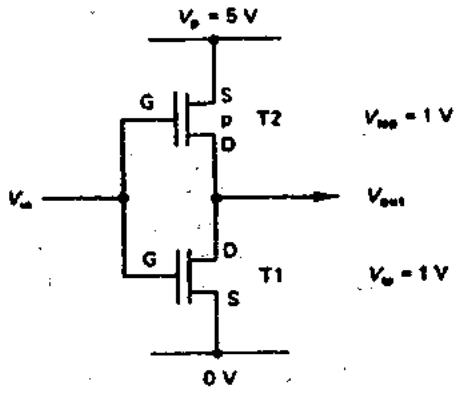


Figure 2.19 CMOS inverter

A low input voltage of 0 V causes T1 to be off since the input is less than T1's threshold. T2 is on as its source–gate voltage is 5 V, which exceeds its 1 V threshold. The current flowing through T2 is the leakage current of T1 which is very small. Thus V_{sd} of T2 is 0 V and V_{out} is 5 V; T2 is in the resistive mode.

For a high input voltage of 5 V, T1 is on since the input exceeds its threshold voltage. T2 is off since its source–gate voltage is 0 V, which is below its 1 V threshold. Again, the only current flowing is a very small leakage current (of T2), so V_{ds} of T1 and hence V_{out} equal 0 V; T1 is in the resistive mode.

If the current capability of T1 and T2 during switching is the same, then the rising and falling edge times are approximately equal. When V_{in} changes from a low to a high level instantaneously, T1 turns on and T2 turns off. The load capacitance on the output discharges from 5 V to 0 V via T1. T1 is saturated until the output falls to 4 V and thereafter is in the resistive region. Its V_{gs} is 5 V throughout the fall time.

Similarly, when V_{in} changes from high to low instantaneously, T1 turns off and T2 turns on. The load capacitance on the output charges from 0 V to 5 V via T2. During the rise, T2 is saturated until the output reaches 1 V and thereafter is in the resistive mode. V_{sg} of T2 is 5 V throughout the output rise.

Equations (2.3) and (2.5) are used to find the saturation current of the transistors and equating these currents yields

$$\frac{\epsilon\mu_n}{2D}\left(\frac{W_1}{L_1}\right)(5-1)^2 = \frac{\epsilon\mu_p}{2D}\left(\frac{W_2}{L_2}\right)(5-1)^2$$

giving

$$\frac{(W_1/L_1)}{(W_2/L_2)} = \frac{\mu_p}{\mu_n}$$

Since the mobility of holes is only approximately half that of electrons, it is necessary to make the aspect ratio of T2 twice that of T1 to obtain equal edge times. It will be assumed throughout the rest of this section that T1's aspect ratio is 1/1 and T2's aspect ratio is 2/1. It is clearly advantageous in terms of the silicon area occupied to use minimum-sized transistors where possible and T1 can have minimum dimensions.

T1 and T2 both act as drivers and accept or supply current to charge the load capacitance when the gate switches. Edge times in CMOS approximate to the falling edge time for an NMOS inverter with a depletion load. Thus

$$t_r = t_f \approx \frac{35C_{out}}{(W_1/L_1)} \text{ ns}$$

where C_{out} is the output load capacitance in pF.

To understand the gate behaviour during switching, it is helpful to examine the gate transfer characteristic (V_{out} versus V_{in}) and the variation of circuit current with input voltage. The easiest way of determining these features is graphically. This is done by first plotting T1's I_{ds} versus V_{ds} curves for different values of V_{gs} above the threshold. Effectively, this is a graph of the circuit current I versus V_{out} for various values of V_{in}. The resulting curves are shown with broken lines in figure 2.20. It will be noted that the behaviour below pinch-off is taken to be linear. This assumption has a negligible effect upon the transfer and current characteristics to be obtained and greatly simplifies the preparation and plotting of the family of curves for the transistors. Table 2.1 illustrates the simple calculations from which the curves for T1 have been drawn in figure 2.20.

Next, the curves for T2 are superimposed on those for T1. Here, V_{gs} of T2 is $5 - V_{in}$, V_{ds} of T2 is $5 - V_{out}$ and the current I_{ds} is the circuit current I. Thus the T2 curves can be expressed in terms of V_{out} versus I for different values of V_{in}. These are shown in figure 2.20 as solid lines and table 2.1 shows in detail the calculations necessary to draw the curves. Again the behaviour below pinch-off is assumed to be linear.

By reading values off the superimposed characteristic plots for T1 and T2, V_{out} versus V_{in} and I versus V_{in} can be found. For example, when $V_{in} = 2$ V figure 2.20 shows that the T1 curve for this value of input intersects the T2 curve for $V_{in} = 2$ V at a V_{out} of 4.5 V and a circuit current of 15 μA. It can also be seen that at this operating point, T1 is saturated and T2 is resistive.

Table 2.1 Current and voltage calculations for CMOS inverter

Transistor T1

$(W_1/L_1) = 1/1$, $\epsilon\mu_n/D = 30\ \mu A/V^2$, V_{te} of T1 = 1 V
Saturation $I_{ds} = 15(V_{gs} - 1)^2\ \mu A$
Pinch-off occurs when $V_{ds} = V_{gs} - 1$
$V_{in} = V_{gs}$, $V_{out} = V_{ds}$, $I = I_{ds}$

V_{gs} (V)	*Pinch-off* V_{ds} (V)	*Saturation* I_{ds} (μA)
2	1	15
2.5	1.5	33.8
3	2	60
4	3	135
5	4	240

Transistor T2

$(W_2/L_2) = 2/1$, $\epsilon\mu_p/D = 15\ \mu A/V^2$, V_{tep} of T2 = 1 V
Saturation $I_{sd} = 15(V_{sg} - 1)^2\ \mu A$
Pinch-off occurs when $V_{sd} = V_{sg} - 1$
$V_{in} = 5 - V_{sg}$, $V_{out} = 5 - V_{sd}$, $I = I_{sd}$

V_{sg} (V)	*Pinch-off* V_{sd} (V)	*Saturation* I_{sd} (μA)	V_{in} (V)	V_{out} *at pinch-off* (V)
2	1	15	3	4
2.5	1.5	33.8	2.5	3.5
3	2	60	2	3
4	3	135	1	2
5	4	240	0	1

Figure 2.21, a and b, show the resulting V_{out} versus V_{in} and I versus V_{in} curves obtained from figure 2.20. It can be seen from the transfer characteristic that there is a very sharp transition in V_{out} at $V_{in} = V_p/2$. During this transition, both T1 and T2 are saturated and the current I rises to 33.8 μA. The frequency of switching therefore determines the overall power dissipation. The power dissipated is also dependent upon the value of V_p. A greater value for V_p increases the current capability, reducing the switching time at the expense of an increased power dissipation. Nowadays, it is common practice to operate CMOS logic from a 5 V supply.

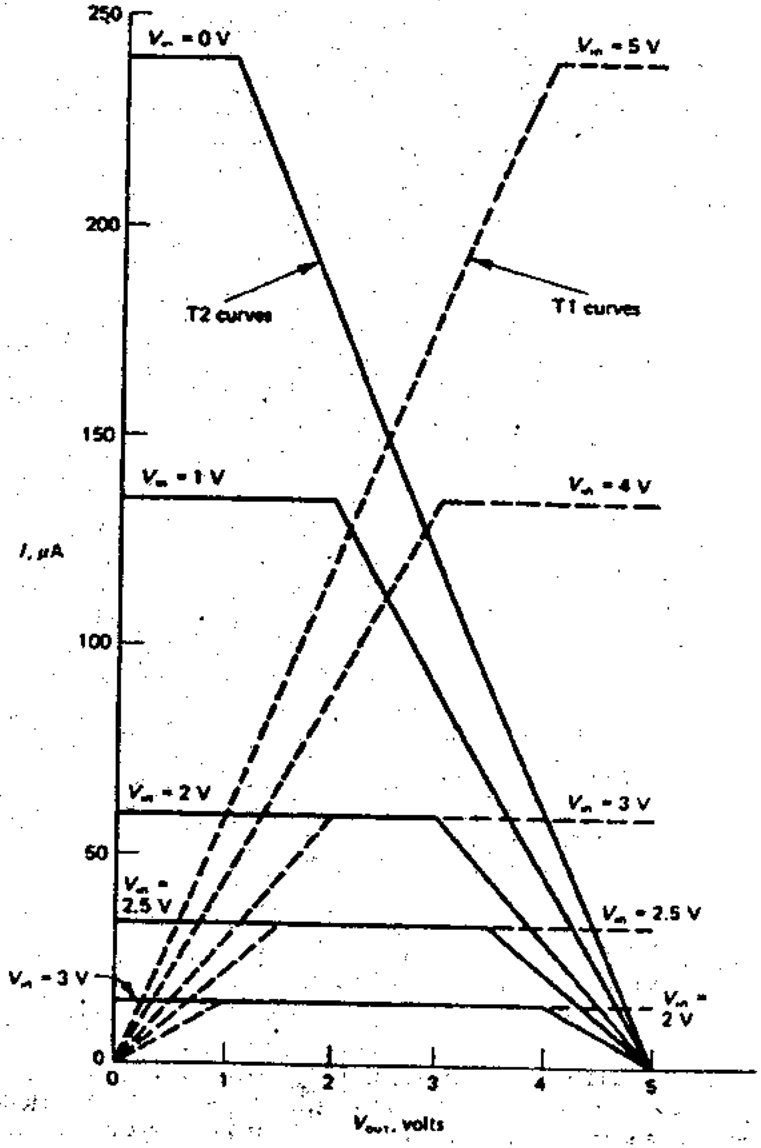


Figure 2.20 I versus V_{out} curves for T1 and T2

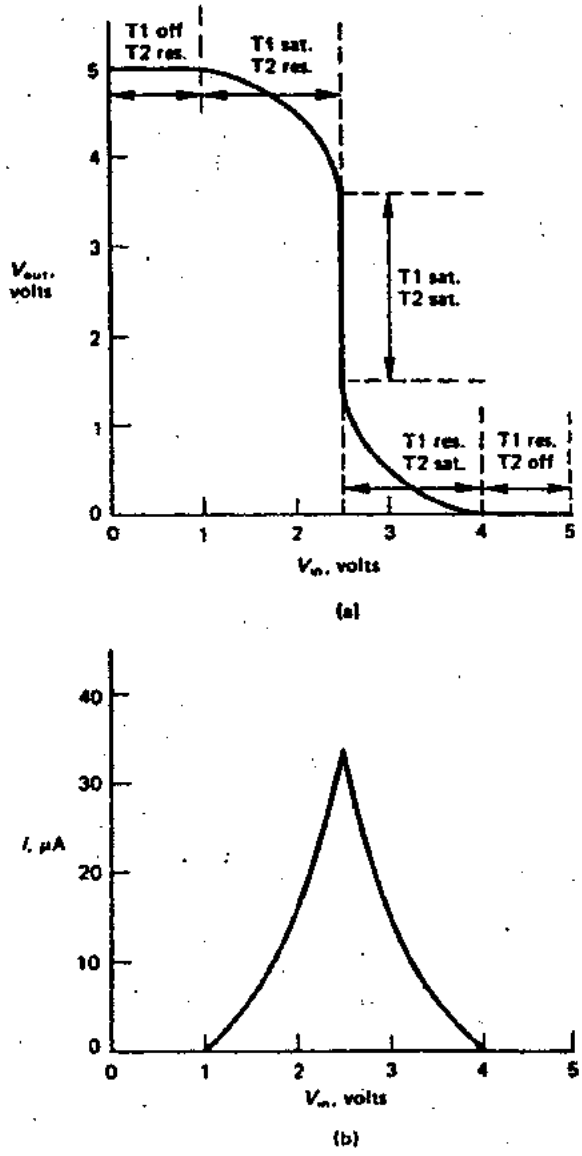


Figure 2.21 Characteristic curves for the CMOS inverter: (a) transfer characteristic, (b) current versus V_{in}

2.13 NMOS Pass Transistors

Another important and fundamental MOS circuit is that of the pass transistor. Here the device acts as a voltage-controlled switch, allowing the device's input to be selectively transmitted to its output. Figure 2.22 shows an NMOS pass transistor; it consists of a minimum geometry n-channel enhancement device. It is usual not to label the input and output terminals, as both can act as the drain or source, depending on the applied and existing voltages. The load capacitance C_{out} is the capacitance inherent in the circuit and will include the capacitance of any gates driven from V_{out}. It should be emphasised that C_{out} is not a discrete capacitor.

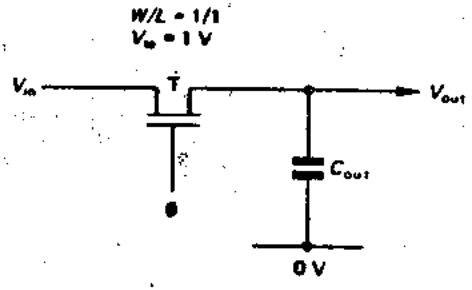


Figure 2.22 NMOS pass transistor

Assume, as before, that a low logic level is 0 V and a high level is V_p. The state of the switch is controlled by the level of ϕ. If ϕ is low then transistor T is off, regardless of the logic levels existing at V_{in} or V_{out}. Since there is no connection between V_{in} and V_{out}, V_{out} remains at its existing voltage. This output level is maintained as charge on C_{out} and, since the leakage current from this point is very small, the voltage will decay towards 0 V at a slow rate.

When ϕ goes high to V_p, the transistor T turns on and the action now depends on the value of V_{in} and the initial value of V_{out}. There are four cases.

(a) $V_{in} = 0$ V, initial $V_{out} = 0$ V

No current flows through the device as its V_{ds} is 0 V. Hence V_{out} remains at 0 V.

(b) $V_{in} = V_p$, initial $V_{out} = 0$ V

Current flows from V_{in} to V_{out}, causing V_{out} to rise as C_{out} is charged. Here the input acts as the drain and the output as the source. Initially V_{gs} is V_p but as the output rises, the gate-source voltage drops and the rise at V_{out} is halted when $V_{gs} - V_{to} = 0$ V. The rise is exactly analogous to the high level output from an NMOS inverter with an NMOS enhancement load. Thus, as V_{out} rises, the

threshold voltage rises because of the body effect (section 2.3) and consequently V_{out} only rises to 3.12 V if V_p is 5 V.

(c) $V_{in} = V_p$, initial $V_{out} = V_p - V_{te}$

No current flows as $V_{gs} - V_{te}$ is 0 V. V_{out} remains at $V_p - V_{te}$.

(d) $V_{in} = 0$ V, initial $V_{out} = V_p - V_{te}$

Current flows from V_{out} to V_{in} causing V_{out} to discharge to 0 V. Here the output acts as the drain and the input as the source. Since V_{gs} is constant at V_p during the discharge time, the fall time at V_{out} is considerably faster than the rise time encountered in case (b).

The pass transistor or transmission gate can be regarded as passing an input value to the output under the control of the input ϕ. Such gates perform the and function during the application of ϕ. Pass transistors allow the designer to implement this function in a much smaller silicon area than other designs allow. This arises because the pass transistor can be of minimum geometry whereas the conventional NMOS implementation is a ratioed design of three devices requiring interconnections between the different layers of the transistor structure; the saving in area is by approximately a factor of 18. However, it should be noted that the pass transistor is a dynamic circuit and once the transistor is off, the output state is held as charge on the capacitor C_{out}. This state is only temporary since there is a small amount of leakage current through the transistor, causing a high level output to slowly decay to 0 V. Dynamic logic circuits therefore have a minimum frequency of operation (typically 5 KHz) in order to avoid a significant deterioration of the output signal.

Pass transistors can be connected to other pass transistors in a chain, as shown in figure 2.23. As the input passes down the chain, the delay gets progressively longer because of the capacitance of each device output. At the end of the chain, the logic level is normally restored by an NMOS inverter with a depletion load. With a V_{in} of 3.12 V, an inverter ratio of 11.3 is necessary to produce a low level output of 0.3 V from the restoring inverter.

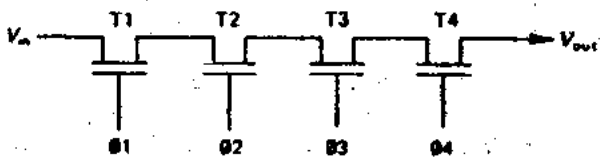


Figure 2.23 A chain of pass transistors

The reader should be aware of one further effect of using pass transistors. Consider figure 2.22 with the control input at V_p and the output at $V_p - V_{to}$. There is a small capacitance, C_{gs}, between the gate and the output as a result of a slight overlap of these features in the transistor structure. Thus, removing the signal ϕ is equivalent to applying a negative step of voltage to two capacitors C_{gs} and C_{out} in series, as shown in figure 2.24.

This causes V_{out} to fall by $V_p[C_{gs}/(C_{gs} + C_{out})]$. C_{gs} for a minimum geometry device where $L = W = 6\ \mu m$ is typically 0.0025 pF and C_{out} is of the order of 0.1 pF if V_{out} drives a restoring inverter. Thus a 5 V fall in ϕ causes a drop of 0.12 V at V_{out}, reducing the high level V_{out} to about 3 volts.

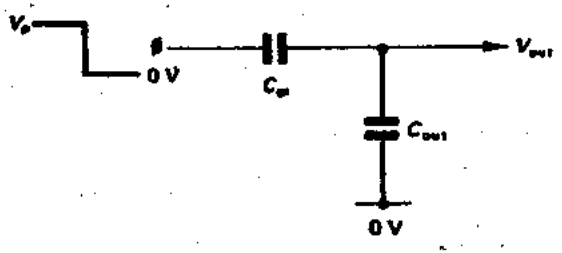


Figure 2.24 Effect of removing the control input in a pass transistor

2.14 Ratioless NMOS Inverter

An example demonstrating NMOS pass transistor features is the ratioless NMOS inverter (see figure 2.25). Referring to figures 2.18 and 2.25, the load switch consists of transistor T3 while the series combination of transistors T1 and T2 form the driver switch. The operation relies on the use of phased clocks which ensure that the load and driver switches are not closed simultaneously. Only one of $\phi 1$ and $\phi 2$ is high at any time, so T3 is off if T2 is on and vice versa. Thus the design is ratioless and all transistors can be minimum geometry NMOS enhancement devices.

Taking $\phi 1$ high turns T3 on. T2 is off because $\phi 2$ is low and hence, regardless of the level at V_{in}, there is no conduction path between V_{out} and 0 V. Thus the capacitance inherent in the circuit at V_{out} is charged via pass transistor T3 to $V_p - V_{to}$. Again, because of the body effect, V_{out} only rises to 3.12 V; the output is always pre-charged high during $\phi 1$. Once $\phi 1$ is removed, T3 turns off. In the period before $\phi 2$ goes high, both $\phi 1$ and $\phi 2$ are low and all transistors connected to the output are off. The output state is indicated by the charge stored on C_{out}.

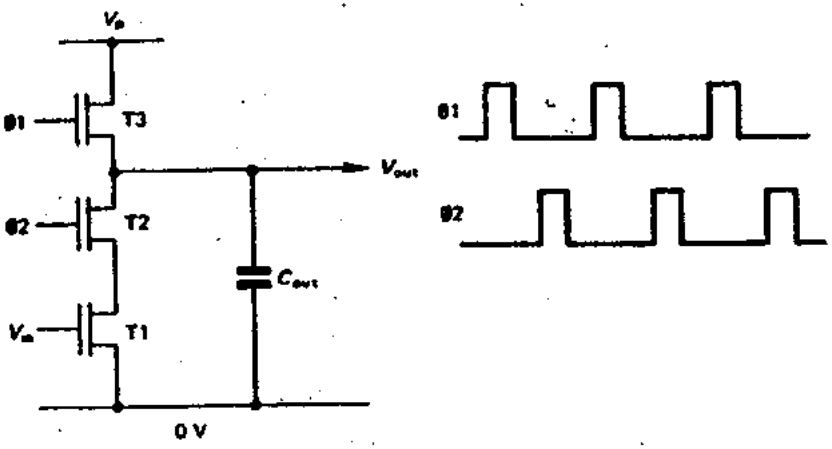


Figure 2.25 NMOS ratioless inverter

When $\phi 2$ is applied, T2 turns on. T3 is off since $\phi 1$ is low, so there is no conduction path between V_{out} and V_p. The circuit action now depends on the level at V_{in} during $\phi 2$. If V_{in} is low, T1 is off so there is no conduction path between V_{out} and 0 V; V_{out} remains high. However, if V_{in} is high both T1 and T2 are on and C_{out} discharges via the series pass transistor chain T1 and T2 to 0 V; V_{out} becomes 0 V.

When $\phi 2$ is removed, both T2 and T3 are off so that the output state is again indicated by the charge on C_{out}. Thus $\phi 1$ and $\phi 2$ have to be regularly applied to avoid a loss of state at the output as a result of charge leakage. It should be noted that the application and removal of the phased clocks will cause small variations in the output voltage, owing to capacitive coupling between them and V_{out}. Clearly, the output is valid – that is, it is the inverse of the input – only between the application of $\phi 2$ and $\phi 1$. Between the application of $\phi 1$ and $\phi 2$, the output is always high. Pre-charging via a single transistor is of particular use in CMOS circuits as it eliminates the need for a complex pull-up circuit.

2.15 CMOS Pass Gate

Figure 2.26 shows a CMOS transmission gate consisting of an NMOS device with a V_{tn} threshold and a PMOS device with a V_{tep} threshold connected in parallel. When ϕ is low, both transistors are off and the output level remains constant.

When ϕ is taken high, both devices turn on. Again there are four cases to consider.

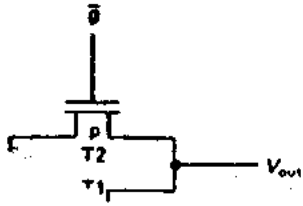


Figure 2.26 CMOS ... gate

(a) $V_{in} = 0$ V, initial $V_{out} = 0$ V
The drain and source potentials of each device are ... V_{out} remains at 0 V.
(b) $V_{in} = V_p$, initial $V_{out} = V_p$
Again the drain and source potentials of T1 and T2 are equal, so no current flows and V_{out} remains at V_p.
(c) $V_{in} = V_p$, initial $V_{out} = 0$ V
Here V_{in} acts as the drain of T1 and the source of T2. Current flows from V_{in} to V_{out}, causing V_{out} to rise as the capacitance C_{out} on the output charges up. V_{gs} of T2 is V_p throughout this rise, while V_{gs} of T1 is initially V_p but decreases as V_{out} rises. At $V_{out} = V_p - V_{tn}$, the NMOS transistor turns off and V_{out} continues to rise to V_p via T2.
(d) $V_{in} = 0$ V, initial $V_{out} = V_p$
V_{in} acts as the source of T1 and the drain of T2. Current flows from V_{out} to V_{in}, causing V_{out} to fall as C_{out} discharges. V_{gs} of T1 is constant at V_p during the fall, while V_{gs} of T2 is initially V_p but decreases as V_{out} falls. At $V_{out} = V_{tep}$, the PMOS device turns off and V_{out} continues to fall to 0 V via T1.

2.16 Buffer Circuits

Some gate outputs in a design, for example clock signals, need to be connected to a large number of gates and thus drive a large capacitive load. The effect of an output driving many inputs directly can be seen from the edge time results obtained in section 2.10. Here it was found that

$$\text{edge time} \propto \frac{C_{out}}{W/L}$$

where W/L is the pull-down transistor aspect ratio.

Thus the edge time increases in proportion to the capacitance driven, and soon becomes unacceptably slow. This speed loss can be avoided by suitably increasing the width-to-length ratio of all transistors in the driving gate; this increases its current capability. Unfortunately, this increase in the dimensions of the driver gate causes its input capacitance to rise. This in turn increases the loading on its preceding gate, causing an unacceptable loss of speed here.

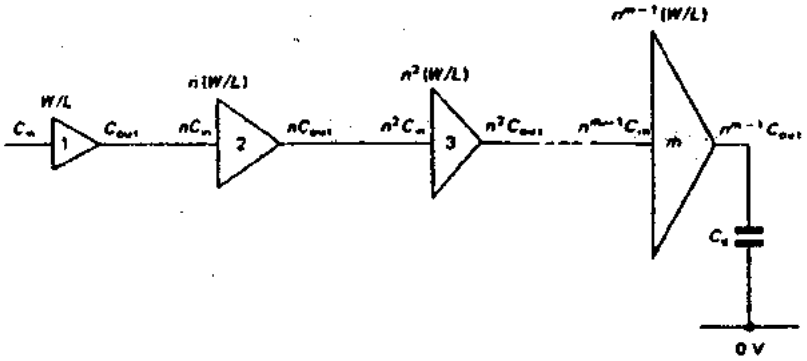


Figure 2.27 Buffer gate chain

For these reasons, a large capacitive load cannot be directly driven from a standard gate output. Instead, a buffer chain is used, as shown in figure 2.27. The aspect ratio of each gate in the chain is larger by a factor n than those of the preceding gate. This increase in dimensions causes the input and output capacitance of a gate to be n times that of the preceding gate. These are indicated on figure 2.27 relative to the input and output capacitance, C_{in} and C_{out}, of gate 1. If the number of stages in the chain, m, is chosen so that the capacitance C_d to be driven by gate m is n times this gate's input capacitance, then the load capacitance on each output increases by a factor of n at each stage and

$$C_d = n^m C_{in}$$

giving

$$m = \frac{\log_e (C_d/C_{in})}{\log_e n}$$

The effect of this cascade arrangement upon the edge times can be estimated from the aspect ratio and load capacitance of each gate in the chain. The load capacitance of gate i is $n^{i-1}(C_{out} + nC_{in})$ and its aspect ratio is $n^{i-1}(W/L)$. So

$$\text{edge time of gate } i \propto \frac{n^{i-1}(C_{out} + nC_{in})}{n^{i-1}(W/L)}$$

$$\propto \frac{C_{out} + nC_{in}}{W/L}$$

This expression is independent of the gate's position in the chain and thus the edge time is the same for each buffer gate. This is to be expected since, although the capacitance increases by a factor of n at each stage, the current capability at each stage increases by the same factor.

The edge time can be regarded as a measure of the delay through a gate and so the delay t_d through the chain is

$$t_d \propto \frac{(C_{out} + nC_{in})m}{W/L}$$

Substituting in $\log_e (C_d/C_{in})/\log_e n$ for m gives

$$t_d \propto \frac{(C_{out} + nC_{in}) \log_e (C_d/C_{in})}{(W/L) \log_e n}$$

$$\propto \frac{C_{out}/C_{in} + n}{\log_e n}$$

Differentiating this expression with respect to n yields

$$\frac{dt_d}{dn} \propto \frac{\log_e n - (C_{out}/C_{in} + n)/n}{(\log_e n)^2}$$

Taking C_{out} as 0.1 pF and C_{in} as 0.05 pF and then equating dt_d/dn to zero to obtain a value for n which minimises the delay, the optimum value for n is 4.3. If this is adopted, the number of stages is $\log_e (C_d/C_{in})/1.46$. Thus a 75 pF load representing 1500 driven gates requires a five-stage chain. In practice, the silicon area occupied by such a chain soon becomes significant and it is necessary to accept an increased delay in order to reduce the area and the number of stages required.

In NMOS, the rise and fall times are unequal. It is therefore usual to implement buffer gates with superbuffer circuits which have approximately equal rise

and fall times. In the non-inverting superbuffer of figure 2.28, T1 and T2 form an NMOS invertor. Hence V_{in} is applied to T3's gate and V_{in} to T4's gate. Thus for a high and low level input, $V_{out} = V_{in}$ and V_{gs} of T4 is approximately 0 V. An inverting superbuffer is obtained by exchanging the connections from T1 and T2 to T3 and T4.

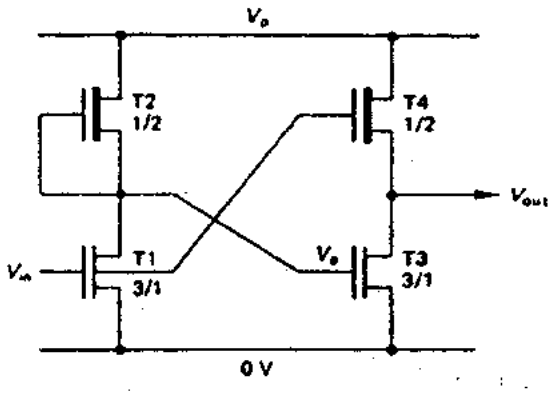


Figure 2.28 NMOS non-inverting superbuffer

The advantage of the superbuffer lies in its action when it switches state. In figure 2.28, when V_{in} is taken high, V_o falls rapidly to 0.3 V, turning T3 off. T4 is on and supplies current to the load capacitance on V_{out}, causing V_{out} to rise. Initially, the gate-source voltage of T4 rises to 5 V and thus T4 can supply more current than the conventional depletion load such as T2 where the gate is connected to the source. Comparing the ratio of the maximum current that can be supplied by T2 and T4 at the commencement of an output voltage rise (see figure 2.29)

$$\frac{I_{ds}\,(T4)}{I_{ds}\,(T2)} = \frac{[(V_{gs} - V_{to})V_{ds} - V_{ds}^2/2] \text{ for T4}}{(V_{gs} - V_{to})^2/2 \text{ for T2}}$$

$$= \frac{(5 + 4)\,5 - 25/2}{16/2} \approx \frac{4}{1}$$

Clearly, this ratio decreases as the source voltage rises. The improved current capability of T4 compared with a conventional depletion load leads to a reduction in the rise time and simulation shows that the edge times for the superbuffer circuit of figure 2.28 are approximately equal.

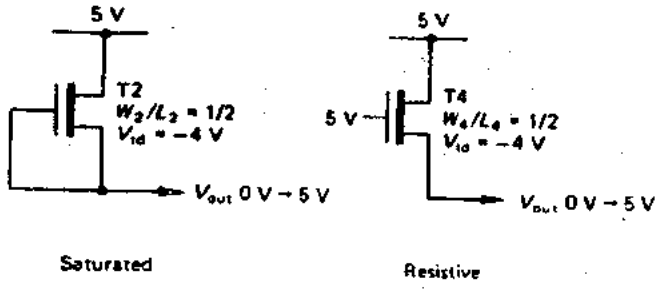


Figure 2.29 Comparing a superbuffer load current with a conventional load current

Other buffer circuits required in a design are the circuits necessary to interface a design to circuitry external to the chip. Here a power and ground pad plus input and output drivers are required. These are normally available as part of a standard library of user functions. Input pads normally consist of a circuit which limits the voltage and current that can be applied to the gate of a transistor. Such protection circuits prevent the breakdown of input transistors arising from large voltages generated by electrostatic charge. Output pads need to drive external circuitry and hence have to be capable of driving a large capacitive load. Thus their circuitry consists of a buffer gate chain, such as that described in this section.

2.17 Further Reading

J. Mavor, M. A. Jack and P. B. Denyer, *Introduction to MOS LSI Design*, Addison-Wesley, 1983.

C. Mead and L. Conway, *Introduction to VLSI Systems*, Addison-Wesley, 1980.

J. T. Wallmark and L. G. Carlstedt, *Field-effect Transistors in Integrated Circuits*, Macmillan, 1974.