Digital integrated circuits with MOS transistors

L. M. van der Steen

The simple structure and low dissipation of MOS transistors allow them to be formed in large numbers per unit surface area on a crystal chip. Moreover, since a MOST has a very high input impedance and allows current to flow in both directions, some MOS circuits require fewer components than the corresponding bipolar circuits. These features make the MOST transistor eminently suitable for use in integrated circuits. In this article we shall describe some examples of digital integrated circuits of this type, made entirely from P-channel enhancement MOSTs.

Little need be said here about the MOST transistor; the N-channel type has been dealt with in detail elsewhere in this issue, and the P-channel type (see fig. 1) differs from this only in polarity. The $I_d-V_{ds}$ family of characteristics and the graph of $I_d$ vs $V_{gs}$ have the same shape as the graphs in fig. 2 of [1], except that currents and voltages are now negative. For application in digital circuits the P-channel MOST may be regarded as a switch that passes current when $V_{gs}$ is sufficiently negative ($|V_{gs}| > |V_{th}|$, the threshold voltage $V_{th}$ being typically -3 to -4 volts) and does not pass current when $V_{gs}$ is between 0 volts and $V_{th}$.

The manufacture of integrated MOS circuits

The voltages applied to a MOST are always arranged so that the drain and inversion layer are reverse-biased with respect to the substrate and therefore isolated from it by a depletion layer (see fig. 1). The source may be connected to the substrate: if it is not, the voltage on the source is always such that this electrode also is surrounded by a depletion layer. The natural isolation provided by this depletion layer presents considerable advantages in the manufacture of integrated circuits made up from MOS devices, since there is now no need to make an isolated island of the appropriate material by means of an epitaxially grown silicon layer and an isolation diffusion for each transistor, as there is with integrated circuits using bipolar transistors [2]. The number of photoetching processes needed for manufacture is thus reduced from six for bipolar-transistor circuits to four for MOS circuits.

The situations after the various photoetching processes [3] are shown schematically in fig. 2. The manufacture of integrated MOS circuits

---

Fig. 1. P-channel MOS transistor on a substrate of N-type silicon. The source S and the drain D are P'-type regions diffused in the silicon. The gate G is of aluminium (shown black) and is isolated from the substrate by a layer of silicon dioxide (white). Substrate and source are earthed; a negative voltage is applied to the drain D. If a sufficiently negative voltage is applied to the gate (lower than the threshold voltage $V_{th}$, which is also negative), a thin inverted P-type layer is formed beneath the gate, enabling current to flow between S and D. Below the source, drain and the channel there is a depletion layer, which acts as an insulator since it contains hardly any free charge carriers.

Fig. 2. Some stages in the manufacture of a MOS transistor. Windows for the source and drain are etched in an oxide layer (O) on the substrate (a). After diffusion of these electrodes, a window for the oxide layer below the gate is etched in the oxide film formed during the diffusion process (b). After the oxide layer has been formed, the windows are etched for the contacts with the source and drain (c). Finally an aluminium layer is deposited, from which the electrodes are formed by etching away surplus aluminium (d).

---

facture of a P-channel MOS circuit starts with a single-crystal wafer of homogeneously doped N-type silicon (of diameter say 50 mm and thickness 250 μm), on which a silicon-dioxide film is formed in a hot oxygen-rich atmosphere. A photoetching process is used to make windows in this oxide film at the places where the source and drain are to be located (fig. 2a). The wafer is then heated in a boron-containing atmosphere, so that the boron diffuses through the windows into the N-type silicon and causes regions of P-type silicon to form. During the diffusion process a new layer of silicon dioxide forms on the wafer, and windows are made in this layer in a second etching process (fig. 2b).

By heating in an oxygen-rich atmosphere a thin layer of silicon dioxide (0.1 μm) is then formed in these windows to give isolation between the substrate and the gate electrodes. In a third etching process windows are made in this oxide layer above the source and drain (fig. 2c). A layer of aluminium is then deposited, from which the various electrodes and connections between the components of the circuit are formed by removing the surplus aluminium in a fourth and last etching process.

Since in this technology there is no isolation diffusion, which takes up a good deal of space, a very large number of components can be formed on a silicon chip. And since the dissipation of MOS transistors is very low, such a high packing density is a practical proposition for many circuits. In a dynamic shift register, for example, which is discussed later on in this article, the local packing densities are as high as 6 x 10^4 components per cm^2, and the average packing density for the complete monolithic circuit is 2 x 10^4 cm^2. A monocrystalline silicon chip measuring a few mm^2 can now accommodate complete MOS circuits of 300 to 1500 components; in the near future monolithic circuits with several thousand components will be possible. At the present, the average packing density that can be achieved in integrated circuits made with bipolar transistors is generally much smaller: about a few thousand per cm^2.

**MOS logic circuits**

Simple logic circuits such as NOR gates or NAND gates are generally used as the "building blocks" for digital circuits, since all digital circuits, from simple bistable circuits to storage elements, shift registers, etc., can be made up from such basic elements [44]. A gate circuit is very easily made with MOS transistors: fig. 3a shows such a circuit, which has four MOS transistors Tr1 to Tr4. Briefly, the circuit operates as follows. The drain and gate of Tr4 are both connected to the negative supply voltage, so that Tr4 always operates in the saturation region. However, current can only flow through

![Diagram](image-url)

Fig. 3. a) The NOR gate, made as an integrated MOS circuit. The supply voltage Vdd is negative. The voltages at the inputs A, B and C can be either zero or negative. The voltage at the output Q is at the zero level when one or more of the inputs are negative, since there is then current in the circuit. When negative logic is used (the negative voltage is then related to the state 1 and the zero level to state 0), the relation between the output level and the input levels is given by Q = A + B + C, which is a NOR function. b) The truth table of the circuit with negative logic.

Tr4 provided one or more of the transistors Tr1 to Tr3 are passing current, i.e. provided the voltage at one or more of the inputs A, B and C is sufficiently negative. The circuit is designed in such a way that in this situation the voltage at the output Q is at the "zero level", i.e. between 0 volts and Vdd. If the voltage at each of the three inputs is at the zero level, no current flows and the output voltage is negative. The transistors Tr1, Tr2 and Tr3 in this circuit thus behave as switches, and so are called switching transistors; whereas Tr4 acts as a resistor and is therefore called a load transistor.

The voltage levels can be related in two ways to the states 0 and 1 used in the mathematical treatment of logical operations by Boolean algebra. In positive logic state 1 denotes the high level and state 0 the low, and vice versa in negative logic. It is the practice to use negative logic for logic circuits built up from P-channel MOS transistors; A = 1 here therefore means that the voltage at point A is negative, and A = 0 means that this voltage is at the zero level. In the circuit given in fig. 3a we then have: Q = 1 if A = 0, B = 0 and C = 0; Q = 0 if one or more of the inputs A, B and C are 1. Thus, Q is not 1 if A or B or C is 1; this NOT-OR relation is represented by the NOR function Q = A + B + C, so that the circuit of fig. 3a, using negative logic, is a NOR gate. Fig. 3b gives the truth table for this case. In the transition from negative to positive logic AND and OR gates change their name, a NOR gate then becoming a NAND gate. This can be seen from the table in fig. 3b by interchanging ones and zeros; Q is then zero (i.e. not 1) only if A and B

and $C$ are 1, so that $Q$ is then indeed equal to $\overline{A.B.C.}$.

The switching transistors, which are in parallel in fig. 3a, can also be arranged in series; the result is then a gate circuit which, with negative logic, acts as a NAND gate. If the output voltage in a current-conducting circuit is to remain within the zero level, the total resistance of the switching transistors must not exceed a particular value. This requirement presents no difficulties if the switching transistors are in parallel; if they are in series, however, an increasing number of inputs requires the use of larger and larger switching transistors, and the circuit then takes up a lot of space on the crystal chip. The circuit with parallel transistors is therefore preferred.

**Inverter circuit**

To describe the behaviour of the gate circuit during the transition between the levels, we can take the simplest situation, for which $B = 0$ and $C = 0$. $A$ can be 0 or 1, and $Q$ is then 1 or 0 as the case may be; $Tr1$ and $Tr4$ then constitute an inverter circuit. We shall now analyse this circuit, referring to fig. 4, and from this analysis we shall derive the requirements that have to be met in the design of the transistors.

Provided that there is no saturation the current in a MOS transistor is related to the applied voltage $V_{ds}$ and the gate voltage $V_{gs}$ by:

$$I_d = \beta(V_{gs} - V_{th} - \frac{1}{2}V_{ds})V_{ds} \ldots (1)$$

Here the threshold voltage $V_{th}$ is the voltage between gate and source at which inversion starts, i.e. at which current begins to flow. If $|V_{ds}| \geq |V_{gs} - V_{th}|$, saturation occurs and the current is given by:

$$I_{ds sat} = \frac{1}{2}\beta(V_{gs} - V_{th})^2. \ldots (2)$$

In these equations $\beta = \mu \cdot \frac{C_{ox}}{w} l$, where $\mu$ is the mobility of the holes in the channel, $C_{ox}$ the oxide capacitance per unit surface area (i.e. of the capacitor formed by gate, oxide and substrate), $w$ is the width and $l$ the length of the channel. (See equations (6) and (9) of the article mentioned in [1].)

For transistor $Tr1$ in fig. 4 we have $V_{gs} = V_i$ and $V_{ds} = V_o$, so that:

$$I_{d1} = \beta_1(V_i - V_{th1} - \frac{1}{2}V_o)V_o \ldots (3)$$

if $|V_i - V_{th1}| > |V_o|$ (no saturation), and

$$I_{d1 sat} = \frac{1}{2}\beta_1(V_i - V_{th1})^2 \ldots (4)$$

if $|V_i - V_{th1}| \leq |V_o|$ (saturation).

Since the gate of $Tr4$ is connected to the drain, so that $|V_{ds}| > |V_{gs} - V_{th4}|$ at all times, this transistor will always operate in the saturation region. In fact $Tr4$ acts here only as a resistor; a transistor is used at this position because a diffused resistor of high enough value would take up too much space on the crystal wafer. For $Tr4$ we have $V_{gs} = V_{th4} = V_{dd} - V_o$ (see fig. 4), so that:

$$I_{d4 sat} = \frac{1}{2}\beta_4(V_{dd} - V_o - V_{th4})^2. \ldots (5)$$

Fig. 5 shows the $I-o$ characteristics of $Tr1$; these have $V_i$ as parameter and at $V_i = V_{th1} = V_o$ they all meet the horizontal part of the curve where the current is saturated. The figure also shows the curved load line which is obtained because $Tr4$ is used as a load resistance and is derived from equation (5). Since at all times

$$I_{d1} = I_{d4}, \ldots \ldots \ldots (6)$$

the points where the load line intersects the $I-o$ characteristics of $Tr1$ are the operating points of the circuit at different values of $V_i$; these operating points set out in a $V_o-V_i$ diagram form the $V_o-V_i$ characteristic of the inverter circuit (fig. 6).

In the characteristic given in fig. 6 three regions can be distinguished. In region 1, $V_i$ is at the zero level, i.e. $0 < |V_i| < |V_{th1}|$. $Tr1$ is now not conducting, so that $I_{d1} = I_{d4} = 0$, and from equation (5) it follows that $V_o = V_{dd} - V_{th4}$. This is therefore the "1 level" of the output voltage. In fig. 6 this situation
corresponds to the horizontal part of the characteristic; in fig. 5 the operating point lies on the \( V_o \)-axis.

In region II \( V_i \) falls below the threshold voltage, so that \( T_1 \) starts to conduct. At the same time, however, \(|V_i - V_{th1}| < |V_o|\), and therefore in fig. 5 the operating point lies to the right of the line \( V_i - V_{th1} = V_o \), so that \( T_1 \) operates in saturation. In fig. 6 a linear transition now arises between the two voltage levels; the region is bounded here by the lines \( V_i = V_{th1} \) and \( V_i - V_{th1} = V_o \).

In region III \( V_i \) is even more negative, and here \(|V_i - V_{th1}| > |V_o|\). The current through the circuit has further increased, and \( T_1 \) is now no longer in saturation. In this situation \( V_i = 1 \) and hence \( V_o = 0 \). To ensure that this is so, the voltage divider formed by \( T_2 \) and \( T_4 \) must be so arranged that \( V_o \) is between 0 and \( V_{th1} \) when the circuit is passing current (as a rule \( V_o \) is put at approximately 0.1 \( V_{th1} \)). Region III lies to the right of the line \( V_o = V_i - V_{th1} \) in fig. 6 and to the left of it in fig. 5.

We shall now take a closer look at the situation in the transition region II, in which both transistors operate in saturation. From equations (4), (5) and (6) it then follows that:

\[
\frac{1}{2} \beta_1 (V_i - V_{th1})^2 = \frac{1}{2} \beta_4 (V_{dd} - V_o - V_{th1})^2
\]

and this gives:

\[
V_o = - (\beta_1/\beta_4)^{1/2} (V_i - V_{th1}) + (V_{dd} - V_{th1}). \tag{8}
\]

In this region the \( V_o-V_i \) characteristic is therefore a straight line with a slope of \( \alpha = - (\beta_1/\beta_4)^{1/2} \). Fig. 7 shows a set of these characteristics for various values of \( \alpha \) and for the same values of \( V_{dd}, V_{th1} \) and \( V_{th4} \). In designing a circuit the choice of \( \alpha \) is partly determined by the requirement noted above that in region III the output voltage should be at the zero level. This means that the slope must exceed a particular limiting value; in practice it is usual to take \( \beta_1/\beta_4 \gg 20 \).

Since the transistors are parts of a monolithic circuit, and are thus formed during the same operation, they have the same thickness of oxide layer below the gates and the same mobility for the holes in the channels. In \( \beta_1 \) and \( \beta_4 \) the factors \( \mu \) and \( C_{ox} \) are therefore identical, so that

\[
\frac{\beta_1}{\beta_4} = \frac{w_1/l_1}{w_4/l_4}.
\]

The ratio \( \beta_1/\beta_4 \) is therefore determined entirely by the \( w/l \) aspect ratios of the transistors. If, for example, the switching transistor has the dimensions \( w = l = 10 \) \( \mu \)m (so that \( w_1/l_1 = 1 \)), and if \( \beta_1/\beta_4 \) should have a value of 20 to give the right slope, then \( w_4/l_4 \) for the load transistor should have a value of 1/20. Thus we could have \( w = 10 \) \( \mu \)m and \( l = 200 \)\( \mu \)m.

The characteristics of figs. 6 and 7 hold for the NOR circuit of fig. 3a when only one switching transistor is conducting. If several switching transistors are activated simultaneously, \( V_o \) adjusts itself to a still lower value; the NOR circuit will operate correctly in any situation if each of the switching transistors has the correct value of \( \beta \) to suit the load transistor.

If the load transistor were to be replaced by a diffused resistor, a widely-used integrated-circuit element, the surface area of this resistor would be much larger than that of the load transistor. The resistance between source and drain of a MOS switching transistor of minimum dimensions \( w = l = 10 \) \( \mu \)m is several tens of \( k\Omega \). This means that a load resistance of several hundred
kΩ should be used to give the correct voltage division. A diffused resistor of this value would be about 10 mm long and 10 μm wide, whereas the corresponding load transistor with the same width is only 0.2 mm long. Although the load transistor has the advantage of being small it has the disadvantage that the full supply voltage can never be obtained at the output of the circuit, since the threshold voltage of the load transistor is always lost. Another disadvantage is that the response of the circuit with the load transistor is slower than when a diffused resistor is used. However, the most important feature is the extra space gained on the chip.

The maximum switching frequency

The maximum permissible switching frequency of a NOR circuit is determined by the speed at which the output voltage switches over, and hence by the rate at which the capacitor connected to this output is charged up. Of course, we must consider the worst case here, i.e. the transition from 0 to 1, at which the switching transistors stop conducting and the output capacitor is therefore only charged up by the load transistor.

In a logic circuit the load on the output of a NOR will usually be several inputs of other NOR circuits. Here we shall take the simplest case where the only load on the NOR is a single input of another NOR (e.g. input A in fig. 3a), so that the load is determined mainly by the capacitance Ca of this input. Then the charging current Iq is equal to \( \frac{\beta}{2} (V_{dd} - V_d - V_{th})^2 \), and is therefore also equal to \( C_d dV_{o}/dt \). From the differential equation obtained by equating these two expressions it can be shown that the RC time constant is proportional to \( C_d/\beta \), and hence to \( \alpha^{2/3}/\mu \).

Since the value of \( \alpha \) is already fixed, we must make \( I_q \) small to be able to reach a high switching frequency. The channels in the switching transistors must therefore be as short as possible and \( \mu \) must be as large as possible.

A much more unfavourable situation is found when the output of a NOR gate is not connected to another NOR circuit but to a point outside the integrated circuit. The load capacitance is then usually determined mainly by the printed wiring of the board on which the integrated circuit is mounted. This wiring capacitance may be several times larger than the input capacitance of a NOR (a few tens of pF against a few tens of fF; 1 fF = \( 10^{-12} \) F), and consequently the switching frequency for the logic circuit as a whole would be 1000 times lower. Measures therefore have to be taken to keep the RC product as small as possible. This is done by using larger load transistors for the final NOR circuit connected to this high capacitance, which make the current 100 times greater than in the earlier circuits. The resistance of the load transistor in this circuit is then only 1/100 of the original value, so that the RC product at the output is now not 1000 times larger but only 10 times. The switching transistors in the last stage must of course be made larger in the same way, and this in turn gives a 100 times greater capacitive load for the penultimate stage. To maintain a reasonable RC time constant the current through this stage must therefore be increased as well, say by a factor of 10, so that larger transistors are also necessary here. The output stages of a circuit thus take up a relatively large amount of space on the chip; in fact it may happen that no transistors at all of minimum dimensions can be used in a small circuit.

MOS shift registers

It can be seen from the above that MOS transistors are most suitable for use in integrated circuits with a small number of outputs. This is the case, for example, in stores with a fixed information content (read-only stores), in "scratch-pad" stores and in shift registers. A shift register consists of a sequence of bistable circuits, or cells whose information content, i.e. the state of the individual cells, moves up one place when a shift pulse or a combination of pulses is applied. Thus, information supplied to the input of the first cell appears, after a number of shift pulses, at the output of the last cell of the register. Shift registers are widely used as storage elements in the arithmetic unit of a computer.

A distinction is made between static and dynamic shift registers. In static shift registers the information content of the register is stored for an unlimited time after a shift pulse. In dynamic shift registers the information content is soon lost and therefore has to be repeatedly replenished. The shift pulses must consequently be repeated at a particular minimum frequency, and this means that the information in a dynamic shift register can never be kept in the same place. Another disadvantage of dynamic shift registers is that they require a more complex combination of shift pulses than a static type. On the other hand, a dynamic shift register is a great deal faster than a static one, and its dissipation is also much lower. Moreover, in integrated-circuit form a dynamic shift register can have a greater packing density than a static shift register.

The static shift register

Fig. 8a gives the diagram of a static shift-register cell built with MOS transistors. The cell consists of a bistable circuit (or flip-flop) formed by two inverters \( Tr_1-Tr_2 \) and \( Tr_5-Tr_6 \) coupled to form a loop by means of \( Tr_7 \) and \( Tr_8 \). The capacitors \( C_1 \) and \( C_2 \), which play an essential part in the operation of the cell, are formed by the capacitances between the gates of \( Tr_1 \) and \( Tr_3 \) and the substrate, and by some stray capacitances. The
output voltage \( V_0 \) represents the information content of the cell. States 0 and 1 correspond, as in the inverter circuit, to the zero level (\( V_0 \) between 0 volts and \( V_{ih} \)) and the negative level (\( V_0 = V_{dd} - V_{ih} \)). The information in the preceding cell of the register is presented as a voltage \( V_2 \) at the input of the cell. The gate signals \( \Phi_1, \Phi_2 \) and \( \Phi_3 \), consisting of shift-pulse trains (fig. 8b) are also fed to the circuit; the pulses \( \Phi_2 \) and \( \Phi_3 \) differ only in the steepness of their leading edges. Whenever a combination of shift pulses arrives, the information content of the previous cell is taken over; in other words, the voltage \( V_0 \) assumes the level of \( V_1 \).

Fig. 8. a) Diagram of a cell of a static shift register. \( Tr_1-Tr_2 \) and \( Tr_3-Tr_4 \) form two inverter circuits, which are coupled to form a closed loop by means of \( Tr_6 \) and \( Tr_7 \). The level of \( V_0 \) (zero or negative) is regarded as the "state" of the cell; \( V_1 \) indicates the state of the preceding cell. b) The control (gate) signals \( \Phi_1, \Phi_2 \) and \( \Phi_3 \) form a train of shift pulses. The information at the input is transferred to the output by each group of three pulses.

Fig. 9a gives a photograph of a 64-bit static shift register built up from cells of the type shown in fig. 8a. An idea of the way in which the various components are joined together to form a circuit can be obtained from fig. 9b, which illustrates a single cell from the register; this is the cell that can be seen at the upper left of the photograph.

The cell works as follows. In the situation where \( \Phi_1 \) is at zero level (see fig. 8a) and \( \Phi_2 \) and \( \Phi_3 \) are negative, \( Tr_5 \) passes no current while \( Tr_6 \) and \( Tr_7 \) are both conducting. Assuming that the output voltage \( V_0 \) is at the zero level, then the gate of \( Tr_1 \), through \( Tr_6 \), is also at this zero voltage (i.e. \( C_1 \) is not charged). \( Tr_1 \) then passes no current, so that point \( A \) is negative, and this negative voltage appears via \( Tr_7 \) at the gate of \( Tr_3 \) (i.e. \( C_2 \) is negatively charged); \( Tr_3 \) passes current, so that \( V_0 \) is held at the zero level. The two inverters are thus cross-coupled via \( Tr_6 \) and \( Tr_7 \), so that the output voltage remains unaltered.

At the instant when the shift pulse \( \Phi_1 \) arrives the sequence of events is as follows. \( \Phi_2 \) and \( \Phi_3 \) go to the zero level, so that \( Tr_6 \) and \( Tr_7 \) no longer conduct and the cross-coupling is broken. The output voltage, however, is maintained because the capacitor \( C_2 \) is negatively charged and \( Tr_3 \) therefore continues to pass current. Since \( \Phi_1 \) has become negative, \( Tr_3 \) starts to conduct. Assuming that the content of the preceding cell is 1, then \( C_1 \) will become negatively charged via \( Tr_3 \), so that \( Tr_1 \) also starts to conduct. If \( \Phi_1 \) returns to the zero level, \( Tr_3 \) becomes non-conductive again, but the negative voltage on \( C_1 \) remains and \( Tr_1 \) continues to conduct. At the same time \( \Phi_2 \) and \( \Phi_3 \) go negative again, so that \( Tr_6 \) and \( Tr_7 \) again start to conduct. Since the leading edges of these pulses are not equally steep, \( Tr_7 \) starts to conduct somewhat earlier, so that \( C_2 \) discharges first through this transistor and \( Tr_1 \); as a result, \( Tr_3 \) stops conducting and \( V_0 \) becomes negative. Meanwhile \( Tr_6 \) also starts to conduct and the cross-coupling is restored, but now with a negative output voltage, with \( C_1 \) negatively charged and \( C_2 \) discharged. The information can now be stored again for an unlimited time.

The principle of this cell is thus fairly simple. During the shifting process the cross-coupling is broken; the old information is held at the output by the capacitor \( C_2 \), which keeps the inverter \( Tr_3-Tr_4 \) in its former state. The new information is meanwhile applied to \( C_1 \), causing \( Tr_1-Tr_2 \) to assume the new state. Next, \( C_2 \) is raised to the new voltage (via \( Tr_7 \)), so that \( Tr_3-Tr_4 \) also assume the new state, and the cross-coupling is then restored through \( Tr_6 \).

During the transfer of information a capacitive storage is used twice. The capacitor \( C_1 \) need only retain the charge during the leading edge of \( \Phi_3 \), that is to say until the cross-coupling is restored, while \( C_2 \) must hold the charge during the time that \( \Phi_2 \) and \( \Phi_3 \) are at the zero level. Both capacitors can only discharge through the leakage current of the P-N junctions of \( Tr_6 \) and \( Tr_7 \). At room temperature the time constant of this discharge can have a value of a few tens of milliseconds; if we compare this with the conventional pulse duration of 1 to 20 \( \mu \)s, we see that the storage action of the capacitors is amply sufficient.

The maximum permissible frequency of the shift pulses depends on the speed at which the circuit changes state. This speed is determined by the time...
Fig. 9. a) A static shift register (64 bits) using the cell shown in fig. 8a. The circuit contains 458 components and measures 1.15 by 2.3 mm. The blue-grey strips are the scribed lines along which the crystal wafer can be broken to separate the circuits from each other. In the blue regions the upper layer consists of oxide, in the white regions it consists of aluminium.

b) Sketch of a single cell from the shift register. The areas covered with oxide are shaded; the shading is light where there is only substrate material under the oxide, and dark where there is a diffused region below the oxide. The areas where aluminium has been deposited by evaporation are left white, whatever is beneath them. The solid lines in the aluminium indicate boundaries between diffused regions and substrate material; the dashed lines indicate regions with a thin oxide film (these are mostly the isolation layers under the gate electrodes). These boundaries can also be seen on the photograph, since there is a difference in height here. The contact windows are shown cross-hatched; these are the locations where the aluminium layer makes contact with an underlying diffused region of source or drain. The various transistors in the diagram are indicated by their number, placed in the channel.

A number of these cells are placed side by side, and the various signals are applied to the strips of aluminium, which are clearly visible in the photograph. The sources of \( T_{i} \) and \( T_{j} \) are earthed by connecting the central diffused layer with the earthed substrate at a point which lies outside this diagram.

needed to charge \( C_{1} \) or \( C_{2} \). In the transition from state 1 to state 0, described above, \( C_{1} \) must be charged via transistor \( T_{4} \) of the previous circuit. During the transition from 1 to 0, \( C_{2} \) must be charged via \( T_{2} \) and \( T_{7} \). Since \( T_{2} \) and \( T_{7} \) are load transistors, they have a fairly high resistance; however, when they are combined with switching transistors of minimal dimensions (\( C_{1} \) and \( C_{2} \) are then about 50 fF), the \( RC \) time constant for charging the capacitors can still be as small as 50 ns. In the situation illustrated, i.e. the transfer of information between two successive cells in the shift register, this gives a maximum shift-pulse repetition frequency of 1 MHz. The speed of the register as a whole is of course adversely affected by the output stage in the same way as described for the NOR gate; in practice the maximum frequency would be about 250 kHz.
The dynamic shift register

Fig. 10a shows the diagram of a dynamic shift register cell built with MOS transistors. The level of the output voltage \( V_0 \) again indicates the state of the cell, and \( V_i \) the state of the preceding cell. \( C_1 \) and \( C_2 \) are the capacitances of the gates of \( Tr_1 \) and \( Tr_4 \), \( C_1' \) is the capacitance \( C_1 \) of the next cell. There are four shift signals \( \Phi_1, \Phi_2, \Phi_3 \) and \( \Phi_4 \) (fig. 10b), consisting of negative pulse trains; after each group of four shift pulses the input voltage is taken over by the output. In fig. 10a it can be seen that both \( \Phi_1 \) and \( \Phi_3 \) are fed to two points of the circuit. Fig. 11a gives a photograph of a part of a dynamic shift register in which this cell is used; fig. 11b again shows a diagram of one of the cells from this register.

To explain the operation of the cell we start from the situation where \( V_1 \) is negative; \( Tr_1 \) is then conducting. At the moment when the first shift pulses arrive, \( \Phi_1 \) and \( \Phi_3 \) go negative, so that \( Tr_2 \) and \( Tr_3 \) also start to conduct. From the source supplying the voltage \( \Phi_1 \) the capacitor \( C_2 \) will now be charged to a negative voltage via \( Tr_3 \) and the series arrangement of \( Tr_1 \) and \( Tr_2 \). The voltage \( \Phi_1 \) then returns to zero, but \( \Phi_3 \) remains negative. As a result, \( Tr_3 \) stops conducting but \( Tr_1 \) and \( Tr_2 \) continue to pass current, and since the gate of \( Tr_1 \) is now at the zero level, \( C_2 \) discharges through these transistors. When \( \Phi_3 \) also returns to zero, \( C_2 \) is thus nearly discharged. Now when the shift pulses of \( \Phi_3 \) and \( \Phi_4 \) arrive, \( Tr_5 \) and \( Tr_6 \) start to conduct but \( Tr_4 \) remains non-conducting because \( C_2 \) is discharged. \( C_2 \) is now charged (through \( Tr_6 \) only), but when \( \Phi_3 \) returns to zero it can no longer discharge. Thus, when \( \Phi_4 \) also goes back to zero, a negative voltage remains at the output.

If the input voltage is at the zero level, \( C_2 \) cannot discharge because \( Tr_1 \) does not go into conduction. Consequently \( C_2 \) remains negative, which then enables \( C_3 \) to discharge, and the output voltage goes to zero. We see that in both cases the voltage that was at the input is transferred toward the output after four shift pulses. Since four phases can be distinguished in this process, a circuit of this type is known as a four-phase shift register.

The dynamic shift register of fig. 10a is based on the storage action of the capacitors \( C_1 \) and \( C_2 \). These capacitors, however, discharge as a result of the leakage currents in the \( P-N \) junctions of the corresponding transistors, and since the dynamic circuit has no cross-coupling as in the static type, the information would be gradually lost. To prevent this, the loss of charge is regularly compensated by the signals \( \Phi_1 \) and \( \Phi_3 \), and this means that the whole shift-pulse pattern must be repeated at a particular minimum frequency. The register therefore continues to transfer the information.

The minimum frequency at which the shift pulses have to be supplied depends on the magnitude of the capacitances and leakage currents. Since leakage currents of \( P-N \) junctions are involved the repetition frequency is temperature dependent: at room temperature a frequency of 40 or 50 Hz may be high enough, but for every 10 degrees increase in temperature the minimum frequency required increases by a factor of 2.

For the static shift registers MOS transistors of different dimensions are needed (switching and load transistors) in order to obtain the desired voltage division in the inverter circuits. This is not necessary for the dynamic registers, and switching transistors of minimum dimensions can therefore be used. The capacitances \( C_1 \) and \( C_2 \) in fig. 10a have about the same value as those in the static shift register in fig. 8a, but during the transfer process they are charged up through transistors whose resistance is a tenth of that of the load transistors in the static register. Because of this the maximum permissible shift frequency for the dynamic shift register is an order of magnitude higher than for the static type; it is about 10 MHz. Another consequence of only using switching transistors is that the component density on the chip is higher in dynamic registers.

![Diagram of a cell from a dynamic four-phase shift register](image-url)

Fig. 10. a) Diagram of a cell from a dynamic four-phase shift register. The output voltage \( V_0 \) indicates the state of the cell, and \( V_i \) the state of the preceding cell. b) The gate signals \( \Phi_1, \Phi_2, \Phi_3 \) and \( \Phi_4 \) consist of negative pulse trains; the information at the input is transferred to the output by each group of four of these shift pulses.
Fig. 11. a) Part of a dynamic shift register built up with cells of the type shown in fig. 10a. b) Sketch of one cell of the shift register. The various areas are indicated in the same way as in fig. 9b.

The cells are located in a row on the crystal wafer and the various voltages are applied to the aluminium strips. The photograph shows four such rows of cells one above the other, the strips for the voltages $\phi_1$ and $\phi_2$ being connected to adjacent rows; the transistor $Tr_a$ in the sketch therefore belongs to a cell in a following row.

The dynamic register also dissipates less heat. In the static shift register there is always current flowing through one of the two inverter circuits, and the dissipation per cell is 2 mW. In the dynamic shift register no current can flow direct to earth; energy is only dissipated when the capacitors are charged. The dissipation is therefore substantially lower than in the static case, and moreover it is proportional to the frequency of the shift pulses. The dissipation per cell is 50 $\mu$W/MHz; even at a pulse frequency of 5 MHz the dissipation per cell is therefore no more than 250 $\mu$W, which is a great deal better than the 2 mW per cell of the static register.

As we noted earlier, there is a penalty to be paid for these advantages: the information cannot be kept in the same place and a more complex combination of shift pulses is required.

Summary. The source and drain in MOS transistors are isolated by a depletion layer from the substrate, so that no isolation diffusions are required. MOSTs can therefore be packed very densely on a crystal wafer. Since mosts also dissipate little heat, they are very suitable for use in integrated circuits. After a brief description of the technology, some examples of such circuits are given. The first is a logic circuit: the NOR gate, in which the MOS transistors are used as switches and also as resistors; the MOST is smaller than a diffused resistor of the same resistance. The other examples, a static and a dynamic shift register, make use of the specific advantages of the MOST: its very high input impedance and the fact that the current through a MOST can flow in both directions. A cell of the static shift register consists of a bistable circuit; the cell of the dynamic shift register is based on the storage action of a capacitor. In the dynamic shift register there is no cross-coupling to prevent the information from being lost, and the voltage on the capacitor therefore has to be continuously replenished by the shift pulses (4 pulses per cycle). This means that the information in the dynamic register must be kept moving through at a specific minimum frequency. Compared with the static type the dynamic register has the two advantages of a maximum pulse frequency that is 10 times higher (10 MHz against 1 MHz), which is possible because only MOSTs of minimum dimensions are used, and a lower dissipation (50 $\mu$W/MHz per cell as against 2 mW per cell).