From Electron Mobility to Logical Structure: A View of Integrated Circuits

WESLEY A. CLARK

Design Consultant, 1572 Massachusetts Avenue, Cambridge, Massachusetts 02138

The microprocessor's rapid development and its appearance in many new theaters of application are bringing computer scientists and programming specialists increasingly into contact with new and largely unfamiliar products of integrated circuit technology. This tutorial paper introduces some of the main ideas underlying the theory, fabrication, and operation of integrated circuits and shows how logically useful functions are derived from the balancing and unbalancing of electric charge in tiny semiconductor-based structures.

Keywords and Phrases: integrated circuits, semiconductor junctions, transistor models, MOS fabrication, logic circuits, microprocessor structures, silicon systems

CR Categories: 6.1, 6.31

INTRODUCTION

Anyone who has used a plastic pocket comb on a cold dry day has at least a basic knowledge of electricity: Electric charge can be moved from place to place, and there is a force between two charged objects that increases as the distance separating the objects is decreased. On a dry day, evidently, the comb removes and retains some of the electric charge found on the surface of hair. This leaves the hair electrified by virtue of a charge deficit while electrifying the comb with a charge surplus. The balance of nature is disturbed momentarily and some observable forces result. Hairs having a similar charge imbalance repel one another, yet are attracted to the comb, which now has an imbalance of the opposite kind. In the further experience of grasping a metal doorknob after walking across a carpet on that same dry day, one is memorably persuaded that metallic objects are certainly excellent charge conductors, and that the human body itself is a fair conductor of electricity.

Electronic computers operate, of course, by coordinating the conduction of charge from place to place along numerous pathways, now opening a gate here or closing a gate there to regulate the flow, now charging or discharging component parts, now merging, amplifying, and so on—always heating up the surroundings and occasionally producing observable forces useful to man.

In today's technology the workhorse, or perhaps demon, of this busy enterprise is the transistor, a carefully made microscopic device that is inseparably integrated with many of its kind into inexpensive electrical circuit modules. These integrated circuits, from simple gates and registers to complete microprocessors, are available in many different forms varying in complexity, utility, and cost. Computer scientists and programming specialists, increasingly active participants in the extension of microprocessor systems into new areas of application, are therefore confronted by a sizable array of possibilities for contemplation [ASPI78].

The intent of this paper is to offer the reader, presumed conversant in the idiom

Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, and notice is given that copying is by permission of the Association for Computing Machinery. To copy otherwise, or to republish, requires a fee and/or specific permission.
© 1980 ACM 0010-4892/80/0900-0325 $00.75

CONTENTS

INTRODUCTION
1. THE DESIGN OF INTEGRATED TRANSISTOR CIRCUITS
  1.1 Physical Properties of Solids
  1.2 Semiconductor Devices
  1.3 MOS Integration
2. THE LOGIC OF MOS CIRCUITS
  2.1 Basic Combinations
  2.2 Logic Arrays
  2.3 Microprocessors
CONCLUSION AND PERSPECTIVE
ACKNOWLEDGMENTS
REFERENCES

of computer science but possessed of limited enthusiasm for the contemplation of its engines, a brief and correspondingly limited glimpse of the "high" technology of large-scale integrated circuits. What is inside an integrated circuit package? How is it made? How are electrons put into logical harness on tiny chips of silicon? These are some of the questions to which at least intuitively satisfactory answers are sought.

Only a basic knowledge of the behavior of electrically charged objects is asked of the reader as we progress from a review of a few principles of theory to a lively explanation of how simple transistors work, then to a description of an important fabrication process, and finally to an examination of some of the logically useful forms from which functionally complex structures are synthesized. The road from electron to logic is well paved and well marked, and those who make the present journey are promised a minimum of Dielectrical Materialism along the way.

1. THE DESIGN OF INTEGRATED TRANSISTOR CIRCUITS

The principles underlying the design and operation of integrated circuits with their myriad transistor devices are easily grasped if one is prepared to begin with a short excursion into the domain of atoms and molecules, from which all of the known properties of matter arise. Since our objective is an understanding of computer *hardware*, we can limit our interest to the properties of appropriate solid objects.

1.1 Physical Properties of Solids

In the atomic domain the attraction of unlike charges holding an atom together is identified as an attraction between a massive positively charged nucleus and a number of much less massive negatively charged electrons, which surround the nucleus and balance the whole thing electrically. When atoms combine to form molecules, crystals, or other matter, the nature of the interatomic attraction is somewhat more complicated. Apparently, each atom has preferred groupings of electrons, loosely thought of as concentric shells, each capable of holding so many electrons and no more. In balancing the nuclear charge, electrons fill inner shells first and become inaccessible but may or may not complete the outermost shell. Molecules, crystals, or other solids form when two or more atoms having incomplete outer shells come just close enough to share their accessible outermost or *valence* electrons with one another in an attempt to complete all shells, locking together by establishing *covalent bonds*. Full removal of electrons from an atom leaves the atom with a net positive charge, of course, while addition of electrons leaves it with a net negative charge. Such electrically unbalanced atoms, called *ions*, also have a role to play in repulsion and attraction.

So much we know from basic physics and chemistry. Now to the substance of the matter.

1.1.1 Insulators, Conductors, and Semiconductors

With regard to electrical substance, solid matter under ordinary conditions can be classified as insulators (dielectrics), conductors, or something in between called semiconductors. In *insulators*, such as combs, carpets, and hair, electrons are so tenaciously bound to atoms that they can be torn free only by relatively large forces, and at that, generally only from the surface of the solid. Electrons within the solid cannot move about in an attempt to compensate such a loss, and the surface of a "stripped" insulator thus remains charged at the fixed

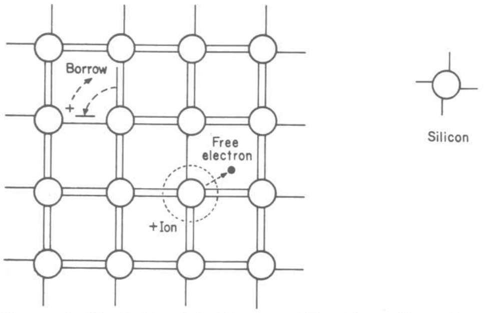


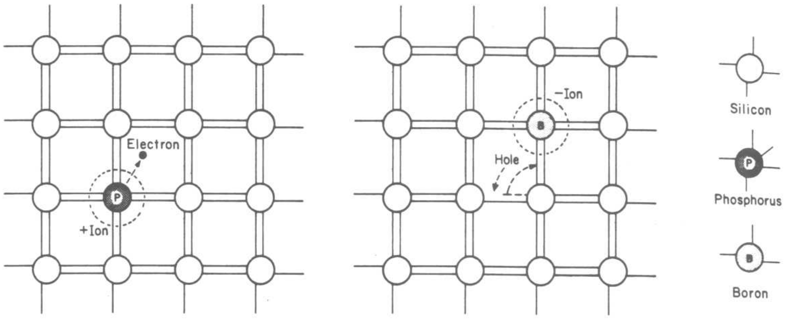
FIGURE 1. Covalent bonds in silicon crystal; breaking and borrowing.

ionized sites created by the loss until new electrons are reintroduced from without.

In *conductors*, on the other hand, valence electrons are quite easily separated from their atoms under the least provocation and travel freely throughout the conductor in all directions at high speed, frequently colliding or otherwise interacting with the atoms fixed within the solid but generally keeping their distance from one another. If extra electrons are somehow crammed into a conductor having no escape pathway, they quickly distribute themselves throughout in the manner of a diffusing gas. But with an escape pathway provided by a closed electrical circuit, electrons injected into one end of a conductor almost immediately push electrons out of the other, and are themselves crowded along and finally out unless the injection stops or the circuit is opened. The *mobility* of the electrons depends on properties of the conductor; the *number* of electrons pushed along per second depends on how hard they are being pushed and how crowded they become—factors measured by the gradient of population pressure from the injection end to the escape end, the cross-sectional area of the conductive corridor, and finally by the electron mobility.

Semiconductors are yet another matter. Certain elements—among them silicon, the one to be discussed here—have valence shells that are half full. Silicon has four valence electrons and readily bonds covalently with four neighbors to form a solid lattice in which interior atoms have completed shells (Figure 1). With the addition of a bit of energy (for example, the thermal agitation energy of free electrons in a comfortably heated room) it will be found that an extremely tiny fraction of these bonds are always breaking apart here and there. The electrons thus released are free to jitter about in search of other bond-deficient ionized sites to fill. But it is vastly easier for a bond-deficient pair simply to borrow an electron from a neighbor, which then borrows from *its* neighbor and so on, deionizing one site at the expense of another.

Semiconductor specialists like to refer to bond deficiencies as *holes.* They observe that these "holes" can be said to wander about and to repel one another just as electrons do, each hole momentarily defining a positive ion which inhibits the intrusion of other holes into its vicinity. In practical silicon crystals the mobility of such holes is about one third that of free electrons, but charge can be "carried" by either or both. Inasmuch as these charge carriers are so very few in number, silicon crystal will conduct only weakly and is therefore said to be a *semi*conductor. Since the breaking of any bond gives rise to exactly one hole and one free electron while the balancing opposite process involves the capture of one free electron by one hole (taking both temporarily out of the action), the number of holes is always equal to the number of free electrons in pure silicon crystals. There appears to be no such thing as completely



(a) (b)

FIGURE 2. Enrichment of silicon in (a) electrons and (b) holes.

pure crystalline silicon (or anything else, for that matter), but then, as pure crystals, semiconductors seem to have no electrical properties of great utility.

Semiconductor crystals can, however, be enriched either in holes or in free electrons by somehow embedding within the lattice just a pinch of atoms of certain other elements. It is this fact that makes it possible to build useful devices.

1.1.2 Charge-Carrier Enrichment of Semiconductors

An atom of phosphorus has five valence electrons; an atom of boron, three. Atoms of either kind can be locked into the silicon lattice, and when they take up residence at even a minute fraction of sites their contribution can become dominant. The awkward fifth valence electron of each phosphorus atom is promptly freed (Figure 2a), while each boron atom's covalence deficit is no less promptly covered by a neighborly silicon atom (Figure 2b), which takes a hole in exchange and passes it on. With enough tenants per crystal, the number of free electrons or holes in general circulation can be increased a millionfold.

This process of substituting other atoms for some of the semiconductor atoms is called *doping*. Semiconductors doped with electron donors such as phosphorus are said to be of n-type, while boron doping, which results in extra holes, produces semiconductors of p-type. (The letters *n* and *p* stand for negative and positive; physicists had already appropriated *e* and *h* for other worthy purposes.) Doping elements, though they give semiconductors desirable characteristics, are referred to as impurities—rather like calling orchids weeds if they are found growing in a cabbage patch. Doping of silicon is easily accomplished by adding just the right amount of the doping element to molten silicon and allowing the result to cool and crystallize, or by diffusing the dopant as a vapor through the surface of the crystalline solid at high temperature, when everything is vibrating so much in all directions that the intruders can find accommodations in minor lattice defects without greatly upsetting the overall structure. In the shorthand of the specialists, heavily doped material is referred to as n^+ or p^+—an unfortunate choice perhaps, but there it is. The heavier the doping, the better—that is, more metallic—the conductivity.

The underpinnings of exposition are now all in place and we are ready to move on to a consideration of their practical consequences. The reader who is disquieted by the informality and imprecision of the foregoing discussion, or who is motivated to learn more of the underlying theory, may refer to GROV67 or CARR72 for a complete exposition.

1.2 Semiconductor Devices

1.2.1 The Semiconductor Junction: CHARGE WARS

Consider, for example, what happens when blocks of n- and p-type semiconductors are joined (Figure 3). If it is a perfect union,

Univ. of Washington Libraries

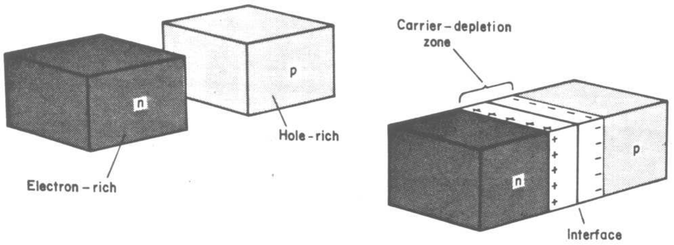


FIGURE 3. Charge-carrier depletion at the semiconductor junction.

electrons and holes will immediately begin to swarm across the interface to annihilate one another. This war would spread throughout the entire composite crystal were it not for the fact that as they enter the fray at the interface, electrons from the n-region leave behind them a thickening barrier of immobile phosphorus ions, whose positive charges are no longer compensated by the departed free electrons. Similarly, holes entering from the p-region leave behind a thickening barrier of uncompensated negative boron ions. These defensive walls quickly become so strong that they cannot be breached in great numbers without extraordinary force; the negatively charged wall on the p-side repels the invading negative electrons, while the positively charged wall on the n-side repels the counterinvading positive holes.

There the matter sits. The battle neatly confines itself to a narrow zone in which the charge carriers are quite depleted and the rate of recruitment just balances the rate of desertion on either side. If one side is more heavily doped than the other, it prevails only to the extent of shifting the battle zone toward, and in the limit just inside, enemy territory.

Fresh troops can be sent in from abroad (by a battery, for example)—electrons into the n-region and holes into the p-region (Figure 4b). After some attrition through skirmishes along the way with oppositely charged "guerillas," which constitute a very small minority in each region, most of these suicidal charge carriers reach the front where they too are annihilated. Now since a backward flow of holes is equivalent to a forward flow of electrons, there is a net flow of negative charge in the n-to-p direction. The increase in population pressure narrows the depletion zone.

On the other hand, an attempt to withdraw troops leads only to a strengthening of the walls because of the further decompensation of fixed ions (Figure 4c). Therefore only a tiny negative charge will flow in the p-to-n direction. We will ignore it.

This remarkable one-way-flow property of semiconductor junctions appears in the operation of practical devices in essentially two different ways: first, under suitable external conditions, to *isolate* one semiconductor region from another region of opposite majority-carrier polarity, thus permitting charge to flow within the one without escaping into the other; second, on the basis of external conditions, to *modulate* the flow of charge between two regions of opposite polarity. Devices of the first kind are said to be *unipolar*; devices of the second kind, *bipolar*. Conduction through a semiconductor junction is a bipolar process.

Devices of either kind can be made in great batches by using the high-temperature vapor diffusion method of doping. Indeed it is the efficiency of batch fabrication that gives us the inexpensive products characteristic of semiconductor technology. Diffused regions are created near the surface of a thin substrate of predoped silicon; windows cut into a covering layer of silicon dioxide permit entry of the vapor (Figure 5a). Phosphorus vapor, for example, will then form a bounded n-region in a boron-doped substrate (Figure 5b) when introduced in a quantity sufficient to overwhelm the contribution of the boron ions already in residence. A few crafty holes will survive

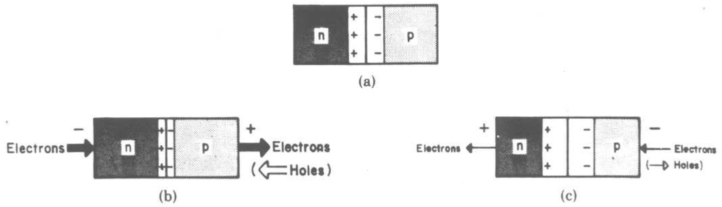

n
+
−
p
(a)
Electrons
−
n
p
+
Electrons
(Holes)
(b)
Electrons
+
n
+
−
p
−
Electrons
Holes
(c)

FIGURE 4. Conduction through a semiconductor junction: (a) at rest; (b) with majority-carrier injection; (c) with majority-carrier removal (charge flow negligible in practice).

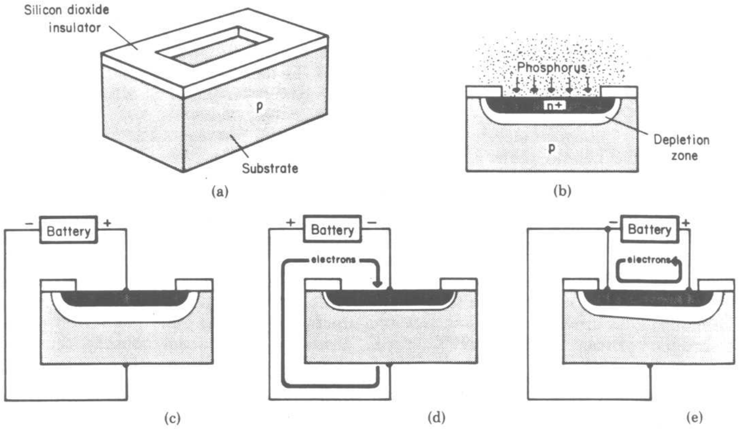

Silicon dioxide insulator
p
Substrate
(a)
Phosphorus
n+
Depletion zone
p
(b)
Battery
−
+
(c)
Battery
+
−
electrons
(d)
Battery
−
+
electrons
(e)

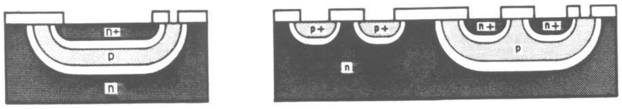
FIGURE 5. The diffused junction: (a) substrate and diffusion window; (b) formation; (c) and (d) bipolar behavior; (e) unipolar behavior.

to become guerillas. Silicon dioxide, by the way, is an excellent insulator.

Additional regions of differing polarities can be formed by changing the diffusion times and diffusing vapor compositions, thus creating more complex layered structures. In practice, substrates are roughly one hundred times thicker than the diffused regions, and therefore the resulting structures can fairly be characterized as *planar*. Two of these more complex planar structures are shown in Figure 6, the diffusion layer structures of TTL (transistor-transistor logic—Figure 6a) and CMOS (complementary metal-oxide-semiconductor—Figure 6b) transistors.

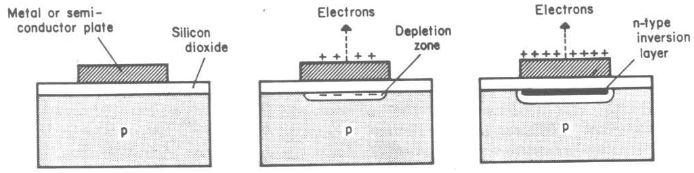
Before we discuss transistor devices, the reader is invited to contemplate one last form of semiconductor junction, a form at once charming in simplicity and extraordinary in utility.

The *field-induced junction* provides the basic unipolar control mechanism of all MOS integrated circuits. It arises in the following way. A semiconductor substrate (p-type in the illustration) is covered with an insulating layer of silicon dioxide—actually *glass*, though in the degree of purity required we know it as quartz. On this is laid a conductive plate (Figure 7a). An electrical force field is now created between substrate and plate by connecting each



(a) (b)

FIGURE 6. Diffusion layers in (a) TTL and (b) CMOS transistors.



(a) (b) (c)

FIGURE 7. Field-induced junction: (a) rest; (b) onset; (c) junction.

of them to opposite poles of a battery, plate-to-plus. Some of the plate's free electrons are conducted away to the battery, leaving the plate with a net positive charge (Figure 7b).

In the substrate beneath the glass, a number of subtending holes are induced to retreat just far enough for their abandoned ions to form a counterbalancing negative image of the plate, thus creating a carrier-depleted zone at the glass boundary. Nearby negative guerillas begin to cluster as well, although their very small minority position at the outset precludes much of a contribution to the formation of the counterimage. But as the battery pressure increases to a threshold value or beyond, almost all of the holes are pushed away from the boundary, and the guerillas suddenly find themselves in the majority within a very thin boundary layer of semiconductor material. The layer is thus, for the moment, inverted from p-type to n-type (Figure 7c). (Although this example is based on the use of a p-substrate, the expected complementary junction would have been formed had we made the appropriate reversal of all charge-mechanism polarities, a fact generally true throughout our discussions.) If the battery is now disconnected, the plate will become electrically isolated and its trapped positive charge will maintain the junction. But if the pressure of a connected battery falls to a subthreshold level, or if electrons are allowed to return to the plate by some other route, the field-induced junction will collapse and disappear.

Very, very nice. The existence of junctions can be controlled dynamically, and there is even a hint of possible memory applications. We have only to discover the means.

1.2.2 TTL and MOS Transistors

In digital circuits, "transistor" primarily means "switch"—a device that either conducts charge to the best of its ability or not at all, depending on whether it is "on" or "off." At present there are many circuit families, each manufacturer developing his own techniques and nurturing his own family and its market. A partial list would include the TTL, MOS, and CMOS families, as well as the CCD (charge-coupled device), ECL (emitter-coupled logic), and I^2L (integrated injection logic) families. Some of these come in either p or n flavors (CMOS in both at once) and some in both high- and low-power or -speed versions, further lengthening the list. The discussion is limited to TTL and MOS transistors, and at that, TTL is given short shrift in order to get on with an introduction to the simpler, unipolar MOS technology.

The MOS transistor was the first to appear. It was invented by J. E. Lilienfeld about fifty years ago in the course of experimentation with the use of the field effect in semiconductors to achieve charge-current amplification [LILI28]. Although the device appeared to work to some extent, an adequate theoretical basis was not possible until the late 1940s; but it was at about that time that J. Bardeen, W. H. Brattain, and W. Shockley invented the bipolar junction transistor [BARD48, SHOC49] and for more than a decade, attention was diverted to the development of this remarkable device.

Early bipolar transistors were clumsy, unreliable, and expensive, yet very promising. Following the lead of the Bell Laboratories' work, commercial interest expanded with increasing rapidity and useful numbers of devices became available to the circuit designer. Germanium rather than silicon was the basic ingredient of these first transistors, and products were far from perfect. The designer's task involved the combination of many transistors and other circuit components and a great deal of wire in an attempt to correct device deficiencies, often with brilliant success but always with added complexity. With the development of the diffusion methods of silicon doping, planar structures became possible and the extremely small sizes necessary for submicrosecond switching were finally achieved. Today's transistors are marvels of reliability and low cost, representing an enormous investment in both engineering effort and manufacturing skill.

Interestingly enough, when attention finally was focused once again on MOS it was found, contrary to theory, that the practical devices failed to work as well as expected. It then took a great deal of study and many years of frustrating and exacting effort to discover that the culprit was common table salt! Salt, nearly ubiquitous on our planet, had been contaminating the silicon dioxide in almost nonexistent amounts, thereby impairing the ability of the oxide to act as a proper insulator. Toughening up the standards of cleanliness and purity—already at exceptionally high levels—finally solved the problem.

TTL transistors are all of the npn form shown in Figure 8 but appear as functionally distinct units embedded in the surface of a broad underlying substrate of p-type semiconductor material not shown. The unit's three participating regions are named the *emitter*, the *base*, and the *collector*. If a battery *B* is connected to these regions as in Figure 8a, a few charge troops will be withdrawn from both base and collector, but the depletion zones at the emitter–base and base–collector frontiers of the eternal charge wars prevent the flow of currents of significant size along any pathway. But if as shown in Figure 8b we now connect another battery *A*, having a pressure or *voltage* of, say, one-tenth that of *B*, then two different currents begin to flow: holes are introduced into the base by *A*, while electrons are sent into the emitter by both *A* and *B*. The emitter–base battle immediately intensifies; energetic electrons pushed by both batteries crowd across the northern frontier into the base region.

The principal electron casualties are those taken out of the action by some of the holes. But the base region is very thin and most of the electrons, pressed forward by endless incoming ranks, quickly pass through enemy territory unscathed and cross over the southern frontier into the collector, where they join their comrades in a withdrawal to battery *B* for another try. If *A* ceases to operate or is removed from the circuit, the two frontier battles become deadlocked once again and all flow immediately stops.

Thus the flow of a very small current in the *A*-loop modulates the flow of a current many times its size in the *B*-loop. In TTL circuits, the *A*-loop of one transistor generally appears as but one of several branches of the *B*-loop of another, which in turn is "driven" as part of the job of a third transistor, and so on.

The symbol used to represent an npn-transistor is shown in Figure 8c. Note that its arrowhead points upstream, like a boat with a bow anchor, rather than in the direction of electron flow.

MOS transistors are unipolar and simple. Figure 9 shows an n-carrier version and is almost self-explanatory. Two separate n-regions, the *source* and *drain*, are diffused

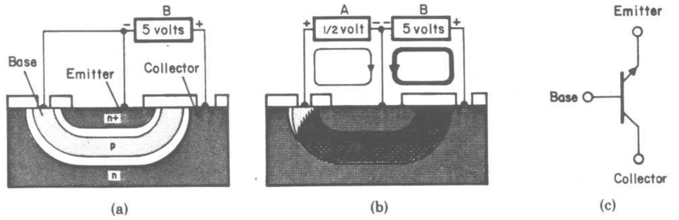


FIGURE 8. The TTL transistor: (a) off; (b) on; (c) symbol.

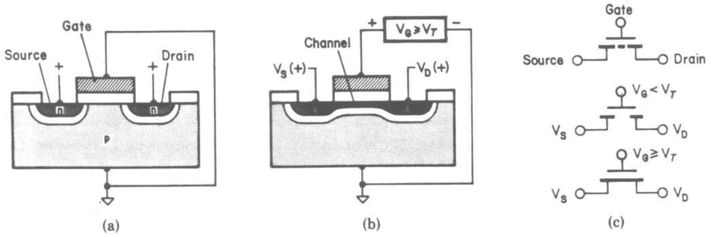


FIGURE 9. The n-channel MOS transistor: (a) off; (b) on; (c) symbols. Voltages: source, V_S; gate, V_G; drain, V_D; threshold, V_T.

into the surface of a p-substrate through windows in a thin covering of silicon dioxide and are bridged by an insulated plate, the *gate*. With substrate and gate both anchored electrically as in Figure 9a, the source and drain will be isolated from the substrate, and thus from one another, by depletion zones—provided that neither is ever forced to become more negative than the anchor when included in external circuits not shown. If we now (Figure 9b) connect a battery having a voltage at least as great as that required to establish a field-induced junction under the gate—about 1 volt—the source and drain will be interconnected by the resulting inversion layer. Electrons will thus have a conductive n-*channel* through which to flow from source to drain when pushed by external means; which is source and which is drain depends only on flow direction.

The channel is not the best of conductors. It is inherently several orders of magnitude more resistive to flow than either the source or drain because it presents a much smaller cross-sectional area to the current and because it supports many fewer charge carriers. Furthermore, as the electrons jostle their way through the channel considerable energy is lost to the crystalline substrate through collisions with its fixed atoms and will appear as unwanted heat. Weakened ranks of electrons emerge from the channel to continue on their way down remaining pathways to an energy-infusing battery or its equivalent, generally through other "resistors" along the way. Long narrow channels are more resistive to flow; short broad ones are less so. There are lower limits to both length and breadth, but no upper limits within the confines of the surface. The general shape of the channel is arbitrary.

1.3 MOS Integration

In 1952 the British engineer, G. W. A. Dummer, made the following remarkable prediction in an invited paper presented at a Washington, D.C., symposium on electronic components (quoted in WOLF76):

> With the advent of the transistor and the work in semiconductors generally, it seems now possible to envisage electronic equipment in a solid block with no connecting wires. The block may consist of layers of insulating, conducting, rectifying and amplifying materials, the electrical functions being connected directly by cutting out areas of the various layers.

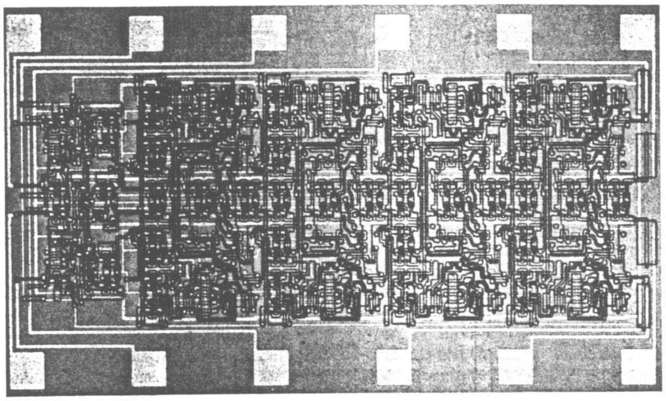


FIGURE 10. Dual-product pipeline multiplier MOS chip. Large squares are aluminum; on the right, broad low-resistance output transistors. [Photograph courtesy of Carver Mead.]

Within a few years work was begun both in England and in the United States in search of Dummer's device. Dummer himself exhibited a model in 1957 that demonstrated how silicon might be used in integration, but the first working monolithic circuit was made by the American J. S. C. Kilby in the following year. The device as we now know it derives from fabrication methods developed by R. S. Noyce. An example is shown in Figure 10, a few-hundred MOS-transistor integrated circuit "chip" designed by E. K. Cheng and C. A. Mead [CHEN75] for pipelined multiplication. It fits on the head of a pin.

1.3.1 The Physical Structure of MOS

The n-channel silicon-gate transistor is currently the preferred form in unipolar integration technology. The name MOS survives an earlier period in which gates were made of metal. Polycrystalline silicon ("poly") is sputtered or otherwise deposited over the surface of oxide covering the substrate. It is then etched selectively to form gates and useful pathways. As we shall see, these polysilicon features are doped during manufacture and this improves their conductivity. Aluminum remains the metal of choice for all high-conductivity pathways, although unlike the older aluminum/oxide/semiconductor sandwich that provided only two topological levels on which to make interconnections, the basic silicon-gate structures provide three and are therefore more compact and correspondingly faster.

The three topologically distinct levels defined by the manufacturing process are shown in Figure 11a. The middle level of polysilicon is insulated by interposed oxide from both the upper aluminum level and the lower diffusion level and therefore may be used to provide pathways that cross under the aluminum without touching it, if required, though the figure shows the poly and aluminum in contact. Such cross-under pathways can also be defined on the diffusion level, but wherever the designer's layout requires a diffusion level path to cross under a polysilicon path as in Figure 11b, the process establishes a transistor! Note that the polysilicon strip, in its glass tunnel, dips down to a physical level almost in contact with that of the substrate (Figure 11c). The thinner the tunnel floor at this point, the lower the field-effect threshold of the transistor; but if the floor is too thin, the transistor will be turned on too easily. The designer, however, follows only a set of simple rules dealing with relative sizes and spacings of planar features, and depends on the manufacturing process to take care of the third dimension. The planar layout in

Univ. of Washington Libraries

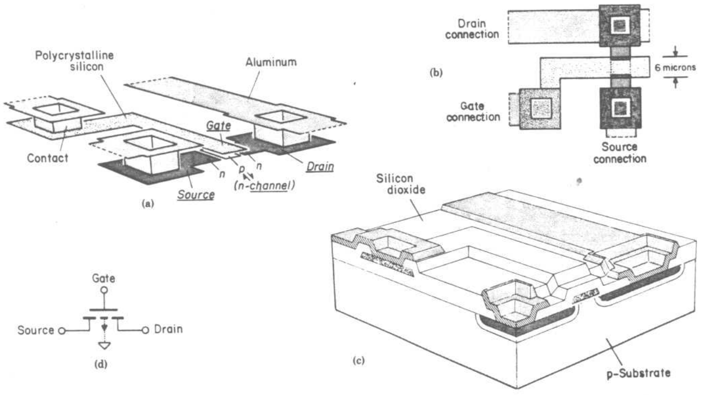


FIGURE 11. An n-channel silicon-gate transistor: (a) topological levels; (b) planar layout; (c) structure; (d) symbol (arrow optional).

Figure 11b is typical. It is scaled to a gate width of 6 microns (a conservative figure in 1980, yet only about 10 wavelengths of visible light) and follows design rules established by Mead and Conway [MEAD80]. In Figure 11c the surface-layer thicknesses are exaggerated for clarity; the thickness of the aluminum layer, for example, is approximately $\frac{1}{2}$ micron. MOS structures are very small and very thin.

Yet not all that planar, our glassy Lilliput; poly-filled tunnels and deeply pitted highways of aluminum run over hill and vale in this fragile, wrinkled domain resting upon a vast plateau of warring crystal, where manifold subterranean battles are fought futilely on and on by agitated electrons against mythical holes. Quite an interesting gadget, actually, even without the battery; but how in the world does one go about making such a thing?

1.3.2 Fabrication of MOS Integrated Circuits

One starts by growing a large silicon crystal ingot, the larger the better. This is done by attaching a small crystal, the *seed*, to the end of a rod, dipping the rod into a vat of molten silicon containing the requisite pinch of boron, and then withdrawing the rod very, very slowly. Oriented, polished wafers are made from this ingot, each one about a third of a millimeter thick and 10 centimeters in diameter. A few hundred replicas of the circuit to be made will be built up simultaneously on each of several wafers—each replica containing many thousands of transistors and their interconnections. Under various kinds of atmosphere and degrees of temperature, phosphorus is diffused, oxide is grown, and polysilicon and aluminum are deposited, each at a different step in the process.

Planar features are created by a combination of photographic and etching processes. A set of photomasks defining the arrayed replica patterns to be etched into the various layers is carefully prepared from the designer's layout, one mask for each level of structure or kind of feature. Levels to be etched are first coated with a light-sensitive emulsion called *photoresist*, then exposed to light through the appropriate mask, and finally immersed in acid or other strong solution to etch away either the exposed or the unexposed part of the pattern, depending on whether positive or negative photoresist had been used. A bath follows.

Microstructure fabrication is a tricky business in which many things can go wrong. Finished wafers are never perfect. Obvious flaws are identified and marked, the wafers are carefully scored with a sharp tool, and with something like a kitchen

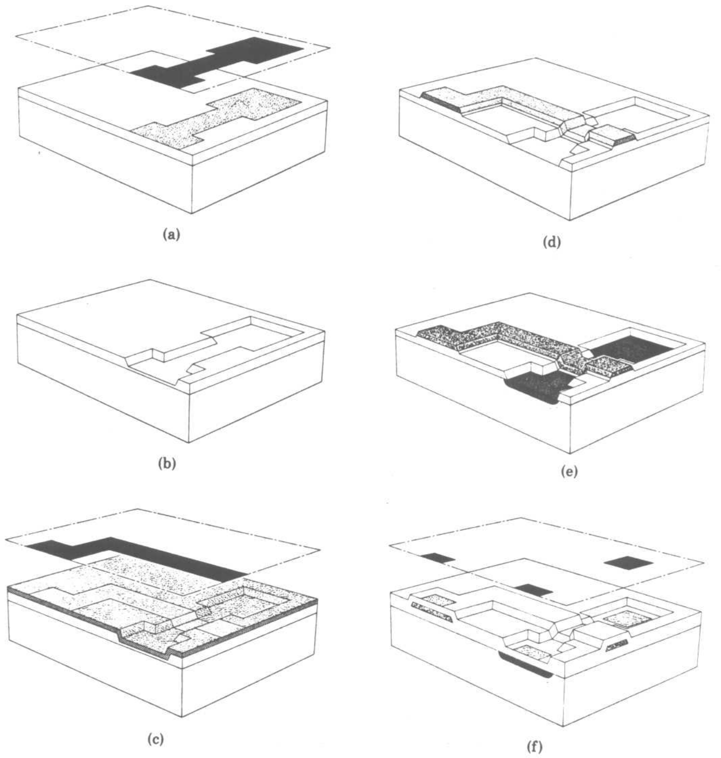
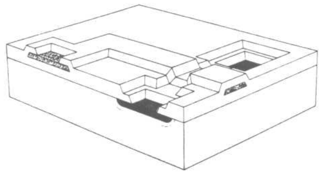
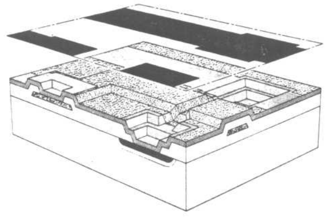


FIGURE 12. MOS fabrication steps. (a) Diffusion mask defines planar features on a photoresist-coated oxide layer. (b) Etching removes unexposed oxide. A thin floor is formed to insulate the gate. (c) Poly mask defines features on a photoresist-coated layer of polycrystalline silicon. (d) Etching of exposed poly forms the gate structure. Additional oxide is removed to uncover the substrate areas defined earlier. (e) Phosphorus is diffused into all uncovered silicon, thus forming source and drain regions and improving the conductivity of the gate structure. Gate-oxide blocks diffusion into the substrate. (f) Contact mask defines cutout areas on a photoresist-coated oxide covering. (g) Etching removes unexposed oxide to uncover contact areas on the poly and diffusion levels for circuit connections. (h) Aluminum mask defines pathways on a photoresist-coated layer of aluminum. (i) Etching of exposed aluminum forms "wires" and their contacts.

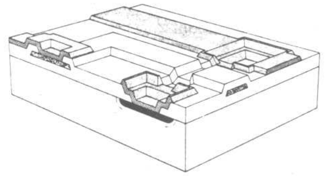
rolling pin the individual replicas—now chips—are broken apart from one another. Marked chips are discarded; unmarked ones are mounted in a small plastic or ceramic case. Tiny gold wires are maneuvered into position under a microscope and then bonded to aluminum pads on the chip, connecting the circuit to convenient metal pins along the edge of the case. A lid bearing the manufacturer's emblem is sealed on and the finished package is tested. Bad packages are discarded or sold at a discount; good ones are shipped to the customer, sometimes filling hollow plastic or metal "sticks" which then cost about a million dollars a cord. Large packages such as microprocessors are sent along in more dignified containers.



(g)



(h)



(i)

Planar features, already near the resolution limit of the light used in their definition, cannot easily be made much smaller by optical means. Even so, the manufacturing tolerances maintained throughout the process are phenomenal. Mask alignment is routinely held to 1 micron in 10 centimeters, an accuracy of one part in 10^5 quite without precedent in industrial practice. If a 10-centimeter wafer is scaled up conceptually to 200 miles—the distance from Washington, D.C., to New York—then 1 micron is about the width of the crown on the head of the Statue of Liberty.

We now follow our simple planar layout through the basic process. Definition of physical structure begins with a masking step on an oxide-covered substrate. The first two steps (Figure 12a and b) establish the tunnel floor and the source and drain areas. The next three steps (Figure 12c–e) form the transistor itself, but leave it bare and unconnected.

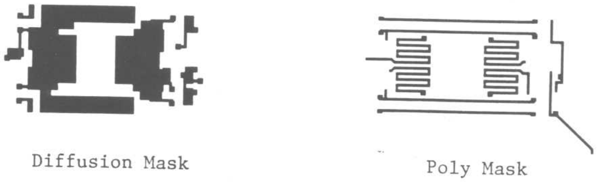
The reader has not failed to notice that the transistor itself is a very small part of the whole, namely, just that part of the gate structure which has the thin tunnel floor, together with a thin slice of substrate just below it. It occupies less than 3 percent of the area of our layout, and much less than 0.03 percent of the volume of underlying substrate. All else serves either to support it, to isolate it, or to connect it into the circuit. Transistors can be packed more closely than this implies if pathways are efficiently routed on all three levels, but most of the area of any integrated circuit remains committed to interconnections or "wires"—and our chip has yet to be wired into its case, and the case into the computer. Electronic computer circuits are *made* of wire.

In any case, our fabrication surface must now be covered once again with oxide in order to insulate the bare semiconductor areas from the final uppermost pathways. Windows must then be cut to provide access to the gate, source, and drain regions wherever required for circuit connections (Figure 12f and g).

We are now ready to build the uppermost pathways. At the places where connections are to be made, areas are enlarged somewhat to assure good interlevel contact even when masks are not in perfect registration; all pathways are otherwise made as small as possible in order to conserve area. Note, by the way, that the task of aligning the poly and diffusion masks is made a bit easier by the fact that it is only their intersections—wherever they occur—that define transistor boundaries. This self-alignment feature is partly responsible for the success of silicon-gate technology.

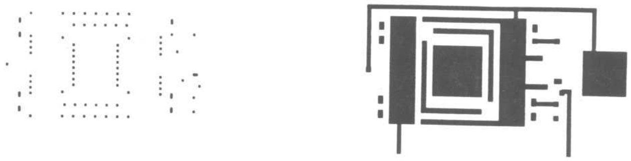
A thin covering of aluminum is deposited over the entire surface of oxide, now quite complex in structure. Etching then removes all but the requisite "wires" and contacts (Figure 12h and i).

The final steps, not illustrated, involve covering the surface with oxide once again and then etching giant windows to gain access to the giant aluminum bonding pads, to which giant gold wires will be attached. On our present scale of things, the final package is Gargantuan.



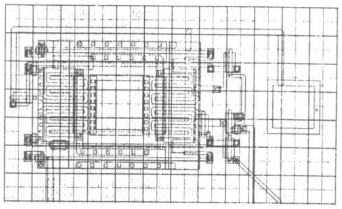
Diffusion Mask

Poly Mask



Contact Mask

Aluminum Mask



Composite Layout

FIGURE 13. A MOS output amplifier circuit: photomask set and composite layout (a computer plot). Smaller staging transistors appear near the left and right edges of a pair of large serpentine-gated output transistors surrounding a bonding pad. A second pad nearby is attached. Multiple contacts share the current load.

Before we leave the subject of fabrication, it is a good idea to point out that MOS transistors do not have to be so tiny even though much effort is devoted to making them as small as possible for greater packing density and higher speed. Indeed, MOS output transistors *must* be large (broad) if they are to supply enough current to operate other packages holding power-hungry circuits such as TTL. Figure 13 shows the four-mask set and a composite computer plot of a pair of giant serpentine-gated output transistors surrounding a bonding pad, all attached to a second pad nearby.

Computer aids to design and layout are essential. Most of the circuit of Figure 13 was obtained by calling forth one plotting program subroutine for TRANSISTOR, another for PAD, and so on. The amplifier itself was designed by I. E. Sutherland.

The word "large" in large-scale integration refers not to the size of the transistors, but to the number of transistors on one chip. Chip size is limited by a consideration of yield: the larger the chip, the greater the probability of surface flaws and therefore the slimmer the chances of finding enough good chips per wafer to make batch production profitable. A good deal of effort thus goes into the development of efficient structures that lead to high packing densities. Chips holding 10,000 transistors are in routine production, and designs of much larger scale are already in progress.

The microcircuit industry is inclined to protect its huge investment by minimizing risk. Mistakes are very costly and progress toward higher levels of integration is made cautiously. But the industry is now at a turning point: x-ray lithography or electron-beam technology must take the place of photolithography if devices are to be made significantly smaller, though the almost incredible problem of mask alignment stands in the way of their full use. One-mask structures already are being made by means of these new techniques and demonstrate the feasibility of feature densities two or three orders of magnitude greater than those achieved under ultraviolet light.

We began the discussion of integration with Dummer's prediction. We will close it with another, and then move on to consider aspects of the organization of microstructural units. Sutherland, Mead, and Everhart [SUTH76], after a very thorough review of the work of many U.S. laboratories aimed at producing devices of submicron dimensions, arrive at the following startling conclusion.

> There is every reason to believe that the integrated circuit revolution has run only half its course: the change in complexity of four to five orders of magnitude that has taken place during the past 15 years appears to be only the first half of a potentially eight-order-of-magnitude development. There seem to be no fundamental obstacles to 10^7 to 10^8 device integrated circuits.

One hundred *million* transistors? Ultrasystem chips in which microprocessors as we now know them, rather than transistors, are the operative units? Powerful storage arrays with integral processing? The designers of complex algorithmic systems face a very important challenge. Let us hope that they do their job well and that we all learn how to make good use of their work. One hundred million is a large number, and devices will be very small indeed. On the Washington-to-New York scale, today's transistor is perhaps the size of a room; tomorrow's transistor may be the size of a book.

2. THE LOGIC OF MOS CIRCUITS

We must now discover transistor combinations that can be made to do the work of logic. Only a few basic forms are required for the synthesis of useful complex structures.

2.1 Basic Combinations

2.1.1 The Inverter

The MOS transistor is essentially an on–off switch with a bad connection; it can be turned completely off but not completely on. The transistor's channel resists the flow of electrons, which thus become less numerous (though a bit speedier) toward the drain end. In bumping through a chain of conducting channels, negative electron ranks become thinner link by link; the relative positivity of the interchannel pathway correspondingly increases. Broad "easy" channels increase the positivity less; narrow "hard" ones, more. The *inverter* is based on these facts.

In the two-link inverter, the upper channel, long and narrow, has a *high* resistance; the lower one, short and broad, a *low* resistance. Uppermost gate and drain are connected via an aluminum pathway to the positive side of a 5-volt battery (Figure 14a); the lowermost source is connected to the negative side and anchored to establish a zero-voltage reference level. The upper transistor serves only to provide a fixed high-resistance link since it is always on by virtue of its highly charged gate, a fact emphasized by the use of a simplifying zigzag symbol. The lower transistor, on the other hand, will be on only when its gate voltage, V_{in}, is given a value above the general 1-volt threshold. When this happens, the output voltage, V_{out}, takes on a *low* relatively positive value as electrons are pushed by the battery through the low-resistance link of the completed chain and flow past the midchain output point, as shown in Figure 14d. V_{out} otherwise takes on a relatively *high* value (but of course not more than 5 volts) as some of the unanchored output pathway's electrons are pulled away to the battery through the fixed "pull-up resistor" (Figure 14c). Thus a *high* V_{in} implies a *low* V_{out}, and vice versa. Proper choice of both battery voltage and channel-resistance ratio ensures that high and low are values, in turn, well above and below threshold so that one inverter will drive another reliably. In logic symbols

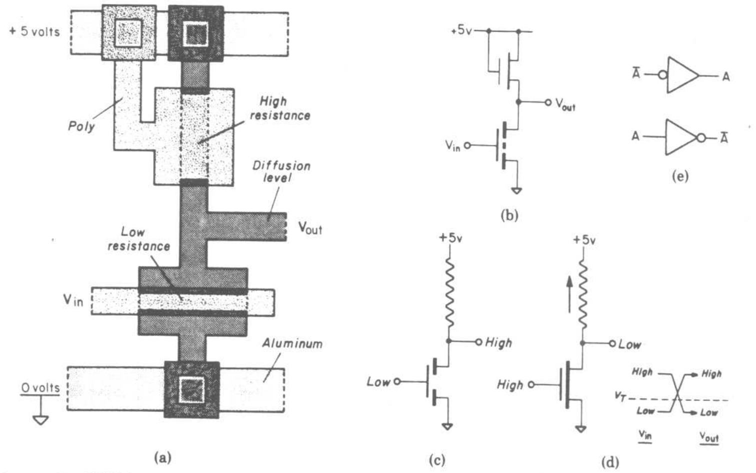


FIGURE 14. MOS inverter: (a) layout; (b) circuit; (c) off; (d) on; (e) logic symbols. The upper-link symbol is simplified in (c) and (d).

(Figure 14e), we identify high with "one" or "true"; low with "zero" or "false." If A is true then its inverse, NOT-A (written $\bar{A}$), is false.

2.1.2 Not-AND (NAND) and Not-OR (NOR) Circuits

Simple elaborations of the inverter chain give us the two basic "logical" gating elements of coordinated charge flow, the NAND and NOR gates. In the NAND circuit of Figure 15a, the output will be low only when both of the inputs, A and B, are high, for only then will a conductive pathway to the zero anchor be established. In the NOR circuit of Figure 15c, the chain is a bifurcated one and a high on either of the inputs, A or B (or on both, of course), will do the trick. The expression "A and B" is written AB; the expression "A or B" is written $A + B$.

The principles observed in designing such combinatorial circuits are simple ones: a single pull-up resistor must be fixed "above" the point of output and a network of variable-gated transistors attached "below"; the ratio of fixed resistance to maximum "on" variable resistance must be made about equal to that of the inverter, about 10 or 20 to 1; the size of the overall structure generally must be kept small for greater speed. Figure 16 shows a somewhat richer example.

Combinatorial circuits based on switching devices have been studied thoroughly since the publication of the classic paper on the subject by Shannon [SHAN38]. In the pretransistor years much activity was devoted to the development of techniques for circuit synthesis, first in the design of telephone relay networks and later in the design of digital computers. In modern integrated circuit layouts the logical is nicely related to the physical, and we see wires and switches at a glance—or perhaps a stare when matters become just a bit complicated, as in Figure 17.

2.1.3 Flip-Flops

Seldom as lively as its name implies, the flip-flop is in fact a bistable circuit that will hold itself in one of two preset states as long as the battery remains in operation. Bistability is provided by a trick of cross-

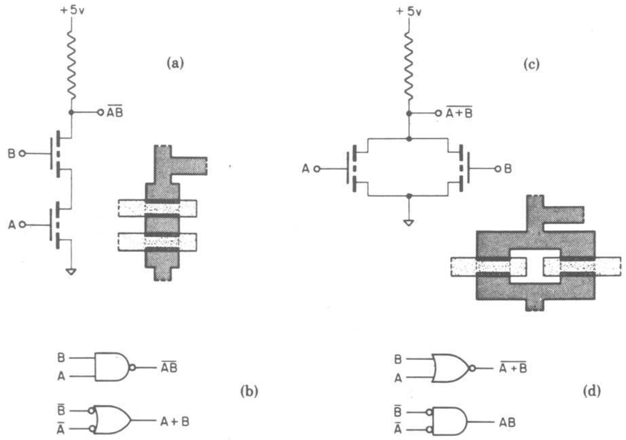

+5v
(a)
$\overline{AB}$
B
A
(c)
$\overline{A+B}$
A
B
(b)
$\overline{AB}$
$\bar{B}$
$\bar{A}$
A+B
(d)
$\overline{A+B}$
AB

FIGURE 15. NAND and NOR gates: (a) NAND circuit; (b) NAND gate and its equivalent low-input-OR gate logic symbols; (c) NOR circuit; (d) NOR gate and its equivalent low-input-AND gate logic symbols.

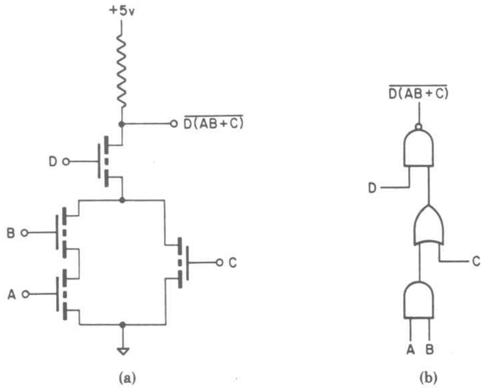

+5v
$\overline{D(AB+C)}$
D
B
A
C
(a)
$\overline{D(AB+C)}$
D
C
A
B
(b)

FIGURE 16. A combination of logic gates: (a) circuit; (b) logic net.

connection between negating logic combinations. Figure 18 shows a version based on a pair of NOR gates, often known as a latch.

Figure 18a shows the circuit initially locked in its Q-low state by the cross-connected $\bar{Q}$ output, which is "unanchored" and therefore high. Circuits external to the figure provide R (reset) and S (set) inputs in such a way that either R or S, but not both at once, can be raised momentarily to a high level. If S is now made high, output $\bar{Q}$ is "anchored" and becomes low; the cross-activated link in the R-side NOR gate is broken and output Q promptly becomes high. This Q-high signal crosses back to establish a second anchoring pathway in

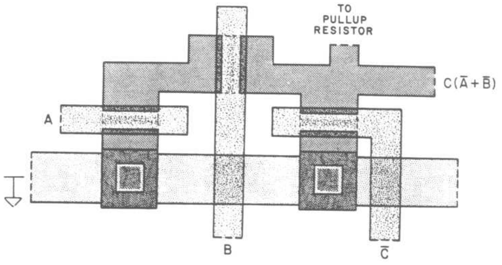


FIGURE 17. A planar layout of a combinatorial circuit with cross-under pathways and inverted logic terms.

the *S*-side NOR gate that will continue to hold $\bar{Q}$ low after the momentary *S* input signal disappears. Circuitous! The flip-flop now appears locked in its *Q*-high state. A momentary high on the *R* input will reset the circuit to its initial *Q*-low state. It is clear that *R* and *S* must never be made high at the same time if the circuit is to be well-behaved. There will be more to say about this shortly.

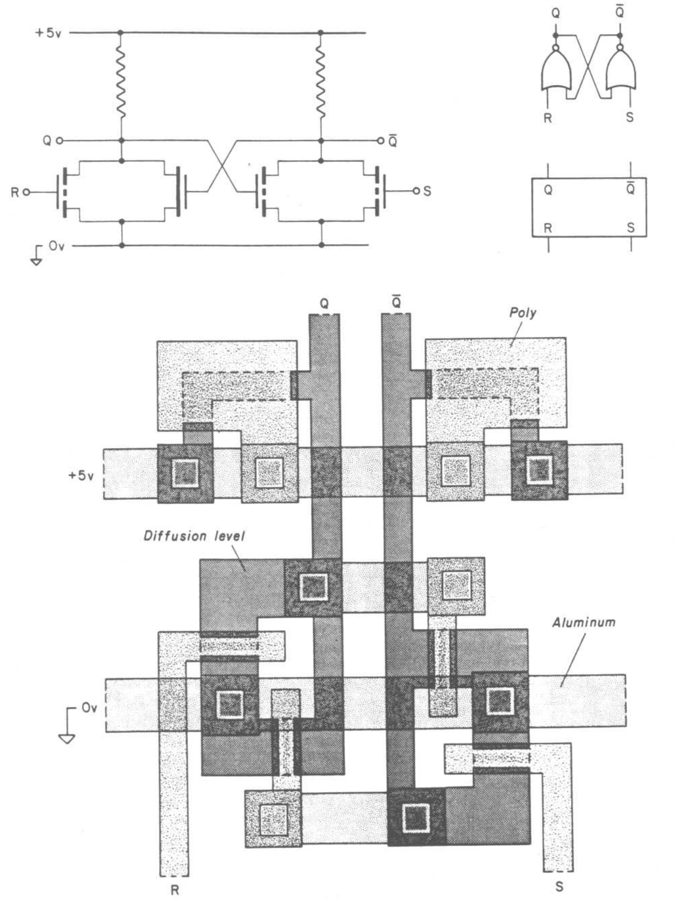
The so-called *D* flip-flop (Figure 19) is often a more convenient design element. It is a bistable circuit in which the changes of state can occur only during specific intervals of time established by an externally generated "clock" signal. The timing diagram (Figure 19d) shows the final transition of the *D* input signal as occurring while the *C* input is high, an event resulting in out-of-step outputs. Clearly, the briefer the "high *C*" intervals, the better the time definition of the outputs. There is a limit, of course: if these high intervals or clock pulses are too brief, the circuit will fail to "flip" or "flop." An alternative way to improve time resolution involves the addition of other elements to the *D* flip-flop circuit so that it will ignore the *D* input signal once the low-to-high *C* transition has been made, and will attend further *D* input changes only after the subsequent "safe" high-to-low *C* transition. These more complex flip-flops, said to be leading-edge triggered, make possible the reduction of timing uncertainty to a minimum—though not to zero, of course, since electron mobilities are finite.

Many other flip-flop forms of even greater complexity are possible. In one robust version, for example, *Q* outputs are led back to input steering gates in such a way that presentation of a "true" value at transition time results in a complementing of the state of the flip-flop. All such complex flip-flops, however, are relatively large and are used only when they simplify, or are essential to, the overall design.

We make a short detour at this point:

Regardless of complexity, no flip-flop is predictably well-behaved when triggered by marginally adequate signals. This is no less true of integrated circuit flip-flops than it is of any other form of bistable device that one is able to contrive. In the view of Holt [HOLT77] the problem is intrinsic to the decision process itself and thus one that challenges our most profound concepts of choice and cause. From some of the earliest perceptions of difficulty [GRAY63, CATT66, LITT66] to the more recent experimental work and theoretical analysis [MARS68, CHAN73, COUR75, HURT75], the subject has grown to maturity and, surprisingly, considerable controversy. An excellent account of the phenomenon in the context of system timing is given by C. L. Seitz in MEAD80.

The following story has been told in many other settings. It is recounted here in the spirit of that alarmed politician who, two thousand years ago, took advantage of every oratorical opportunity to call for the precautionary destruction of Carthage. It does seem that the potential for instability in what we would like to regard as bistable mechanisms cannot be brought to our attention often enough.



(c)

FIGURE 18. An *RS* flip-flop (latch): (a) circuit; (b) logic form; (c) planar layout. Poly level and diffusion level features are connected to one another via the aluminum level. Q and $\bar{Q}$ outputs appear on all three levels, thus simplifying later integration.

In a clocked flip-flop, marginal triggering can occur whenever a data input, for some reason, happens to change just at the moment of clock transition. In an *RS* flip-flop, marginal triggering can occur whenever reset and set signals are presented concurrently and then withdrawn nearly simultaneously. Instead of recording YES or NO with convincing authority, the marginally triggered flip-flop can become vacillatory or perhaps metastable, its confused outputs interpretable as "neither yes nor no," or

Univ. of Washington Libraries

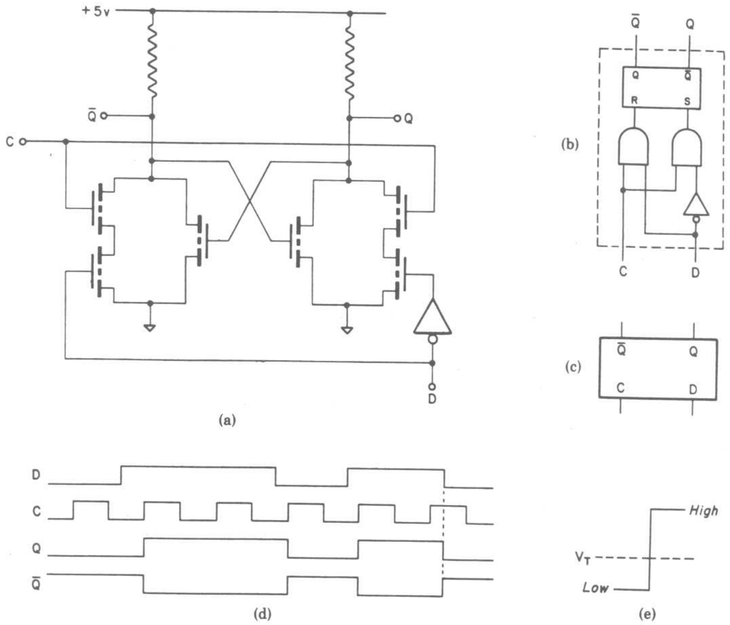


FIGURE 19. The *D* flip-flop: (a) circuit; (b) logic form; (c) symbol; (d) timing diagram; (e) voltage "swing." When the *C* (clock) input is high, *Q* and $\bar{Q}$ outputs match the *D* (data) input and its inverse.

worse, "yes *and* no." Furthermore, it is impossible to predict the durations of any such periods of indecisiveness. One can know only that such periods are more likely to be very short than very long, the odds at best falling exponentially with length in a way that varies from circuit to circuit. The logician groans.

These occasional dicey outcomes, known in some cultures as "glitches," are unavoidable in systems having more than one clock or timing basis. Synchronizing circuits can be interposed between independently timed subsystems to trap most of the misbehavior and thereby reduce the scope of unwanted effects. Figure 20 illustrates one method of synchronization in the context of interaction between a randomly thrown switch and a clocked subsystem.

Actually, two independent processes are depicted in Figure 20. In the first, an $\bar{R}\bar{S}$ latch is used to tidy up the signals produced by operation of the switch when its mechanical arm bounces against the "target" contact before coming to rest; the resulting filtered output, *A*, is indeed quite tidy but is not at all in step with the clock pulses. In the second, a pair of edge-triggered *D* flip-flops attempts to reconstruct an authoritative clock-cycle-quantized approximation of signal *A*; successful or not, it sends its result, *C*, to the clocked logic subsystem.

The timing diagram in Figure 20 shows small delays between causes and effects in order to clarify sequential dependency. For example, the latch (cross-connected NANDs) responds to the first appearance of a low on either the $\bar{R}$ or $\bar{S}$ input lines

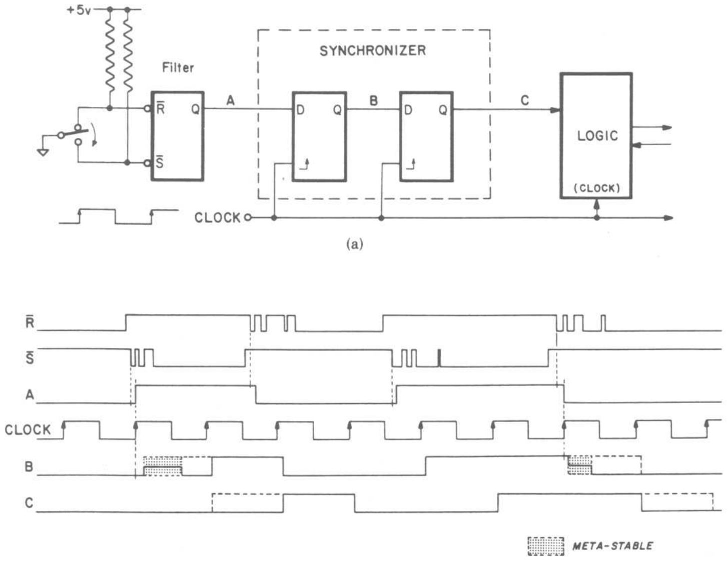


FIGURE 20. Clock synchronization of a randomly timed signal: (a) logic; (b) timing.

after a characteristic small delay. In other respects the A, $\bar{R}$, and $\bar{S}$ tracks are grossly foreshortened in the diagram; mechanical switching and bouncing epochs typically span many tens of thousands of clock cycles. The first rising and last falling A transitions are shown as happening to coincide with clock triggering edges, thereby creating two B glitches, both of which happen to resolve to a final low after a fairly short period of instability.

The principle of glitch-trapping synchronization is this: Owing to its exponential time dependency, the probability that a B glitch created on one clock transition will survive long enough to a cause a C glitch on the following clock transition can readily be reduced to any arbitrary acceptable value by setting the clock rate sufficiently low. By the same consideration, synchronizers run at increasingly high rates will fail dramatically. In any event, it is clear that synchronization introduces both delay and pulse-width uncertainties.

A self-timed arbiter [SEIT80] is a circuit combining synchronizers and decision elements in such a way that it can withhold a change in, or validation of, its output signals until stability appears to be assured. For this service it demands the sacrifice of decision time predictability. Not all self-timed arbiters work well.

End of detour.

2.1.4 Drivers

Where speed of operation is important, fairly powerful (i.e., low-resistance) combinations are required to supply charge current rapidly to the many gates often found connected to a clock line or other pathway. Three of the forms such drivers take in current practice are shown in Figure 21. In the simplest form (Figure 21a) both upper

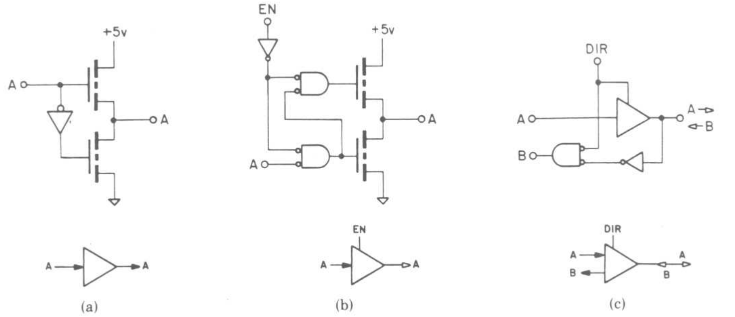


FIGURE 21. Drivers: (a) simple driver (noninverting); (b) a three-state-output driver; (c) a three-state-output driver/input receiver. EN—enable; DIR—direction.

and lower transistors of a tandem pair are given low-resistance channels, an inverter ensuring that only one of the two will be on at a time to force the output voltage to a high or low level. All CMOS elements are based on this tandem-pair principle but avoid the heat-producing inverter altogether by including one low-on and one high-on transistor in each pair, as foreshadowed in Figure 6b.

If the gating network of Figure 21b is now added, it becomes possible to disconnect the output line from the output transistors altogether, thereby giving us a useful "third state." Another such three-state-output driver can be attached to this same output line, and another, and another, so that signals from many different originating nodes can now be sent along the same signal bus pathway, provided that only one node is licensed or enabled to drive at any one time. The final form (Figure 21c) provides for both sending and receiving by combining a three-state-output driver with an inverter and a low-input AND gate.

While we are on the subject of driving, the reader surely will excuse the need for a few steps backward at this point in order to see how a trick of fabrication—*ion implantation*— results in a significant increase in driving capability and a general decrease in physical size. The trick involves the deliberate alteration of the characteristics of the transistor so as to produce an inherently inverted rather than a field-inverted channel. This is accomplished early in the fabrication sequence for n-channel MOS by using an elaborate ion-accelerating gun to shoot phosphorus ions into some of the channel areas of the substrate. Additional masking and processing steps are necessary. Ion-implant mask areas overlap those of the partially completed transistor candidates; the additional steps create temporary access windows through which ion bullets can be shot into the substrate. The ions are then accelerated to a velocity just high enough so that they lodge themselves in the selected future channels to a carefully controlled depth. Subsequent standard processing steps will then build transistors that are normally on rather than normally off around each of these implanted channels. The standard MOS transistor, normally off, is turned on by application of a gate voltage that is greater than some positive threshold value. The implanted-ion transistor, normally on, is turned off by application of a gate voltage that is less than some negative threshold value. But inasmuch as negative gate voltages never appear in our circuits, one can read "always on" for "normally on."

Why is it advantageous to use these normally on transistors in design? The simple answer is that they form better pull-up resistors when used in a certain way, but a full explanation of the improvement would take us farther back than we care to go. We would have to examine the notion of chan-

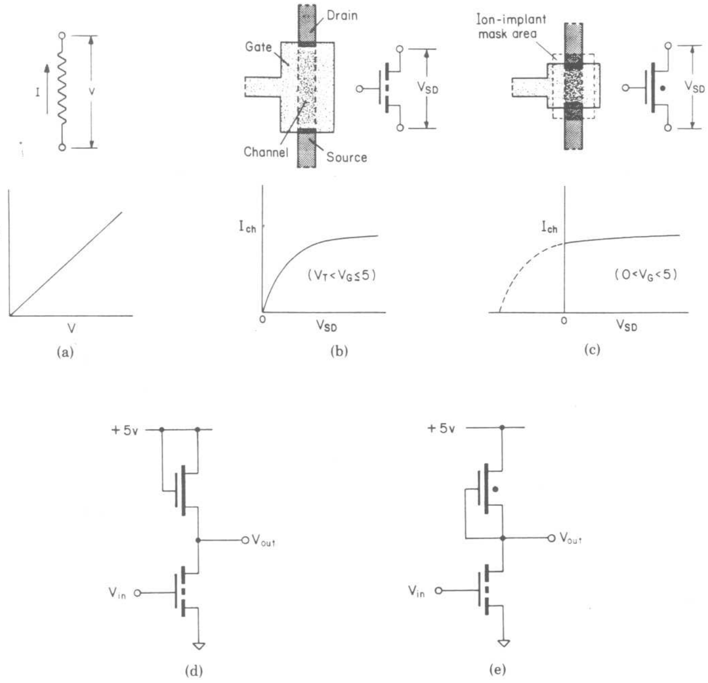


FIGURE 22. Comparison of "resistors": (a) pure resistor and linear current/voltage characteristic; (b) a standard MOS transistor and characteristic; (c) an implanted ion MOS transistor and characteristic to source (the dotted symbol denotes "normally on"); (d) inverter circuit with standard MOS pull-up "resistor"; (e) inverter with implanted ion pull-up "resistor."

nel resistance in more detail, for example, and would immediately find that our pull-up resistor is not in fact a purely resistive component (for which, in a careful exposition, the zigzag symbol would strictly have been reserved). A purely resistive component would exhibit the linear relationship between current *I* and voltage *V*, known as Ohm's law (Figure 22a), whereas our pull-up resistor, formed by connecting together the gate and drain of a standard MOS transistor, would be found to exhibit a markedly nonlinear relationship (Figure 22b), as would the resistor based on an implanted-channel transistor in which the gate is connected to the source (Figure 22c). We would then have to call upon the mathematical tools of circuit analysis for a clear understanding of the input/output behavior of basic inverter configurations with nonlinear components. Our reward, however, would be the discovery that the implanted inverter (Figure 22e) requires an upper-to-lower size ratio of only 5 to 1 for an adequate output voltage swing, while the standard inverter (Figure 22d) requires a ratio of 10 to 1 or greater. We would also learn that the implanted inverter "swings" more crisply, and therefore drives more capably, than does its counterpart.

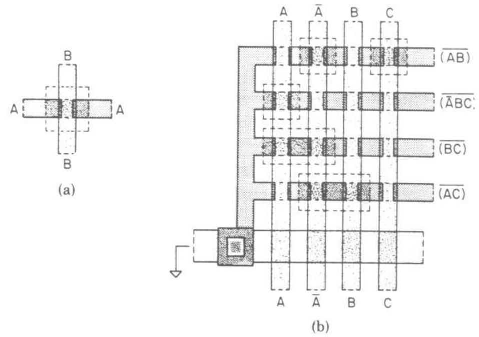


FIGURE 23. Decoder matrix using implanted ion transistors: (a) an *ion cross*, showing continuity of signal pathways on diffusion (AA) and poly (BB) levels; (b) a combined matrix. Dashed lines enclose the ion implant mask areas.

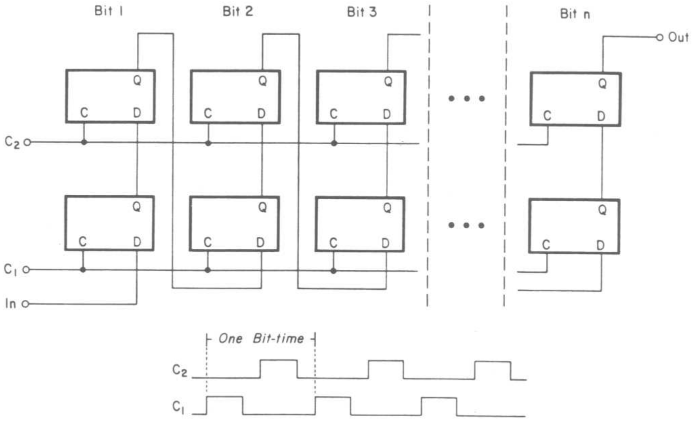


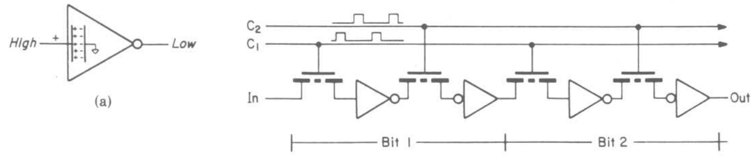
FIGURE 24. *D* flip-flop static shift register.

2.2 Logic Arrays

The basic circuits can be combined in many different ways to form more or less regular arrays of logical structure. Combinatorial logic circuits may or may not assume regular form; storage elements almost always do. But in any case, simple linear shift registers, as well as complex memories and logic, are all teased into geometrical forms having felicitous planar projections. The catalog of structures is a very large one, and we can look at only a few of its illustrations.

2.2.1 Decoders

In its simplest form, a decoder is essentially a set of logic gates with several outputs, each output asserting the presentation of a



(a)

(b)

FIGURE 25. Simple dynamic shift register: (a) charge storage at the inverter input gate; (b) circuit.

specific pattern of input values. A delightfully compact decoder structure has been illuminated by Sutherland. It is based on the observation that the implanted ion transistor defined by poly and diffusion pathways that cross in an implanted zone as in Figure 23a will nicely preserve the continuity of signal flow on both levels, since it will always be on for all proper signals on the crossing poly level. This structural feature—what we may call an *ion cross*—can then be used to great advantage in decoder matrix forms when combined with standard transistor "crosses" as in Figure 23b.

2.2.2 Shift Registers

Because of their simplicity and regularity of structure, shift registers were among the earliest candidates for large-scale integration. Two main families, termed *static* and *dynamic*, have been developed. In the static shift register family information can be held for indefinitely long periods in the absence of clock pulse activity. In the dynamic shift register family periodic or occasional clock pulses are required to prevent loss of information, an inconvenience justified by a concomitant reduction in the overall size of the structure.

One *static shift register* makes use of *D* flip-flops in pairs, with one pair per bit (Figure 24). Trains of nonoverlapping clock pulses, C_1 and C_2, alternately transfer data from the input and the upper rank of flip-flops into the lower rank, and from the lower rank into the upper and to the output. Each pair of clock pulses shifts the information along one bit position. Time must be allowed between C_1 and C_2 pulses to permit the flip-flops to stabilize before passing information onward. Effective but big: each bit inhabits a 20-transistor "cell."

The *dynamic shift register* shown in Figure 25 has a six-transistor cell. Its operation is based on the storage of charge on electrically isolated pathways. Recall that when a gate pathway is somehow disconnected from all other pathways after some of its electrons have been removed by application of a high signal, the positive charge thus trapped above the gate oxide will sustain the field-induced channel in the substrate below until electrons are reintroduced by application of a low signal. Unfortunately, such isolation is never perfect; electrons wander back onto the isolated structure along a number of unplanned and unwanted yet unavoidable minor routes, thus leading to a gradual loss of the gate's positivity and the collapse of its field. Any information held in place by the field must be passed on or otherwise regenerated in a timely fashion. Charge is retained long enough, however, to make possible the simple arrangement shown in Figure 25.

Two ranks of inverters are operated by two trains of nonoverlapping clock pulses. In Figure 25 the two ranks are shown interleaved in order to emphasize structural linearity. C_1 pulses control "pass" transistors which establish pathways across the cell boundaries; C_2 pulses establish pathways within the cells. Alternation of pulses therefore shifts information along; failure to apply pulses often enough results in information loss.

Simplification of the static shift register follows directly (Figure 26) by extension of this technique. C_1 and C_2 pulse pairs shift information along as in the previous example. Bistability is provided by adding a *feedback* pass transistor to each cell; it is turned on by a third clock pulse, C_3. Small delays between pulses allow inverters to stabilize. So long as C_2 and C_3 are both held high, the inverters within each cell are cross-connected and information will remain in place without further attention.

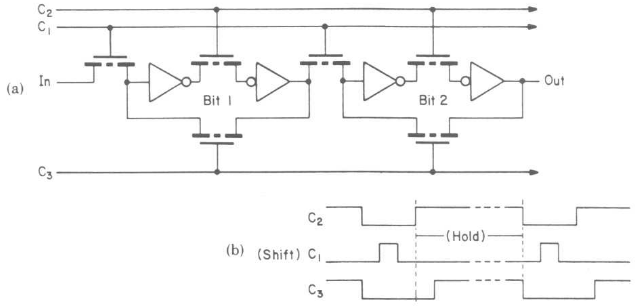


FIGURE 26. Simple static shift register. Information remains in place as long as C_2 and C_3 signals are held high.

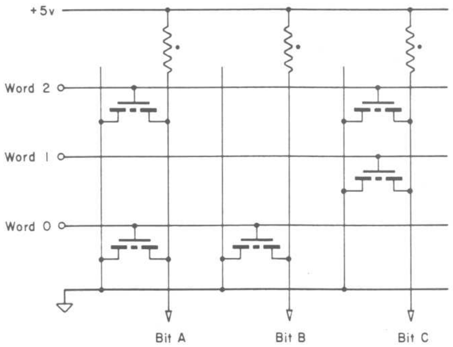


FIGURE 27. Read-only-memory (ROM) matrix. Word lines select partially filled rows of transistors arranged in NOR columns. Absence of a transistor corresponds to high output.

2.2.3 Read-Only Memory (ROM)

Structurally very regular, ROM is made in many forms. Figure 27 shows one simple form in which stored information is represented by the presence or absence of transistors in a storage matrix. When a high signal is applied to a selected word line, high signals will remain on only those bit-output lines for which the corresponding matrix-element transistor sites are unfilled. In *mask-programmable* ROM manufacture, variants of one basic chip are produced by specifying different patterns of filled and unfilled sites, a matter of relatively easy alteration of just the diffusion mask.

Other ROM forms of greater complexity provide for alteration of stored patterns by means of externally applied "programming" signals. In one popular form selected matrix elements are permanently disabled by energetic signals. In another ROM form the information storage mechanisms are such that stored patterns can be erased by ultraviolet light, which shines down upon the transistor plane through a transparent window thoughtfully built into the lid of the integrated circuit package.

Univ. of Washington Libraries

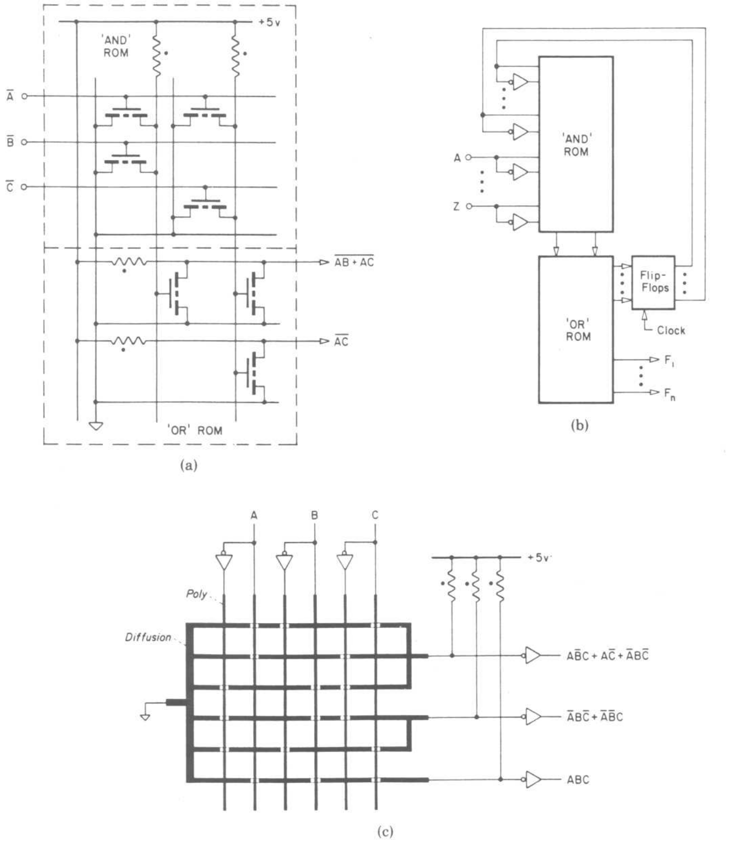

'AND' ROM
+5v
$\bar{A}$
$\bar{B}$
$\bar{C}$
$\overline{AB + AC}$
$\overline{AC}$
'OR' ROM
(a)
'AND' ROM
A
Z
Flip-Flops
Clock
'OR' ROM
F_1
F_n
(b)
A
B
C
+5v
Poly
Diffusion
$A\bar{B}C + A\bar{C} + \bar{A}B\bar{C}$
$\bar{A}B\bar{C} + \bar{A}\bar{B}C$
ABC
(c)

FIGURE 28. Programmed logic arrays (PLA): (a) ROM pair; (b) typical PLA form; (c) transistor-cross matrix with "wired-OR" outputs. "White" crosses, standard transistors; all others, *ion crosses*.

2.2.4 Programmed Logic Array (PLA)

By connecting two ROMs in tandem, we obtain a structure that generates outputs in canonical sum-of-products form in response to the presentation of sets of input signals. Although both ROMs have the same NOR gate composition as that of the previous example, we view the input ROM as the AND (logical product) generator for low input signals and the output ROM as the OR (logical sum) generator with low outputs as in Figure 28a. PLA combinations typically provide inverters so that both true and false values can be included in the static logic expressions developed by such a pair of ROMs. They can be extended to sequential operation—in a direct implementation of the well-known finite state automaton—by incorporating flip-flops having outputs fed back to some of the PLA inputs, as in Figure 28b. An appealingly simple AND/OR structure based on the transistor-cross matrix is shown in Figure 28c. It omits the OR ROM altogether but may require some duplication of AND terms, as illustrated on the middle two rows of the matrix.

2.2.5 Read–Write Memory (RWM)

Both static and dynamic forms of read-write memory arrays are made. The *static RWM* cell shown in Figure 29 illustrates the use of a bistable pair of inverters as the storage element. Pass transistors connect opposite sides of the pair to bit lines for reading and writing. Quite indifferent to the direction of flow, those pass transistors turned on by a high row-select signal will conduct information into the cell if either the bit or the NOT-bit line is driven low externally, but will otherwise conduct information out of the cell. The bistable pair either assumes the state forced by the bit line signals or else puts its own signals onto the bit lines without changing state.

In order to expose some of the ideas involved in the simplification, we will take only the first step down the path that leads ultimately to a one- or two-transistor storage cell. The six-transistor static RWM cell can be simplified directly by removing all inverter pull-up components to a zone outside the array, from which a single pair can be time-shared by all cells within the same bit column. The resulting *dynamic RWM* cell array is shown in Figure 30. The occasional high signals on the row-select lines which are presented during normal readout operations temporarily reconnect the selected row of cells to the set of shared pull-ups, and thus refresh the information stored along the row. Logic circuits not shown provide background reference signals to ensure that each row is selected often enough to prevent information loss.

Addressing the foregoing arrays is a simple enough matter of decoders and further pass-transistor gating of bit lines (Figure 31).

2.3 Microprocessors

We have come almost to the end of the road. A consideration of structures at the microprocessor level of complexity exceeds the scope and purpose of our exposition, while at the same time the reader's interest naturally begins to shift toward the larger questions of functionality and architecture. Yet one important point remains to be made: MOS microprocessors are little more than further combinations of the basic elements and arrays that we have already seen.

Figure 32 shows the composite mask set for one such microprocessor, the 16-bit OM1, a forerunner of the OM2 design beautifully presented by D. Yohannsen in Mead80. Several arrays can be seen in the interior: to the left, two 16 × 16-bit dynamic RWM subarrays, with address decoders above and below and read–write circuits on either side; to the right of center, a 16 × 16 array of data pathway offset gates for high-speed shifting; farther to the right, the arithmetic logic unit (ALU), with other PLA structures above and below. Drivers and driver/receivers are seen near the edges, as are the names of the designers—Mead, Carroll, Johannsen, and Matsumoto—which are registered as minute engravings on the polysilicon level.

The management of complexity in such large circuits is an enormous task, one that will demand new approaches as the scale of integration increases [Suth77, Seit79]. New modes of chip design and implementation are being sought. In a recent report, L. Conway, A. Bell, and M. E. Newell [Conw80] describe the production of a multiproject chip whose designs were the result of 82 separate projects simultaneously undertaken at universities throughout the United States. Broadening the base in this way makes it economically feasible to try a variety of design experiments, and we can expect the emergence of many novel microprocessing structures to add to the list of important silicon systems.

CONCLUSION AND PERSPECTIVE

There you have it: out of the ceaseless, irregular bustling of jittery electrons, the statistical orderliness of controlled circuit flow; out of the happy coincidence of just the right properties of matter and the innate inventiveness of man, a minuscule but powerful apparatus to serve the methodology of logic. It all hangs by a circumstantial

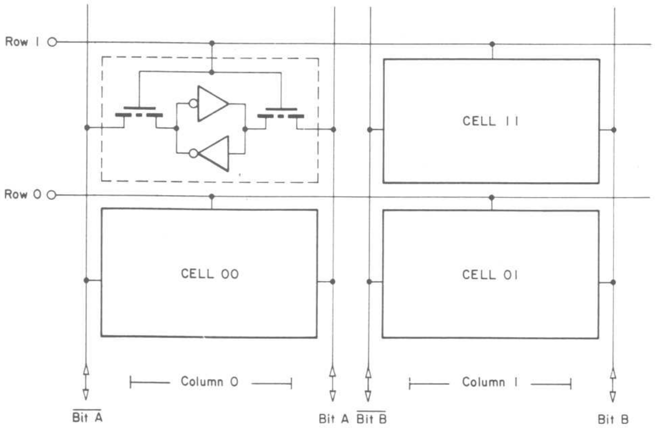
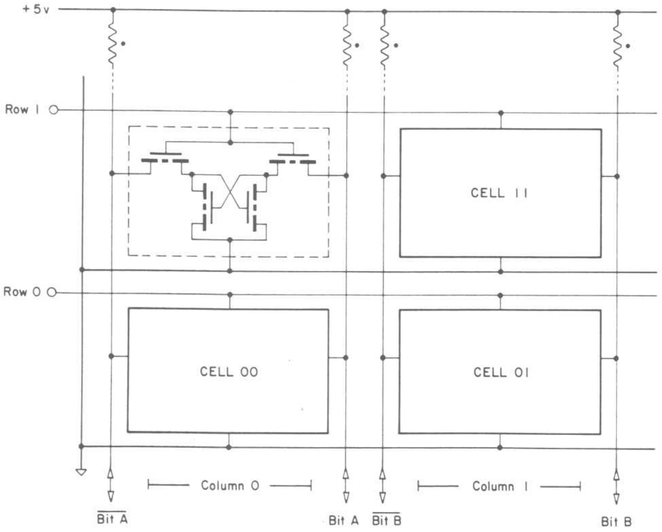

Row 1
Row 0
CELL 11
CELL 00
CELL 01
Column 0
Column 1
$\overline{\text{Bit A}}$
Bit A
$\overline{\text{Bit B}}$
Bit B

FIGURE 29. Static read-write memory (RWM) array. In addition to row-select and bit in/out lines, 0- and +5-volt lines (not shown) connect all cells to operate the inverter pairs.

FIGURE 30. Dynamic read-write memory (RWM) array. Occasional high signals on the row-select lines refresh information. Background row-select signals are provided by circuits not shown.


+5v
Row 1
Row 0
CELL 11
CELL 00
CELL 01
Column 0
Column 1
$\overline{\text{Bit A}}$
Bit A
$\overline{\text{Bit B}}$
Bit B

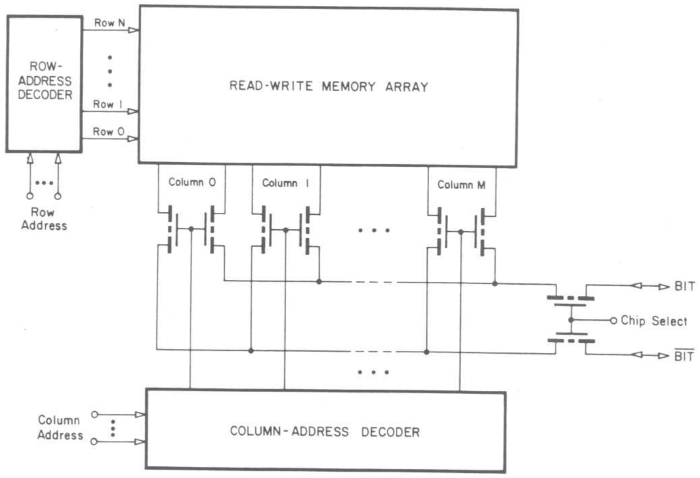


FIGURE 31. Addressing an $M \times N \times 1$-bit memory array.

chain: silicon just happens to have an accommodating lattice structure that can be altered in useful ways; insulating glass just happens to bond well to silicon; some chemical etchants just happen to eat glass while others eat nonglass; the right sort of aluminum just happens to bond well to both silicon and glass, the right sort of gold to aluminum.

In learning how to create and use the tiny no-man's land between opposed regions of enriched semiconductive material, we have contrived to build practical devices of extraordinarily small size. In learning how to partition and distribute electron population pressures, we have developed a set of techniques for controlling the movement of charge in minute physical structures. A grasp of control principles has given us the ability to analogize increasingly complex expressions of logic, from the "and" and "or" of propositional calculus to the rich machinery of algorithmic systems. The family of microprocessors lives on top of this growing peak of technical achievement.

The tools of fabrication are costly and elaborate marvels of technology: crystal growers, orienters, slicers, and polishers; diffusion furnaces, vacuum deposition chambers, and ion accelerators; high-precision cameras with ultrastable indexing mechanisms; micromanipulators and microwelders; injection molders; inspection and test equipment; accounting departments.

The integrated circuits themselves are amazingly cheap and almost too plentiful. Through the sorcery of a very advanced technology we will soon be able to produce powerful circuit modules in quantities far greater than we can assimilate. And remember, *the integrated circuit revolution has run only half its course*! We have much to learn and much to keep in careful perspective. Yet with the relentless quickening of the pace of development and the almost inevitable decline in our ability to master our most complex artifacts, will not we sorcerer's apprentices make too many brooms anyway? What are we going to do with them all?

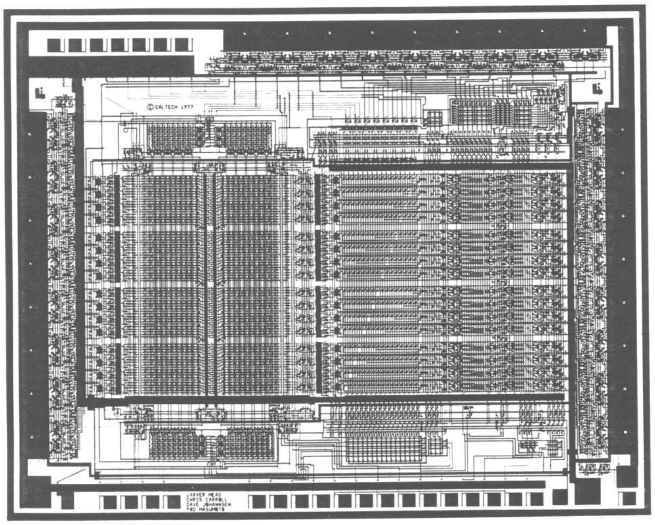


FIGURE 32. Composite n-MOS photomask set for the OM1, a 16-bit microprogrammable microprocessor. [Photograph courtesy of Carver Mead.]

ACKNOWLEDGMENTS

I am indebted to David Aspinall for his invitation to participate in the Advanced Course on Microprocessors, and for the challenge and fun of introducing integrated circuit technology to a gathering of exceptional students with diverse backgrounds in computer science. In an auspicious early discussion, Amar Archbold, a man conversant in many idioms, suggested the title and therewith established my theme. I happily credit whatever range and accuracy I have achieved to the stimulus, criticism, and guidance proffered with characteristic generosity and enthusiasm by Ivan Sutherland and Carver Mead of the California Institute of Technology, where I had the good fortune to complete my notes for the course record as a visitor under the Sherman Fairchild Distinguished Scholars program. I have also had the advantage of discussing various aspects of presentation with Charles Molnar and Charles Seitz. In bringing the paper to its final form, I have benefited greatly from the advice and help of Douglas Clark and the editorial encouragement of Adele Goldberg. To all these good scientists and scholars goes my deepest gratitude.

This paper is based on lectures given by the author at the Advanced Course on Microprocessors, University College of Swansea in the University of Wales, September 1977. With minor alteration, it is reprinted here from the course record, *The Microprocessor and Its Application*, D. Aspinall, Ed., Cambridge University Press, Cambridge, England, 1978, with the kind permission of Cambridge University Press.

REFERENCES

ASPI78	ASPINALL, D., ED. *The microprocessor and its application*, Cambridge University Press, Cambridge, England, 1978.
BARD48	BARDEEN, J., AND BRATTAIN, W. H. "The transistor, a semiconductor triode," *Phys. Rev.* **74** (1948), 230.
CARR72	CARR, W. N., AND MIZE, J. P. *MOS/LSI design and application*, McGraw-Hill, New York, 1972.
CATT66	CATT, I. "Time loss through gating of asynchronous logic signal pulses," *IEEE Trans. Elect. Comput.* **EC-15** (Feb 1966), 108–111.
CHAN73	CHANEY, T. J., AND MOLNAR, C. E. "Anomalous behavior of synchronizer and arbiter circuits," *IEEE Trans. Elec. Comput.* **EC-22** (Apr. 1973), 421–422.
CHEN75	CHENG, E. K., AND MEAD, C. A. "Single-chip cursive character generator," in *Conf. Digest 32*, ISSCC Symposium, Philadelphia, Pa., vol XVIII, 1975.
CONW80	CONWAY, L., BELL, A., AND NEWELL, M. E. "MPC79: The large-scale demonstra-

tion of a new way to create systems in silicon," *Lambda* **1,** 2 (Q2, 1980), 10–19.

COUR75 COURANZ, G. R., AND WANN, D. F. "Theoretical and experimental behavior of synchronizers operating in the metastable region," *IEEE Trans. Elec. Comput*, **EC-24** (June 1975), 604–616.

GRAY63 GRAY, H. Y. *Digital computer engineering*, Prentice-Hall, Englewood Cliffs, N.J., 1963, pp. 198–210.

GROV67 GROVE, A. S. *Physics and technology of semiconductor devices*, Wiley, New York, 1967.

HOLT77 HOLT, A. W. "Research on information system specification," *Rep. CADD-7708-0911*, Computer Associates, Inc., Wakefield, Mass., Aug 1977.

HURT75 HURTADO, M. *Dynamic structure and performance of asymptotically bistable systems*, D.Sc. dissertation, Dept. of Electrical Engineering, Washington Univ., St. Louis, Mo., 1975.

LILI28 LILIENTHAL, J. E. U.S. Patent no. 1900018, filed March 1928. (Compare Heil, O., British Patent no. 937457, 1935.)

LITT66 LITTLEFIELD, W. M., AND CHANEY, T. J. "The glitch phenomenon," *Tech. Memo. No. 10*, Computer Systems Lab., Washington Univ., St. Louis, Mo., Dec 1966.

MARS68 MARS, P. "Study of the probabilistic behavior of regenerative switching circuits," *Proc. IEE* (London) **115** (May 1968), 642–668.

MEAD80 MEAD, C., AND CONWAY, L. *Introduction to VLSI systems*, Addison-Wesley, Reading, Mass., 1980.

SEIT79 SEITZ, C. L. "Self-timed VLSI systems," in *Proc. Caltech Conference on Very Large Scale Integration*, Pasadena, Calif., Jan. 1979.

SEIT80 SEITZ, C. L. "Ideas about arbiters," *Lambda* **1,** 1(Q1, 1980), 10–14.

SHAN38 SHANNON, C. E. "Symbolic analysis of relay and switching circuits," *AIEE Trans.* **57** (1938), 713–723.

SHOC49 SHOCKLEY, W. "The theory of P-N junctions in semiconductors and P-N junction transistors," *Bell Syst. Tech. J.* **28** (1949), 435.

SUTH76 SUTHERLAND, I. E., MEAD, C. A., AND EVERHART, T. E. "Basic limitations in microcircuit fabrication technology," *Rep. R-1956-ARPA*, Rand Corp., Nov. 1976.

SUTH77 SUTHERLAND, I. E., AND MEAD, C. A. "Microelectronics and computer science," *Sci. Am.* (Sept. 1977), 210–228.

WOLF76 WOLFF, M. F. "The genesis of the integrated circuit," *IEEE Spectrum* (Aug. 1976), 45–53.

RECEIVED DECEMBER 1979; FINAL REVISION MARCH 1980; ACCEPTED MAY 1980