Characteristics of faults in MOS arrays

by H. R. LAMBERT

North American Rockwell Corporation
Anaheim, California

INTRODUCTION

Before discussing characteristics of faults in MOS arrays, it appears desirable to review briefly MOS technology, and to present some discussion on the use of MOS devices in arrays. Not only will this benefit those not too familiar with MOS devices, but it will also serve as a basis for the subsequent discussion on failure modes and mechanisms.

Metal-Oxide-Semiconductor (MOS) device fabrication employs the same basic technologies used in the fabrication of bipolar devices. However, MOS devices are quite different from bipolar devices in operation. To understand some of the problems that may be encountered in MOS devices, it is thus necessary to know something of their construction and operation.

Figure 1 shows a schematic cross-section of the type of MOS field effect transistor (FET) that is the basic element for most MOS arrays. This figure also shows the symbolic representation for a MOSFET. This type of MOSFET, called a P Channel Enhancement MOSFET, is made by diffusing two P regions into an N type silicon substrate. A metal film control element, called a gate, is evaporated onto an insulating film over the region separating the two P regions. It is from this feature that the term Metal-Oxide-Semiconductor is derived. This can be somewhat of a misnomer, however, as the insulating layer between the gate and the substrate need not be silicon oxide. It may be silicon nitride, or a silicon oxide/nitride combination. This latter type of device may be called an MNOS device, for Metal-Nitride-Oxide-Semiconductor.

If the Source P region of a P Channel device is grounded with the substrate, as shown in Figure 2, and negative voltage is applied to the Drain P region, there will be no appreciable drain to source conduction as long as the gate is unbiased. This off state has the megohm impedance of a reverse biased P–N junction. When negative voltage is applied to the gate, minority (positive) carriers are attracted to the surface of the substrate that is covered by the gate. At the same time, majority (negative) carriers are repelled from the surface. When sufficient negative gate voltage has been applied, the N type material under the gate will "invert" and a P type path or channel is formed between the drain and source. The gate voltage at which this occurs is called the threshold voltage and is generally defined as the gate voltage required to allow a specified drain to source current at a specified drain to source voltage. As the gate voltage is made still more negative,

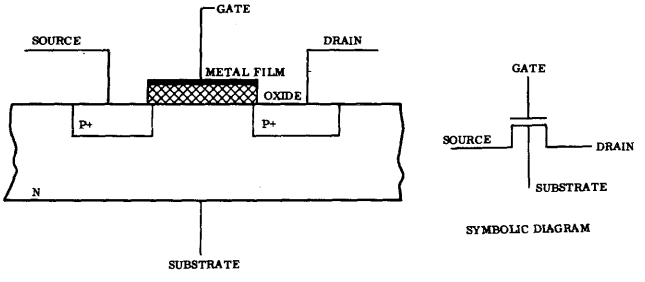

Figure 1—Cross section of P channel enhancement mode MOSFET

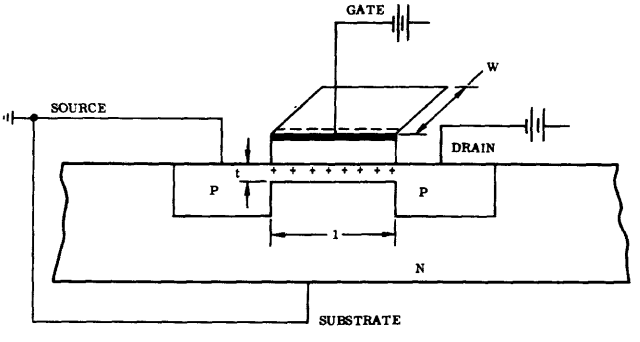

Figure 2—Biasing and operation of MOSFET

the induced channel is deepened and the drain to source conductivity is increased. This increase of conductivity by increase of gate voltage is referred to as enhancement mode operation. When the gate voltage level is subsequently reduced below the threshold voltage, the channel reverts to its initial N type and the device no longer conducts. The control of conductivity by an electric field gives rise to the descriptive name of field effect transistor. Because gate is separated from all other portions of the device by an insulating layer, there is essentially no gate current required for operation.

It is also possible to fabricate N channel devices by diffusing N regions into P type substrates and to make devices with a thin diffused channel between the drain and source. This latter type results in a normally on device in which the drain to source conduction can be either increased (enhanced) or decreased (depleted). Such a device is usually called a depletion type device to distinguish it from the above described enchancement mode devices. Junction type FETs are also fabricated, but will not be discussed because of their current unimportance in integrated circuitry.

Figure 3 shows what are commonly called drain characteristics, in order to present some idea of the electrical characteristics of MOS devices. This figure plots the variation of drain-to-source current with gate and drain voltages for a typical P channel enhancement device. Other typical electrical characteristics are threshold voltages of 3–5 volts and on resistances of on the order of 100 ohms.

More should be said about on resistances at this point, as this is one of the characteristics of MOSFETs which makes them particularly applicable to integration into arrays. The drain to source on resistance is determined by the common resistance formula:

$$R = \rho \frac{L}{Wxt} \quad (1)$$

L and W are the channel length and width; t is the thickness and ρ is the resistivity of the induced channel. A MOSFET fabricated with a long narrow channel has the characteristics of a load resistor when turned on. This technique is utilized in the fabrication of MOS integrated circuits to eliminate the area consuming diffused resistors required on bipolar integrated circuits.

The switch like characteristics of MOSFETs makes them ideally suited for digital applications. Because of low gate current requirements, MOSFETs can be made very small. Good isolation between elements allows dense packing of MOSFETs in integrated circuits. These features result in the ability to fabricate very complex MOS digital arrays on very small dice. As an example of the degree of complexity possible, Autonetics has fabricated 1024 bit shift registers containing over 6000 MOSFETs on 160 by 135 mil dice. Table I lists some of the type of MOS arrays currently available from semiconductor manufacturers so that the impact of MOS on the computer industry can better be appreciated.

Although generalizations are difficult to make because of the many and varied types of MOS arrays available, for the purposes of this discussion an attempt will be made to categorize at least the logic arrays into static and dynamic types. The former use flip-flop type elements to store information as long as desired and are capable of d-c operation. Dynamic arrays, on the other hand, utilize the high input impedance characteristics of gates (10^{14} to 10^{16} ohms) to store information for brief periods of time. As an example, consider what happens when the voltage is removed from the gate of a turned on device. If the voltage is removed by grounding, the gate which acts like a capacitor is discharged and the device will be turned off. If, however, the voltage is removed by opening the gate line (as by turning off a series MOSFET) the gate will stay charged until the charge leaks off. As long as the gate stays charged, the device will remain turned on. In this manner it is possible to store information for brief periods while other operations are being performed.

This technique is illustrated in Figure 4 which shows one bit of a dynamic, multibit shift register. When an input zero is applied to Q1 coincident with clock pulse 1 (CP_1), the gate of Q5 will be charged to -18V from V_{DD} (through Q2 and Q3). Q5 will stay turned on after Q_2 and Q_3 are turned off by CP_1 returning to ground. When CP_2 turns on Q_4 and Q_6 any charge on Q_7 will discharge

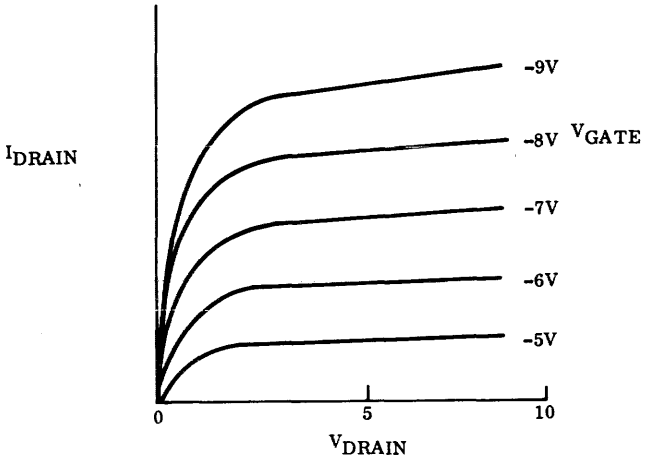

Figure 3—Drain characteristics of MOSFETS

Table I—Typical types of MOS circuit currently available

1—Shift Registers
- Static and dynamic registers ranging up to Quad 256 bit registers.

2—Memories
- Read only memories, nondestructive, to 1024 bits.
- Random access memories to 32 bits.

3—Converters
- D to A converters.
- A to D converters.

4—Counters
- Binary up-down counters.
- Variable modulus counters.
- Binary, ripple carry type counters.

5—Flip-Flops
- Dual J-K flip-flops.
- RST flip-flops.

6—Gates
- Various types of Dual and Quad multi input NAND/NOR, AND, and OR/NOT gates.

7—Miscellaneous
- Parallel accumulator.
- Full adder.
- Parity detector.
- Level shifter.
- Counter timer.
- Clock generator.
- Address decoder.
- Serial/parallel—parallel/serial converter.
- Other specialized types of arrays.
- Numeric character generator.
- Arithmetic comparator.
- Servo Adder.
- Frequency divider.
- Decimal point display.
- Digital differential amplifier.
- Core memory interface.

to ground through Q_5, and a zero will have been propagated from the input of Q_1 to the input of Q_7 by the subsequent application of CP_1 and CP_2 pulses.

Although input gates of MOS arrays are generally protected from static electric discharges by low voltage diodes, the high input impedance of internal devices if not affected. The lower frequency limit of operation of dynamic arrays depends on the RC time constant of the internal gates and is on the order of 1 kHz.

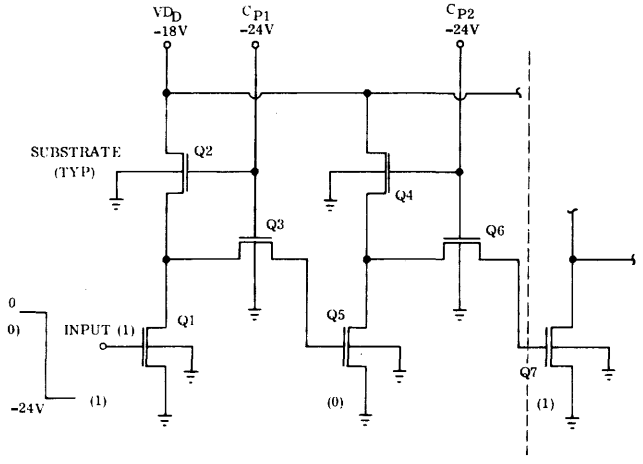

Figure 4—One bit of typical shift register

Discussion of fault characteristics

It was previously noted that the same basic methods are used for fabricating both MOS and bipolar devices. These include crystals growing, wafer preparation, oxide formation, masking and etching, diffusion, die and wire bonding, and packaging. The materials used for the substrates and doping are also similar. It would be correctly expected that processing and material related failure mechanism on both types of devices would be similar. The justification for discussion of MOS fault characteristics, when so much has been written on those of bipolar devices, is based on degree and manner that MOS devices are affected. These different degrees and manners are a result of the field method of operation of MOS devices and the necessary dimensional differences

required by MOS devices for this type of operation. Discussion will be confined to fault characteristics related to these differences although tabular listings of all normally encountered failure mechanisms will be presented. An attempt will later be made to relate these discussed mechanisms to various types of MOS array malfunctions.

The discussion will confine itself to the types of faults that cause failure in services; that is, mechanism that cause apparently good arrays to malfunction at some subsequent time after they have been put into operation. These malfunctions, neglecting misapplication and overstresses, relate back to fabrication anomalies or design deficiencies, as such, they also relate to mechanisms responsible for yield loss during fabrication.

For the purpose of organization, the faults will be grouped into the following three categories.

1. Catastrophic faults that result in permanent complete device malfunction. Examples are opens for shorts caused by one or more of the conditions shown in Table II.
2. Intermittent faults that cause temporary device malfunction or that cause loss of any stored information. Examples are intermittent opens or shorts caused by one or more of the conditions shown in Table III.
3. Degradation faults that gradually cause complete device malfunction or malfunction under some conditions of operation. An example is leakage current degradation which could be caused by one or more of the conditions shown in Table IV.

Table II—Catastrophic failure mechanisms

Failure Indicator	Related Failure Mechanism	
Opens	1—Bonding	a. Poor wire bonding (weak, off pads, etc.).
		b. Intermetallics at pads.
		c. Damaged wires.
		d. Mechanical overstress (shock, etc.).
	2—Metallization	a. Unopened contact windows.
		b. Breaks in the metallization
		c. Handling damage to metallization.
		d. Metal migration.
	3—Mechanical Damage	a. Cracked or broken die.
	4—Shorting	a. Excessive current drain at short and elements subsequently burned open.
Shorts	1—Oxide defects	a. Metallization penetrates oxide.
	2—Metallization	a. Metallization lines touching as a result of mask defects.
	3—Bonding	a. Bonding wires touching.
	4—Mechanical Damage	a. Cracked die.
	5—Electrical Over Stress	a. Oxide rupture.
		b. Excessive power drawn causing melting of elements.
	6—Foreign Material	a. Bridging between elements.

Table III—Intermittent catastrophic failure mechanisms

Failure Indicator	Failure Mechanism	
Opens	1—Bonding	a. Loose bond, die
	2—Mechanical Damage	a. Cracked die.
Shorts	1—Bonding	a. Wires touching each other or die.
	2—Mechanical Damage	a. Cracked die.
	3—Foreign Material	a. Loose foreign material shorting between elements.

Table IV—Degrading failure mechanisms

Failure Indicator	Failure Mechanism	
High Leakage Currents	Contamination	a. In the oxide. b. On the oxide. c. On the package.
Shifted Threshold Voltage	Contamination	a. In the oxide. b. At the oxide—silicon interface.

The failure mechanisms listed in Tables II, III and IV are common, except to degree and effect, to both MOS and bipolar devices. Some mechanisms are dependent upon die and/or diffused region areas, others are dependent on the number of internal wire bonds used, still others are dependent on packing densities of elements and spacing of metallization lines. Some mechanisms are also related to the particular structure and operation of MOS devices. Table V illustrates these various dependencies.

As an example of what is meant by mechanisms being dependent on various geometric factors, consider the effect of die area and diffused (active) area on devices with oxide defects. The die area obviously does not cause oxide defects; however, if there exists a certain density of oxide defects per square cm, then statistically a large die is more likely to have defects on it than a small die. Similarly, a die with closely packed elements or large diffusion areas is more apt to have oxide defects in critical locations than the same size die with a lower packing density and smaller diffused areas. These geometrical considerations indicate that it is reasonable to expect a large, complex array to have a higher failure rate than a small, simple array. This would normally be offset by system requirements for fewer of the more complex arrays, with the resultant fewer interconnect and mounting problems. At any rate, geometrical effects are not peculiar to MOS devices and will not be discussed further.

Table V indicates that oxide defects, surface contamination, and metal migration are all influenced by the MOS structure and method of operation. Experience has also indicated that these three mechanisms are the largest causes of MOS failures under stress. These mechanisms and their relation to MOS devices will be discussed below.

Oxide defects can be holes, cracks, or weak spots in the oxide insulating layer. If these defects occur under

Table V—Failure mechanisms and dependency factors

Failure Mechanism	Dependent Factors					
	Geometry				MOS	
	Die Area	Diffusion Areas	Packing Density	No. of Bonds	MOS Structure	MOS Operation
Bond discrepancies				X		
Metallization defects			X			
Metallization migration			X			X
Cracked or broken dice	X					
Oxide defects	X	X	X		X	
Foreign material			X			
Surface contamination	X					X

interconnect lines whose metallization subsequently penetrates the defect, the interconnect line becomes shorted to the underlying element or substrate. Most oxide defects appear to be caused by surface irregularities present on the silicon prior to oxidation.[1] Assuming a uniform distribution of silicon surface irregularities, more oxide defects will occur in regions of thin oxide and at abrupt oxide steps than in uniform regions of thick oxide. The threshold voltage of MOS devices is governed by the following relationships:[2]

$$V_T = (Q_{SS} + Q_{BB}) \frac{t_{ox}}{e_{ox}} \quad (2)$$

where t_{ox} is the oxide thickness, e_{ox} is the dielectric constant of the oxide, and Q_{SS} and Q_{BB} are the charge densities at the silicon-silicon oxide interface and in the substrate respectively. In order to obtain thresholds in the usable range of 3 to 6 volts, it is necessary to have a gate oxide thickness on the order of 1500A°. In comparison, the thinnest oxide normally found on bipolar devices is in the 5000A° to 10,000A° range. There is also an oxide step around the periphery of the gate oxide where the oxide thickens to several thousand Angstroms over the P regions. Inasmuch as the gate oxide and surrounding step are covered by metallization any oxide defects in these regions are potential shorts between the gate metallization and underlying substrate and P regions. The effect of such shorts will be discussed later in the paper.

Contamination in and on the insulating layer or at the silicon-silicon oxide interface can result from impurities in the various cleaning and etching solutions or volatile impurities in the oxidation and diffusion furnaces. The various exposures to high temperatures during fabrication tend to uniformly distribute the contamination ions and the effect of their presence is usually not detectable by just electrically testing the completed device. Small amounts of contamination can, however, result in considerable device performance degradation in subsequent use.

The operation of MOSFETs by application of an electric field at the gate was discussed in the introduction. The formation of a conducting drain to source channel by inversion of the substrate surface under the gate was described. This inversion was seen to be caused by attraction of minority carriers and repulsion of majority carriers by applied gate voltage. The gate voltage required to initiate drain to source conduction across the inverted region was defined as the threshold voltage.

Any spurious charge concentrations under the gate affect Q_{SS} (reference equation 2), thereby, inhibiting or enhancing the channel formation process and shifting the threshold voltage. If gate bias is applied at ambient temperatures above 100°C mobile contamination ions become concentrated at the substrate surface as shown in Figure 5. This figure shows the tendency of positive contaminant ions to increase threshold voltages in P channel MOS devices by inhibiting the inversion process. It appears that the spurious positive ions result from contamination by sodium present in the various fabrication materials and equipment. Considerable effort is normally expended to reduce the level of sodium contamination to an acceptable level.

In addition to inhibiting channel formation, positive ions can cause surface inversion of P regions by attracting negative minority carriers and repelling positive majority carriers. In most types of MOSFET's there is some gate metallization overlap of the drain P region. If at high temperatures the drain is biased negatively in respect to the gate, positive charges will accumulate in the gate overlap region as shown in Figure 6. This will cause a distortion of the drain to substrate junction in this region giving rise to an increase drain leakage current when the device is in the off state.

The effects of increased threshold voltages and leakage currents will be discussed later.

Metal migration occurs when metal ions are knocked

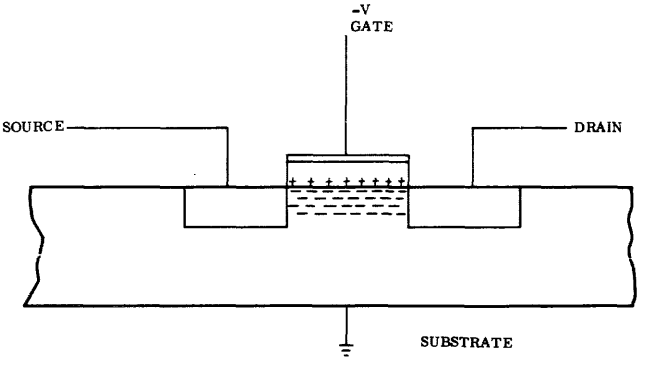

Figure 5—Accumulation of channel surface

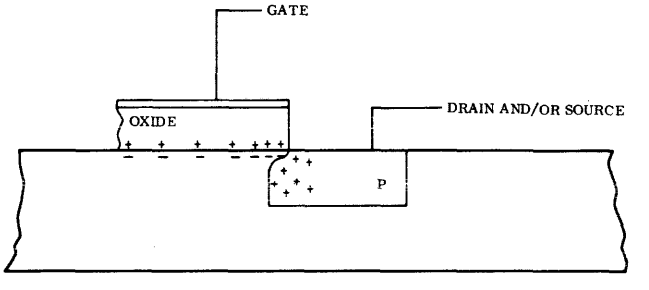

Figure 6—P region inversion

out of their crystal lattice by high velocity conduction electrons. The migration rate is accelerated by high temperatures and high current densities and is influenced by the crystal structure of the metal. Considerable investigation has been done on the migration of aluminum films such as are used for interconnects in integrated circuits.[3,4,5] The small cross sectional area of these interconnects can result in current densities of 10^5 to 10^6 amps/sq cm for only a few milliamps of current. By way of comparison, this current density is about two orders of magnitude higher than that which occurs in high power transformers. As elements become more densely packed in complex arrays, the interconnect lines must be made narrower. Care must be taken to limit current densities within safe limits by application of appropriate design rules.

The foregoing applies to any arrays employing aluminum interconnects, but a special problem exists on MOS arrays in lines used to charge large numbers of gates. Normally gate lines are considered to draw very little current. However, gate capacitance charging by fast rise time clocks gives rise to large instantaneous currents with resulting instantaneous current densities of 10^7 to 10^8 amps/sq cm. Although no significant relations have been obtained between these high instantaneous current densities and migration rates, conservative design rules must be employed to anticipate the potential problem.

Previously it was noted that the three failure mechanisms most influenced by MOS structure and operation are the same ones that are responsible for most MOS array malfunctions under stress or use. These mechanisms result in gate shorts, threshold and leakage increases and open interconnections. The question to be examined is how such discrepancies can be expected to affect operation of typical digital circuits.

The function of all types of digital arrays is to store, read out or perform some operation on digital information at the command of some input. The resultant digital outputs are in the form of "ones" and "zeros." When an array malfunctions it can be expected that the outputs will be in error in one of the following ways:

1. Outputs are all "ones," all "zeros" or some intermediate level for any input or operation.
2. Outputs that should be "ones" are "zeros" and vice versa, either permanently or intermittently.
3. Outputs are correct but are degraded to the extent that other arrays will not correctly recognize them.

In some cases these malfunctions may be temperature, voltage or frequency dependent.

It can be expected that gate shorts and interconnect opens will cause permanent (as opposed to intermittent) arrays malfunctions such as 1 or 2 above. These malfunctions will in addition be quite insensitive to temperature, voltage or frequency variations about their nominal values. The location at which the discrepancy occurs will govern the type of malfunction. A short in one place may cause an all "zero" output while a short at another place may cause an all "one" output. In addition a short at one location may produce the same effect as an open interconnect at a different location. Certain types of circuits may have weak areas in fixed locations so that devices tend to fail in the same manner, causing repetitions of the same circuit malfunction.

If shorts are responsible for the array malfunction, loading down of clocks, inputs or suppliers may occur. This loading may be to the extent that other good arrays fed from these sources may not function properly. Open interconnects will generally not cause such loading down, and in fact may reduce the normal load.

Arrays with degraded threshold voltages or leakage currents may exhibit any of the three above listed malfunctions at nominal temperature, voltages and frequency. However, these malfunctions will generally be sensitive to variations of temperature, voltages and frequency about the nominal values, as will later be discussed.

It was shown in Figure 3 that for a given drain-to-source voltage, increasing the gate voltage above its threshold results in an increase of drain to source current. That is, the channel resistance is reduced. For a constant gate voltage, as the threshold voltage increases, the gate is turned on less hard and the channel resistance increases. Thus, any gate charging and discharging through another device whose on resistance has increased as a result of threshold degradation will charge and discharge at a slower rate. The device will therefore turn on and off more slowly. If a number of threshold degraded devices are employed in series, as for example in a multibit register, the cumulative effect may impair high frequency operation of the array. Low frequency operation in which the devices are allowed enough time for this slower turn on and off will not be affected. At intermediate frequencies intermittent operation and incomplete transitions between the "one" and "zero" levels may occur. Devices affected by degraded thresholds will not be greatly sensitive to temperature or voltage variations.

As shown in Figure 4 devices such as Q3 and Q6 are used to control the gate charging of other MOSFETs. When the controlling devices are turned off, any charge present on the gates of MOSFETs they control will be stored until it leaks off. This means that the controlled

device will stay turned on as long as its gate stays charged. Information may be stored in this manner. The length of time a gate will maintain its stored charge under such type of operation determines the lower cutoff frequency of operation. If there is an enhanced charge leakage path through the control device as a result of P region inversion, the gate being controlled will loose charge faster than intended and low frequency operation of the device and array will be impaired. At higher frequencies charges are required to be stored for shorter durations and array operation may be normal. Low frequency operation may be improved by increasing the voltages to the array as the gates are then charged to higher levels and a longer time is required for them to discharge. Operation at high frequency may be impaired by elevated ambient temperature which increase leakage currents. At frequencies, voltages and temperatures intermediate between those causing normal operation and inoperation, outputs may become degraded so that "ones" appear only as sharp spikes. Such degraded "ones" may not be recognized by other arrays into which they feed.

The leakage currents on an array may degrade to the extent that considerable loading down of signals occur. In addition enough power may be consumed by the array to significantly increase its temperature. As leakage current increases with temperature a thermal run-away condition may occur with the resultant destruction of the device. Even if thermal runaway is not encountered, the array temperature may increase enough to further impair low frequency operation.

It is characteristic of contamination associated failure mechanisms that the degradation introduced over a period of time in use can be baked out at 200°C in only a few hours time.

It should be remarked that some oxide defects may cause high resistance leakage paths which will have the same affect on device performance as high leakage currents resulting from contamination. It is expected that such type failures would be less temperature sensitive, and could not be returned to normal operation by baking.

SUMMARY

The fabrication techniques and mode of operation cause oxide defects, contamination, and metal migration to be the dominant failure mechanisms of MOS devices.

Shorts caused by oxide defects and opens caused by metal migration give rise to permanent voltage, frequency and temperature insensitive array malfunctions. Array outputs may be all "ones," all "zeros," or outputs that should be "ones" may be "zeros" and vice versa.

Contamination, notably positive sodium ions, cause thresholds and/or leakage currents to increase. High frequency operation is affected by the former and low frequency operation by the latter. The latter is also sensitive to high temperatures and low voltages. Outputs on contaminated devices may be similar to those for opens or shorts or may be degraded correct outputs.

REFERENCES

1 J E MEINHARD
Process techniques study of integrated circuits interim scientific
Report No C5-147/ 12/501
Prepared under NASA contract NAS 12-4 May 1968

2 P J BESSER J E MEINHARD P H EISENBERG
Factors influencing dielectric defects in silicon oxide layers
Presented at Electro Chemical Society Philadelphia Pa October 1966 Autonetics Report # X6-3072/501

3 J R BLACK
Mass transport of aluminum by momentum exchange with conducting electrons
IEEE Sixth Annual Reliability Physics Symposium Proceedings 1967 Vol. 6

4 I B BLECH H SELLO
The failure of thin aluminum current carrying strips on oxidized silicon
Physics of Failure in Electronics Vol 5 1966

5 R V PENNY
Current induced mass transport in aluminum
Journal Physics of Chemical Solids Vol 25 1964