1972 WESCON Technical Papers

Volume 16

Western Electronic Show and Convention

Papers presented at the
Western Electronic Show and Convention in
Los Angeles, California, September 19-22, 1972

Schiff, L.	WES-16-20/2-72	Stockebrand, Tom	WES-16-17/4-72
Schmidt, W.G.	WES-16-28/1-72	Suffield, F. G.	WES-16-3/2-72
Schnathorst, V.	WES-16-26/2-72	Terhorst, Kas	WES-16-13/2-72
Schwartz, Morton D.	WES-16-12/1-72	Tickle, Andrew C.	WES-16-4/2-72
Semon, W.L.	WES-16-8/4-72	Tkal, Oleh	WES-16-2/2-72
Sepe, Raymond B.	WES-16-14/1-72	Topol, Sidney	WES-16-22/2-72
Shopbell, M.L.	WES-16-19/3-72	Valentine, Don	WES-16-18/1-72
Shoshani, Arie	WES-16-7/4-72	Van De Water, J. M.	WES-16-12/4-72
Silverman, Paul B.	WES-16-14/4-72	Walker, A.C.	WES-16-28/3-72
Singleton, J.B.	WES-16-14/3-72	Walsh, William J.	WES-16-9/2-72
Skornia, Thomas A.	WES-16-18/5-72	Wanlass, Frank M.	WES-16-4/2-72
Smith, Lowell W.	WES-16-13/3-72	Wasserman, Jerry	WES-16-22/4-72
Sontag, Frederick B.	WES-16-11/2-72	Wheeler, P.C.	WES-16-27/2-72
Sorensen, A.A.	WES-16-27/2-72	Young, Leo	WES-16-11/3-72
Staras, Harold	WES-16-20/2-72	Zangrando, Don	WES-16-25/4-72

ELECTRICALLY ALTERABLE NON-VOLATILE SEMICONDUCTOR MEMORIES

Andrew C. Tickle
Nitron Corporation, 927 Thompson Place, Sunnyvale, California 94086

Semiconductor memories are claimed to be able to exceed the performance of magnetic memories, which they are steadily replacing, on almost every parameter, whether it is speed, cost, density or reliability. However, in applications where it is necessary to store electrically alterable data without the continuous drain of power, there has been, until recently, no substitute for magnetic memories, whether they be tapes, cores or films.

With the increasing use of semiconductor memories the need for a non-volatile version is becoming more acute. For example, in a computer using a large number of micro-instructions in a ROM, it is almost the rule that a modification is required during the prototype development. Unless an electrically alterable ROM is used, new memory components have to be built. The flexibility of computers using large amounts of ROM is greatly increased if EAROM's are used, since they need no longer be dedicated to a single purpose. They can even by electrically reconfigured, without incurring the cost of component replacement and wiring changes. Look-up tables can be routinely updated.

Non-volatile RAM's have the advantage of being able to retain data during a spurious power interruption. For battery operated and portable systems, power drain is greatly reduced since power may be disconnected when the system is not in use. For very large memory systems, non-volatile storage again provides advantages in power (and reduced heating problems) since power may be consumed only in the section of the memory being accessed.

Electrically alterable, non-volatile, semiconductor memories are now in production in both systems and component houses. This session reviews the two main technologies used and their application to memories and computer systems.

ELECTRICALLY ALTERABLE NON-VOLATILE SEMICONDUCTOR MEMORY TECHNOLOGY

Andrew C. Tickle and Frank M. Wanlass
Nitron Corporation, 927 Thompson Place, Sunnyvale, California 94086

SUMMARY

This paper describes some of the features and limitations of the technologies most commonly used in electrically alterable non-volatile semiconductor memories. Some new approaches are also described.

1. INTRODUCTION

Although semiconductor memories have been in widespread use for many years, it is only recently that they have been available in versions that will store electrically alterable data without power. Many such non-volatile devices[1-5] have been reported, but only two versions are in production. This paper reviews some of the features and limitations of these technologies.

Semiconductor devices capable of storing alterable, permanent, binary data all have two common features:

i) the use of MOS transistors modified so that charge may be stored in the dielectric so as to cause a permanent (but alterable) threshold change;

ii) a non-linear conduction mechanism (such as tunneling or injection) through a dielectric layer such that charge is effectively transferred only when a large or "full select" voltage is applied.

Storage devices fall into two classes depending on the storage mechanism:

i) charge is stored in deep energy states at the interface between two dielectrics in the gate structure;

ii) charge is stored in a floating electrode buried in the gate dielectric. The storage mechanism is analogous to an RC network with an extremely long time constant.

2. THE WRITING MECHANISMS

MNOS

The most commonly used Metal Nitride Oxide Silicon (MNOS) storage transistor uses a thin layer of oxide (20 to 60 Å) between the silicon nitride gate dielectric and the substrate, as shown in Fig. 1. The dielectric constant of the silicon nitride is two (or more) times that of the thin oxide, causing the applied field to be doubled in the oxide. This field enhancement and the small thickness permit tunneling of charge to the oxide conduction band once a suitable field has been reached. This charge is stored in traps at or near the oxide/nitride interface. Several years of real time storage have been demonstrated[6], and ten years may be reasonably predicted.

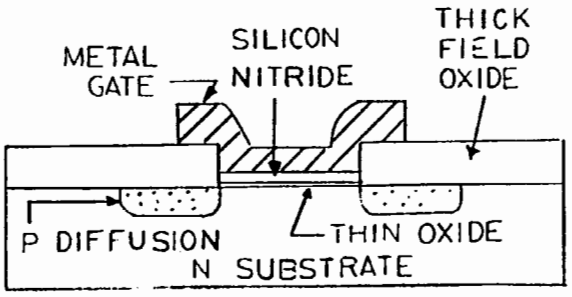

Fig. 1. Cross-section of a typical MNOS transistor

In the P channel storage array (Fig. 2) all MNOS transistors in a given row will experience a threshold shift when the negative writing voltage is applied to the gate, except where a negative voltage is applied to the sources and drains. When the writing voltage is applied to the gates, an inversion channel forms at the silicon surface, linking the source and drain electrodes. Since the channel forms an ohmic connection to the source and drain, it is at the same potential when no current flows (i.e., open circuited drain). Hence the source potential may be used to control the channel potential, and hence the potential difference across the gate dielectric when the writing voltage is applied to the gate. This is known as "channel shielding" and is the means by which binary data is selectively entered into an array during programming.

In order to reset the transistor back into the original state, the simplest method is to bias the substrate at a large negative voltage with the gates at ground potential so as to reverse the writing process. This does require that the gate drivers and substrate under the memory array must be isolated from each other to prevent the array gates from shorting to the substrate through the junctions of the array drivers. In this mode there is not selective "clearing" of the storage transistors; all are reset simultaneously. This is known as "block erase."

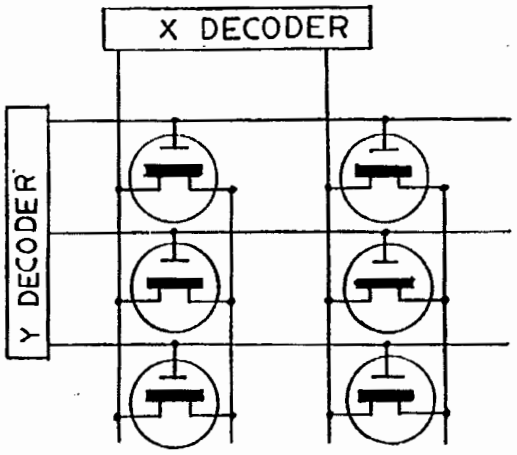

Fig. 2. MNOS memory array

Figure 3 shows the simplest configuration for interrogating and sensing the binary state of a MNOS storage transistor. The circle around the transistor symbol indicates that it is a storage transistor. If the interrogate voltage, Vr, applied to the gate is, for example, -8 volts and the transistor has threshold voltage of, say, -3 volts in the "zero" state and -9 volts in the "one state, then the output voltage will be either V_{dd} or ground when interrogated.

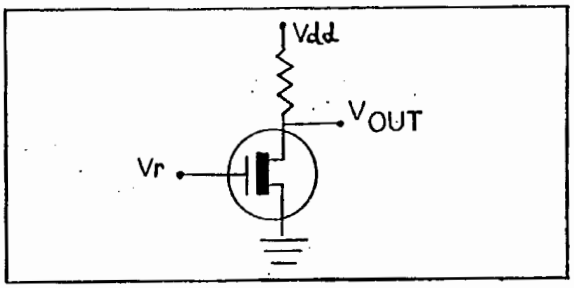

Fig. 3. Equivalent circuit of MNOS transistor during readout

The writing pulse length can be decreased, and the threshold shift increased, by increasing the writing voltage. However, excessive writing voltage causes an eventual degradation, or wearout, of MNOS storage transistors, reducing their ability to store charge. This is associated with the production of new surface states[7] which enable charge to tunnel through the thin oxide under zero bias. However, useful threshold shifts for EAROM's can be achieved over 10^6 or 10^7 writing cycles since longer duration, lower voltage pulses can be used. For a non-volatile RAM which may be required to survive 10^{13} or 10^{14} cycles, smaller threshold shifts have to be tolerated in order to live with small enough writing voltages to avoid wearout. In order to detect these small threshold changes, a sensitive readout method such as differential sensing (see Section 5) is required.

FAMOS

The FAMOS storage transistor (Fig. 4) has a floating silicon gate into which charge is injected through the gate oxide when the drain diffusion is in the avalance breakdown condition. Electrons are trapped in the electrically isolated gate, causing (for P channel devices) the transistor to be held permanently "on." There is no electrical method of removing the stored charge from the basic device, so ultraviolet or X-radiation have to be used to reset devices.

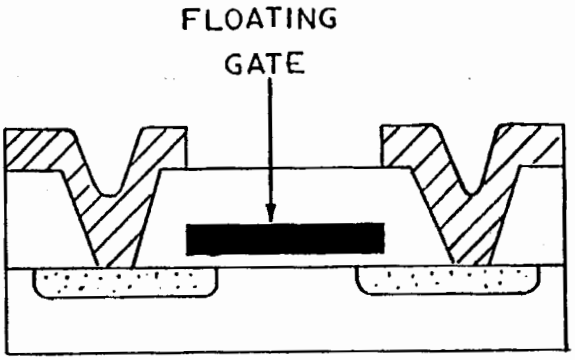

Fig. 4. Cross section of FAMOS type memory transistor

There is no natural mode of XY selection of the transistors, as with MNOS devices, so they are selected by an auxiliary transistor, as shown in Fig. 5. Compared with MNOS storage transistors, they are less versatile and require relatively high writing voltages. However, their simplicity and reliability have enabled them to reach quantity production sooner than MNOS.

3. PROBLEMS IN MEMORY CIRCUIT DESIGN

The basic problem in a non-volatile storage array is caused by the writing voltage levels. Of necessity, the writing voltage is larger than the normal logic levels of a MOS circuit, in order to produce the required threshold shift, and yet it is desirable to generate this voltage with MOS circuits on the same chip. Since either the power supply or clock supply (if used) in a convential MOS circuit has to be at least one MOS transistor threshold greater than the

output voltage, which is itself greater than a normal logic level, the writing voltage circuits are under considerably more stress than normal logic gates. This requires careful layout and processing to avoid drain breakdown, field inversion, and punch through conditions. A second problem is the need for the gates of the memory transistors to swing both positive and negative (for writing and clearing) with respect to the substrate. This problem has been solved in several ways.

Part of the substrate including either the gate drivers or the memory array may be isolated using, for example, diode isolation diffusions in an epitaxial layer (as used in bipolar integrated circuits). This permits the array to be cleared by applying substrate bias without this bias being shorted to the memory transistor gates through the drains of the gate drivers. This approach requires two additional processing steps to produce the isolation diffusion and the ohmic contacts to the top of the substrate. Since one extra mask is already required to produce memory devices, the process is considerably more complex than basic MOS processes.

A second method of dealing with the need for both polarities of voltage for writing and clearing is to avoid it by using a non-electrical clearing method such as ultraviolet irradiation of the array[3]. This simplifies circuit design and does not add any processing steps except for the addition of an ultraviolet window in the package.

A third method is to drive the gate lines of the memory array from circuits off the chip. The gate lines, of course, cannot be connected to any diode diffusions; otherwise, they will not support voltages of both polarities. This approach again simplifies the chip design and adds no processing steps. It also makes the memory completely clearable electrically. The problems, however, are shifted to assembly and system design. Gate protection devices cannot be conveniently made with the basic process, so precautions have to be taken in assembly and testing to avoid damaging electrostatic charges on the floating gate electrodes. Also, the decoding of the gate address is performed off the chip which requires more connections to the chip. The "pin count" may be minimized to some extent by deviating from the optimum "square" array and, instead, using fewer undecoded gate lines and increasing the number of address bits for the other axis. The source-drain lines carry only one voltage polarity and lower voltages than the storage gates, so they may be decoded on the chip.

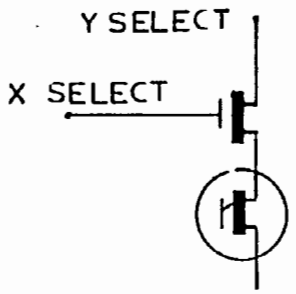

Fig. 5. Selection of a FAMOS transistor

4. OPERATING SPEED

The factors determining the speed of an EAROM may be considered most easily by comparing it with a conventional semiconductor ROM in which the pattern of stored data is represented by "missing" transistors in the array. Actually, the "missing" transistors are "field transistors" where the thick field oxide has deliberately not been removed from the gate region. The field transistor cannot turn on until the metal gate potential exceeds the field inversion voltage, which, of course, does not happen in normal circuit operation. This means that when a gate line is interrogated, (i.e., both thin oxide and field transistors on the array) effectively unlimited overdrive is permitted on the thin oxide transistors without turning on the field transistors. The resultant high output current permits rapid charging of the capacitive nodes and a high operating speed.

In an EAROM array the difference in the thresholds of '1' and '0' state storage transistors under the same gate line is typically much smaller than the difference between the thin oxide threshold and the field inversion voltage in a ROM. Hence, the available overdrive of the lower threshold transistors (without turning on those of the opposite state) is reduced and, consequently, also the output current and operating speed.

5. DIFFERENTIAL SENSING

One way in which relatively small difference in '1' and '0' thresholds may be overcome is by differential sensing. Each memory cell contains two transistors, one in the high threshold state and one in the lower state. The sensing circuit detects the polarity of the difference, the actual threshold values being of little importance. A sensitive differential amplifier will discriminate between '1' and '0' states with a much smaller threshold difference than the one

transistor per bit organization. The penalty for the two transistors per bit is not as severe as it might first appear when one remembers that in a volatile active MOS memory cell the minimum number of transistors per bit is three.

Since MOS circuits are not naturally suited to analog applications, the sense amplifier may be an external bipolar circuit. If this amplifier is operated in a current sensing mode, the speed can be increased since the need for substantial voltage swings is removed.

An ideal method of differential sensing on the chip is to use a flip-flop.[8,9] When power is applied to a flip-flop in the configuration shown in Figure 6, it will switch to a state determined by the voltage asymmetry at the inputs. This asymmetry is determined by the threshold difference in the pair of transistors in the selected cell. The flip-flop acts as a bidirectional trigger, and after each readout cycle its power supply must be removed so that voltages may be discharged to restore it to the sensitive condition. The voltage at one output will remain close to ground whilst the other output will charge up to the power supply voltage, less the load resistor drop, irrespective of the threshold difference sensed.

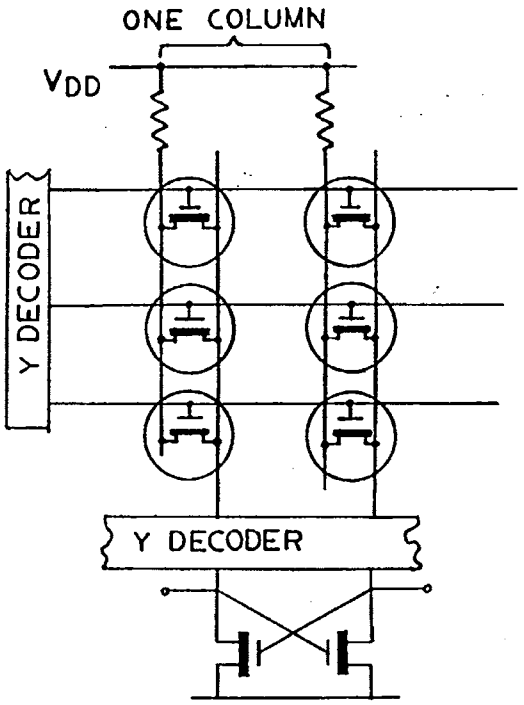

Fig. 6. Differential sensing with a flip-flop

The bidirectional trigger circuit also acts as an amplifier for the differential inhibit voltages during the writing cycle. When thin oxide MNOS storage transistors are used, the flip-flop is triggered during readout into the desired state to rewrite the same data. If the readout pulse on the storage transistor gate is allowed to continue to rise to the writing threshold, it will automatically rewrite into the higher threshold storage transistor (the one with the source grounded by the flip-flop). The other storage transistor of the pair is protected from the writing voltage by the inhibit voltage of the other flip-flop output.

6. N CHANNEL MOS

N channel MOS devices are inherently faster than P channel because of the higher mobility of electrons than holes. Also N channel circuits may be diffused with arsenic, which diffuses much more slowly than phosphorus, enabling smaller devices and higher packing densities. For example, changing from P channel to N channel increases the transconductance by about a factor of two. Reducing source drain spacings from 0.4 mils to 0.2 mils increases the transconductance by another factor of two. The effective length of the conducting channel of the transistor is actually less than the nominal length. Also, the reduction of devices' geometries reduces the capacitive loading. These two factors yield an additional increase in operating speed. Thus a high density N channel MOS process provides a valuable speed advantage to EAROM's, enabling their basically slower operation to be made compatible with normal RAM operating speeds.

However, N channel MNOS devices have a basic drawback in that the data-representing threshold shift is (as with P channel) mainly in the N direction. For P channel devices, an enhancement mode device (non-conducting at zero gate to source bias) is still enhancement mode after shifting the threshold. This enables large numbers of memory transistors to be organized in parallel in a memory array with only the selected device conducting when interrogated. With N channel MNOS storage transistors, the normal enhancement mode device (in the virgin state) becomes a depletion mode (conducting at zero bias) device after the threshold shift. To prevent the depletion mode storage transistors from conducting when their gates are grounded, they may be connected in series with enhancement mode MOS transistors with the gate common to the storage transistor as shown in Fig. 7. The double gate structure requires no extra processing steps and only slightly increases the packing density.

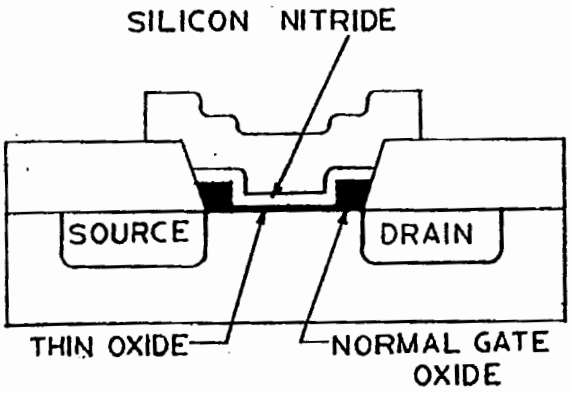

Fig. 7. N Channel MNOS with split gate

7. A NEW APPROACH TO THE HIGH VOLTAGE PROBLEM

Depletion Shielding

In the devices described in Section 1, channel shielding was used to prevent the full writing voltage from appearing across the gate dielectric, in order to prevent a threshold shift in the unselected transistors. In the transistor structure shown in Fig. 8a, selection is achieved by "depletion shielding." When a writing pulse is applied to the writing gate, such as to cause an inversion channel, the channel forms immediately if the necessary carriers can be supplied by the source or drain. However, if the source and drain are isolated by the control gate from the region under the writing gate, as shown in the equivalent circuit in Fig. 8b, then the necessary carriers to form the inversion channel will be supplied slowly by thermal generation in the substrate. The time required for this is similar to the carrier lifetime -- several microseconds. Thus for a few microseconds after the writing pulse is applied, the inversion channel will not have formed and the writing voltage will appear mostly across the thick depletion region under the writing electrode, and only a fraction of it will appear across the dielectric. As long as the writing pulse is removed before the channel forms (and the thick depletion layer collapses) no threshold shift will occur. In order to produce the threshold shift in a selected device during writing, the control gate has to be turned on to allow minority carriers to flow in and produce the inversion channel. This channel will be at the same potential as the source; namely ground potential, enabling the full writing voltage to appear across the gate dielectric. With N channel devices, for example, this is achieved by making the control gate appropriately more positive than the source.

Figure 9 shows XY writing selection of depletion shielded devices where the MOS transistor threshold is 1.5 volts. The writing gates of all transistors can now be common and pulsed unconditionally. This has the advantage that the high writing voltage does not have to be decoded on the chip. The voltage levels for controlling the depletion shielding are all normal logic levels. The common writing electrode supports

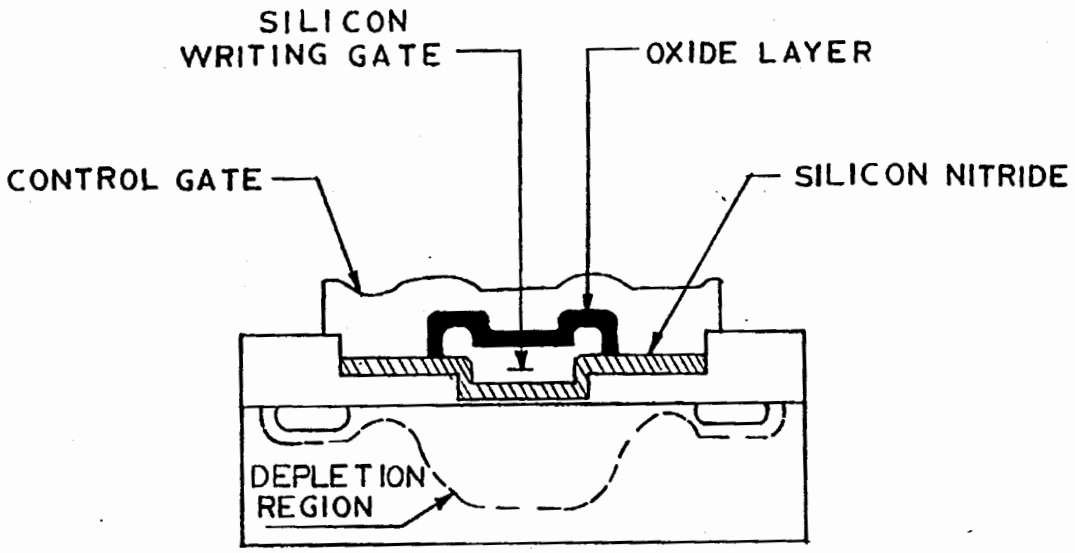

Fig. 8a. Depletion shielded writing mode

voltages of both polarities since it does not connect to any diffusions. Thus the "clear" function can be achieved without the need for electrically isolating different regions of the substrate. Readout selection is achieved by conventional XY selection, and a suitable bias may be applied to the writing electrode. Because of the shorter writing pulse used with depletion shielding, the threshold shifts are not as large as with conventional DC systems. Hence differential sensing is an ideal method for readout.

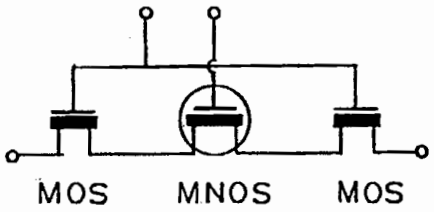

Fig. 8b. Equivalent circuit of a depletion shielded transistor

8. THE ELECTRICALLY ALTERABLE RESISTOR (EAR)

Figure 10 shows the configuration of an EAR. The gate is floating, as in the injection devices. However, charge is introduced by non-linear conduction along a short length of dielectric film, such as silicon nitride, embodied in a structure (Fig. 11) which can be external to the transistor.

The writing voltage is applied, in the normal manner, between the writing electrode and the substrate. The floating memory gate is the center tap of a capacitive voltage divider (Fig. 12). Since the gate to substrate capacitance is much greater than the stray capacitance at the overlap of the non-linear

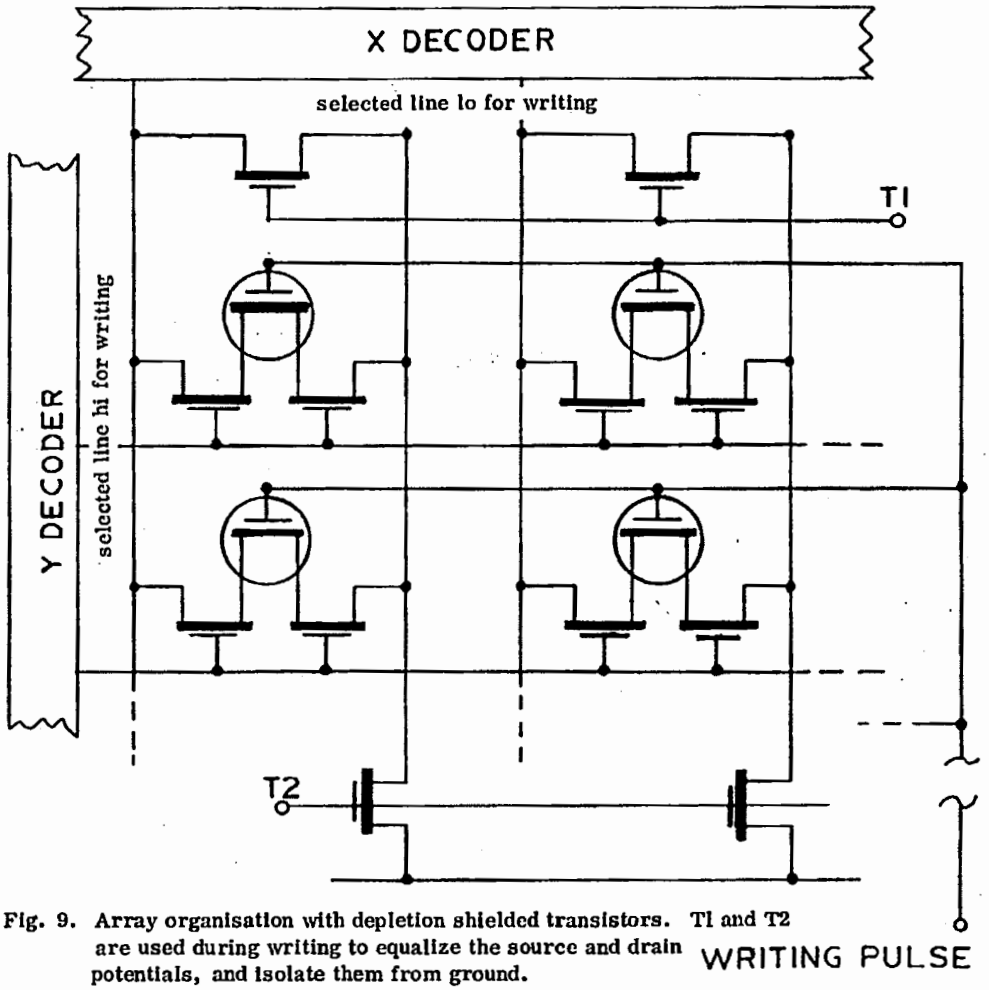

Fig. 9. Array organisation with depletion shielded transistors. T1 and T2 are used during writing to equalize the source and drain potentials, and isolate them from ground.

resistor, the gate is always close in potential to the substrate surface and the majority of the voltage drop is across the resistor when capacitive voltage division is the only mechanism in control. However, if the substrate surface potential is raised by channel shielding, as in the MNOS devices, the floating gate potential will follow it closely, reducing the potential difference across the non-linear resistor to below the critical value.

For binary storage arrays the EAR may be fabricated in a double gate configuration as shown in Figure 13 and an array organization with a common writing gate, similar to that shown in Fig. 9. As with the depletion shielded devices, the need for decoding high voltages on the chip is avoided.

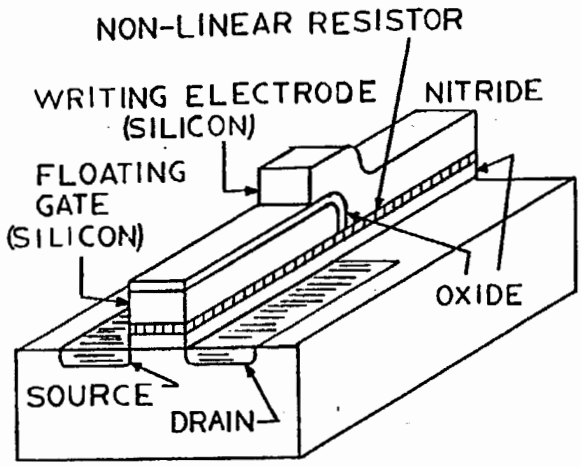

Fig. 10. Structure of the EAR

9. ELECTRICALLY RESETABLE AVALANCHE INJECTION TRANSISTORS

The active part of the EAR in Fig. 10 is the same as the FAMOS device of Fig. 4. However, instead of the gate being completely isolated it is connected to the writing electrode through the non-linear resistor. The floating gate may be charged by avalanche injection at the drain and discharged electrically by a clearing electrode, common to all storage transistors on an array.

10. APPLICATIONS

The most natural mode of use for storage transistors is as an electrically alterable read only memory (EAROM) with block erase, (i.e., not single word or bit realterable as in a RAM) either electrically or by irradiation. The reasons are that several problem areas are avoided, namely:

1) no complex selection sequences

2) lower voltage, longer duration writing may be used

3) writing wearout is avoided since an EAROM is unlikely to be reprogrammed more than a million times (e.g. 300 times a day for 10 years).

This application covers a large proportion of the general needs for memories and enables fixed ROM's to be replaced by alterable ones. This provides a great advantage in computers relying on microprogrammed instructions since their function may be easily changed electrically.

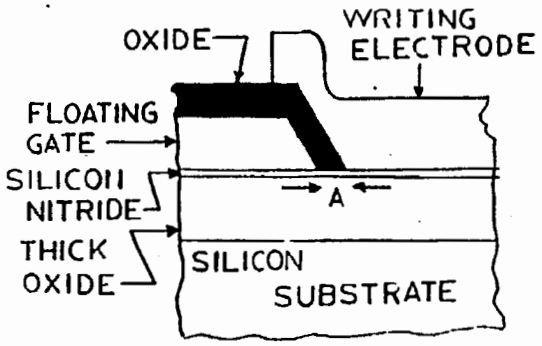

Fig. 11. Cross section of the non-linear resistor

11. ANALOG STORAGE

All of the devices described exhibit analog storage to some extent. However, the devices that store charge in traps have a rapid decay immediately after writing. This occurs at a much greater rate than the long term decay after the device has "settled." The EAR does not have this effect since the storage decay is characterized by a very long RC time constant. If the EAR is used without channel shielding, and positive or negative potentials are applied to the writing electrode to charge and discharge the floating gate, then the device conductance can be smoothly varied without disturbing the conductance during the change. This is not possible with the other types of storage devices. To achieve this with the EAR, the stray capacitance at the overlap of the writing electrode and the writing gate must be made negligible compared to the gate to substrate capacitance in order to avoid capacitive feedthrough of the control signal while resetting the device impedance. Also the length of the non linear conducting path may be increased to give

a slower response, if desired.

The resultant device is the analog of the motor driven potentiometer. This is the first solid state electrically alterable analog resistor in a single device.

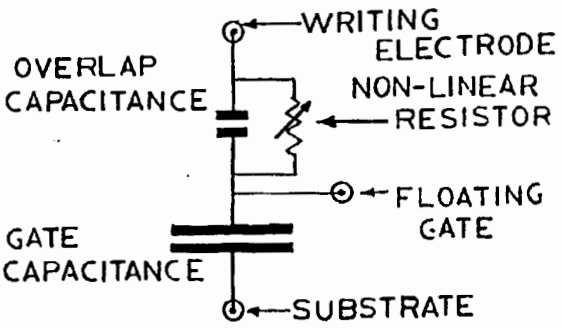

Fig. 12. Equivalent circuit of the EAR

12. CONCLUSIONS

Non-volatile semiconductor memories show reliable long term data retention, and may be used in place of conventional MOS memories. However, the operating speeds are lower. Also the higher voltages required impose design restraints and more difficult processing techniques. In order to avoid wearout effects, lower writing voltages are necessary, and differential sensing techniques are needed for reliably discriminating between small threshold differences. The use of N channel MOS will improve operating speeds; and the use of common charging electrodes, with EAR's or depletion shielded MNOS, may solve the problem of decoding high voltages on the chip.

Because of the difficulties of selective clearing and the wearout which occurs with high speed, high voltage writing pulses, present non-volatile storage transistors are best suited to electrically alterable ROM's with block erase. Non-volatile RAM's are considerably more difficult.

REFERENCES

1. H.A. Wegener, "MNOS memories," Digest Intermag. Conf., Apr. 1970.

2. D. Frohman-Bentchkowsky, "An integrated metal-nitride-oxide-silicon (MNOS) memory," Proc. IEEE (Lett.), vol. 57, June 1969, pp. 1190-1192.

3. D. Frohman-Bentchkowsky, "A fully decoded 2048-bit electrically programmable FAMOS read-only memory," IEEE Journal of Solid State Circuits, vol. SC-6, No. 5, Oct. 1971.

4. S. Nakanuma et al., "A read-only memory using MAS transistors," ISSCC Digest Tech. Papers, Feb. 1970, pp. 68-69.

5. H.G. Dill and T.M. Toombs, "A new MNOS charge storage effect, "Solid-State Electron. vol. 12, 1969, pp. 981-987.

6. Y. Hsia and G.S. Holland, "Silicon-nitride non-volatile semiconductor storage," Special Report, LSI Memories Session, 1970 WESCON, Los Angeles, Aug. 26, 1970.
Also:
Y. Hsia, et al., "Non-volatile, electrically alterable memory," Technical Report No. AFFDL-TR-7U-151, USAF, Wright-Patterson Air Force Base, Ohio, Nov. 1970.

7. M.H. Woods and J.W. Tuska, "Degradation of MNOS memory transistor characteristics and failure mechanism model," IEEE Reliability Physics Symposium, April 1972.

8. J.G. Mark and A.C. Tickle, U.S. Patent 3,636,530. Jan. 18, 1972.

9. A.C. Tickle, U.S. Patent 3,660,827. May 2, 1972.

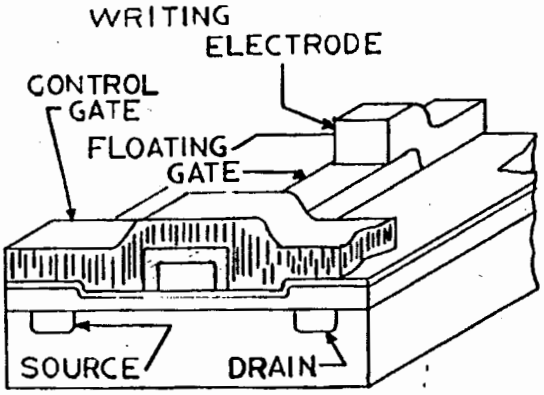

Fig. 13. Double gate EAR for memory arrays