Superior Retention Programmable Memory Transistor

EXECUTIVE SUMMARY:

OVERVIEW:

A novel Superior Retention Programmable Memory Transistor (PMT) design is described which is intended for 1.2µm Smart Power production and submicron Smart Power SP2 production. The existing PMT used in MOD CMOS designs consists of a Poly1/Poly2 capacitor integrated with an NMOS device. The Superior Retention Programmable Memory Transistor (SRPMT) is identical to the Poly1/Poly2 PMT except that the Poly2 layer is replaced with a lightly doped N-diffusion (NHV) having N+ and P+ contacts. The N+ layer forms an ohmic contact to the NHV. The P+ layer forms a "stitch" contact that sources holes into a P-type inversion layer at the surface of the NHV diffusion when the device is being programmed. The P+ stitch contact is the novel element in this invention. In addition, measurements demonstrate that the SRPMT has superior data retention when compared to a Poly1/NHV PMT that does not have a P+ stitch contact. Furthermore, all layers used in the SRPMT are core process layers in Smart Power (Poly1, NHV, P+, N+), so there is no need to add Poly2.

EXPERIMENTAL RESULTS:

Poly1/NHV PMTs with no P+ stitch (control PMTs) and Poly1/NHV PMTs with a P+ stitch (SRPMTs) were processed side-by-side on a Smart Power PMT test array. All devices were erased with a deep UV bake and were then programmed from an intial Vt of about 2V. After programming, the SRPMTs reached a Vt = 8.5V, while the control PMTs reached Vt = 7.5V. All PMTs were then subjected to a standard data retention bake at temperatures of 160C, 180C and 235C. The control PMTs experienced an initial drop in Vt of between 1.5 and 2.0V after the first hour of baking. After the initial rapid Vt drop, the control PMTs stabilized and Vt began to drop at a much slower rate which would be sufficiently slow to certify the devices as reliable over time and temperature. Under the same stress conditions, the SRPMTs experience no initial drop in Vt but rather Vt decayed at a slow rate through the entire data retention bake in a similar manner to the control PMTs after the initial drop in Vt. Given that the programmed Vt was initially higher and that the rate of decay was overall slower, the SRPMT exhibited superior data retention with respect to the control PMT.

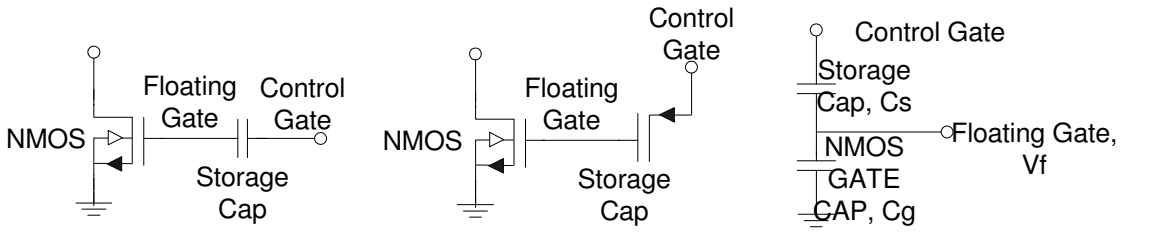
DISCUSSION:

Since the layers used to create the SRPMT are standard in most Smart Power processes and since there has been no discussion of this idea in the open literature, we want to patent the use of the stitch concept to provide a superior data retention PMT. We cannot as yet explain why the SRPMT does not experience the initial drop in Vt during data retention bake. However, since the initial drop is undesirable, the lack of this phenomenon in the SRPMT is not a cause for concern with regards to reliability. We will continue to study the phenomenon to determine the root cause. A more detailed explanation of the device design and testing procedures follows in the next section.

THEORY OF OPERATION OF PROGRAMMABLE MEMORY TRANSISTOR

STRUCTURE:

A PMT consists of a NMOS transistor with its poly gate extended to connect with and form the upper plate of a separate storage capacitor. Refer to Figure below. For optimal PMT programming, the storage capacitor is generally made many times larger than the NMOS gate capacitance. The bottom plate of this storage cap, located beneath the dielectric of the storage capacitor, is the control gate; i.e. positive voltage



applied to these gate controls the drain NMOS current. This bottom plate can be formed in two conventional ways. To make this clearer we will first discuss IC capacitor technology.

IC CAPACITORS:

An IC capacitor consists of two key elements: a dielectric and two plates. The dielectric can be an oxide or a depletion region. Generally, the best oxide is the thinnest that meets the voltage and reliability requirements. Typically it is the MOS gate oxide. The main requirement for the plates is that they are electrical conductors. Materials that are typically used are good conductors, i.e. aluminum and poly.

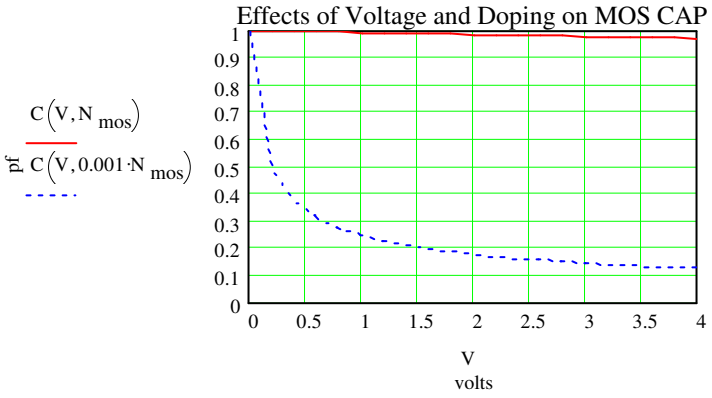
A common capacitor consists of a bottom poly 1 plate with a thin gate oxide grown on top of this and a top plate of deposited poly 2. A nitride layer and associated mask may be needed for oxide integrity. Also a "protect" mask may be needed to remove poly 1 to poly 2 stringers.

There is a potential cost avoidance if an alternative lower plate can be used in place of poly. At the very least, the capacitor's nitride layer and mask can be eliminated. Tradeoffs with process and layout topologies are complex issues. Eliminating the need for the bottom poly layer certainly results in increased flexibility and offers potential for additional savings.

N DIFFUSION FOR LOWER PLATE CAPACITOR FOR HIGH SPEED PROGRAMMING:

An n-type diffusion, which we call NHV, can function as the lower plate. This allows us to use the common pwell for both the NMOS and storage cap. The pwell then serves as isolation to the N- substrate. Contact to the NHV, is made by a subsequent more highly doped diffusion, called an +, that is self-aligned to the poly 2. Refer to Figure "PMT WITH DIFFUSION PLATE".

However, when the positive control voltage is applied to the NHV plate, which underlies the capacitor dielectric, it causes a depletion region to form. (The depletion is frequency dependent. Consider the case for programming where the frequency > 1000Hz). This depletion region forms a capacitor in series with the oxide capacitor. This is the depletion capacitance found in all MOS structures. At 0V, the MOS capacitance equals the oxide capacitance alone. Increasing the doping level can minimize the effect of this depletion capacitance. The solid line in the plot below shows the effect for a typical ($1.5x10^{19}$ atoms/cm^3) CMOS process doping level. This effect is less than 5% for voltages less than 6V. The dashed line shows that reducing the doping level by a factor of 100 drops the capacitance by 90% at 6V. Additionally, it would have



a much higher plate contact resistance. Thus, a storage capacitor using this reduced doping level generally would not be used. However, as we shall see, this capacitor with modification, can result in a PMT with superior retention and testability.

CHANNEL HOT ELECTRON INJECTION PROGRAMMING:

In the programming mode, drain and gate voltages are applied to the NMOS for a few milliseconds or less. A positive voltage less than the NMOS BVdss is applied through an optional current limiting resistor to the drain and a positive voltage on the order of $5x10^6$V/cm applied to the gate. The drain voltage accelerates electrons to the drain. Some small fraction of the electrons cross the channel without collisions and gains kinetic energy close to that of the applied drain potential. The poly to silicon dioxide barrier height for silicon dioxide is about 3.2ev. Some small fraction of these "hot" electrons near the drain collides with the silicon lattice and are scattered thus changing their direction toward the dielectric. Electrons in the upper tail of the distribution have sufficient energy to overcome the silicon dioxide barrier energy. Most of these tail electrons will then drift toward and are collected by the floating gate. A small fraction of the injected electrons are trapped in the oxide.

Programming is the process of injecting these hot electrons, which then become stored on the common poly floating gate. Subsequently, when a positive voltage is applied to the control gate, it's effect on the channel is screened by the negative charge on the floating gate resulting in a much higher effective gate threshold voltage.

The threshold in the uncharged state is Vto =Vth (Cf + Cs)/Cs. The magnitude of the increased threshold voltage is ΔVth =Qf/Cs, where Qf is the charge injected onto the floating gate. Thus the control gate threshold is Vto + ΔVth. The effective Vth then increases during programming; this in turn causes the current to decrease resulting in current self-limiting.

RETENTION:

Because of the large interfacial barrier energy, once a charge is stored onto the floating gate, it has a long intrinsic storage time. For PMTs, the measured Vth mean decay is 0.2V/decade_hours at 160C and the initial programmed mean Vth is 8.1V. Thus it would take 10^{21} years for the PMT to discharge to a Vth of 3V. At the end of ten years the leakage has dropped to an average of one electron per day.

PHASES OF CHARGE DYNAMICS:

PMT Vth degradation is the result of and limited by physical processes. The magnitudes of the electric field and temperature dictate what conduction processes will be dominant. There are three distinct phases of Vth degradation for nominal PMTs, each associated with a different possible physical mechanism of charge distribution/conduction and each having its own empirical "activation energy".

First there is an initial period of rapid Vth loss. We shall assume that this is associated with the depolarization/dielectric absorption behavior observed to a lesser or greater degree in all capacitor dielectrics. We have observed this phenomena in polystyrene, ceramic and now silicon dioxide dielectrics. Experimental data on PMTs for the CMOS technology process revealed that the Vth drop during the initial 1.6 hour period was 0.3 to 1.6V (from an initial Vth value of 8.3 volt) for both NHV and poly 1-poly 2 storage capacitors. The magnitude of the "activation energy" for this initial phase (as observed in our "B" storage capacitor design) is so small, ~ 0.027 ev, that the thermodynamic concept of an activation energy is not really applicable.

Second, there is an intermediate period of charge loss associated with a high (but less than 6Mvolt/cm where F-N tunneling is dominant), but decaying electric field. It is possible that there is movement of trapped electrons. This has an "activation energy" of about 0.2 ev. Thermodynamically, this is a region described by the Eyring stress level dependent activation energy model.

Ultimately, we arrive at the long period of low field leakage through the oxide. The low field conduction mechanism that is generally accepted for EPROM is conduction by a thermionic emission. The "activation energy" projected from our low temperature data is about 0.4 ev.

The retention data also shows a difference in concavity between the Vth versus log time plots for low (concave up) and high (concave down) temperatures, indicating different dominant conduction mechanisms.

CHANNEL COUPLING PMT:

LOW VTH CHANNEL COUPLED PMT, SUPERIOR RETENTION:

The low doping NHV PMT can be improved. One capacitor structure that is known to us is an N extension with a P+ contact diffusion. See the attached Figure “Channel Coupling PMT Design”. The P+ contact functions as the source of a PMOS. For V> Vth, the P+ source floods the depletion/inversion region under the control gate dielectric with holes that form a relatively low resistance channel. Under these circumstances, it is an advantage to reduce NHV doping density because this reduces the threshold voltage. In this application because we are supplying just gate current, the high NHV resistance is not a limitation.

Attached at the end are plots that compare different PMT designs. The plots are a matrix of columns of three storage capacitor sizes with four rows of different capacitor designs designated, A, B, C and D. Cap type A has the PSD source, B is A without the PSD source, C has a PSD source and a smaller NMOS. The first two plots of D have PSD; the third does not. The plots reveal that the channel coupled design is free from the initial rapid Vth loss phase.

BENEFITS OF CHANNEL COUPLED PMT:

INCREASED RETENTION

Eliminates the rapid charge loss phase. Eliminates 1.5V drop, which with a 0.2Vper hour decade, long term retention is equivalent to 5 decades > 10,000 hours increased retention.

IMPROVES TESTABILITY INCREASING RELIABILITY OF VDD MARGINING:

Allows quality level to be screened with VDD/5V margining. A 1.5V initial drop plus process variations could reduce Vth below 5V

BETTER PPM SCREENING/RELIABILITY:

Allows more sensitive discrimination for Vth initial loss.

NHV PERMITS COST AVOIDANCE BY ELIMINATING THE NITRIDE AND INCREASES THE FLEXIBILTY OF THE PROCESS.

Thomas W. Kotowski
Leap Cad Systems