Intel StrataFlash™ Memory Technology Development and Implementation

Al Fazio, Flash Technology Development and Manufacturing, Santa Clara, CA. Intel Corp.
Mark Bauer, Memory Components Division, Folsom, CA. Intel Corp.

Index words: StrataFlash, MLC, flash, memory.

Abstract

This paper will review the device physics governing the operation of the industry standard ETOX™ flash memory cell and show how it is ideally suited for multiple bit per cell storage, through its storage of electrons on an electrically isolated floating gate and through its direct access to the memory cell. The device and reliability physics aspects of the three key technology features of multiple-levels-per-cell (M.L.C.): precise charge placement, precise charge sensing, and precise charge retention are discussed. The mixed signal design implementation of these features is reviewed along with challenges for low periphery circuit overhead and standard flash memory product performance. Lastly, process manufacturing aspects are reviewed and it is shown how Intel StrataFlash™ memory is manufactured on the same process flow and at the same high yields as standard flash memory.

Introduction

The concept of M.L.C. is ideally suited to the flash memory cell. The cell operation is governed by electron charge storage on an electrically isolated floating gate. The amount of charge stored modulates the flash cell's transistor characteristic. M.L.C. requires three basic elements: (1) Accurate control of the amount of charge stored, or placed, on the floating gate such that multiple charge levels, or multiple bits, can be stored within each cell, an operation called placement; (2) accurate measurement of the transistor characteristics to determine which charge level, or data bit, is stored, an operation called sensing; and (3) accurate charge storage, such that the charge level, or data bit, remains intact over time, an operation called retention. These elements are achieved by exploiting stable device operation regions and by the direct cell access of the ETOX flash memory array.

Flash Cell Structure and Operation

An explanation of M.L.C. first requires a review of the flash memory cell. The ETOX flash memory cell and products[1] have a long manufacturing history, having evolved in the late 1980's from EPROMs, which had been an industry standard from the early 1970's.

Cell Structure

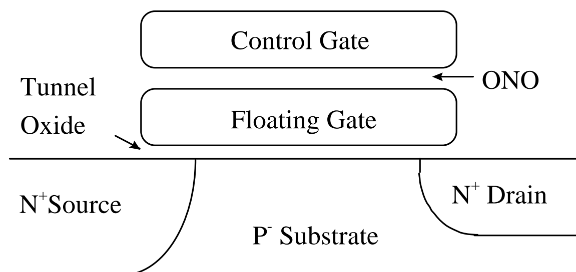

Figure 1: ETOX flash memory cell cross section

Figure 1 shows a cross-sectional view of a flash cell. It consists of an N-channel transistor with the addition of an electrically isolated poly-silicon floating gate. Electrical access to the floating gate is only through a capacitor network of surrounding SiO_2 layers and source, drain, transistor channel, and poly-silicon control gate terminals. Any charge present on the floating gate is retained due to the inherent Si-SiO_2 energy barrier height, leading to the non-volatile nature of the memory cell. Characteristic of the structure is a thin tunneling oxide (~100Å), an abrupt drain junction, a graded source junction, ONO (oxide-nitride-oxide) inter-poly oxide, and a short electrical channel length (~0.3µ). Because the only electrical connection to the floating gate is through capacitors, the flash cell can be thought of as a linear capacitor network with an N-channel transistor attached. The total capacitance of the cell (C_{TOT}) is equal to the additive

capacitance of the network. For convenience, coupling ratio terms, which are defined as the ratio of terminal voltage coupled to the floating gate, can be defined as follows:

GCR =control gate coupling ratio,

DCR = drain coupling ratio, and

SCR = source coupling ratio.

Therefore, a change in control gate voltage will result in a change in the floating gate voltage, $\Delta V_{FG}=\Delta V_{CG}*GCR$. The basic equation for the capacitor network is

$$V_{FG}=Q_{FG}/C_{TOT} + GCR*V_{CG} + SCR*V_{SRC}+DCR*V_{DRN} \quad (1)$$

where Q_{FG} = the charge stored on the floating gate.

A simple first-order transistor equation of drain current says

$$I_D=G_M*(V_{FG}-V_{CG}-V_{DRN}/2)*V_{DRN} \quad (2)$$

where $G_M=q\mu eC_{OX}Z_E/L_E$

This equation is very inexact for the small geometry of the flash cell, but nevertheless the conclusions hold. Substituting V_{FG} of the basic coupling ratio Equation (1) into the basic transistor I-V Equation (2) leads to the conclusions that the transconductance of the transistor (and also the pre-threshold slope) degrades by GCR, while the threshold voltage, V_T, depends upon Q_{FG}, the charge stored on the floating gate. Therefore, the V_T depends upon Q_{FG}, while the I-V shape does not. Very simply, the flash cell can be thought of as a capacitor which is charged and discharged, the charge value being determined by the amplification of the transistor I-V. To give an idea of the amount of charge, every volt of cell threshold corresponds to approximately 10,000 electrons of floating gate charge.

Cell Operation: Programming

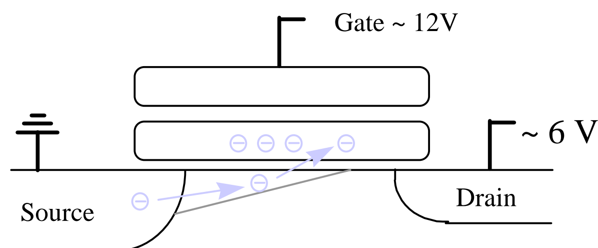

Figure 2: Cell bias conditions during programming

Programming a flash cell means that charge, or electrons, are added to the floating gate. Figure 2 shows the cell bias conditions during program operation. A high drain to source bias voltage is applied, along with a high control gate voltage. The gate voltage inverts the channel, while the drain bias accelerates electrons towards the drain. Programming a flash cell, by channel hot electrons, can be understood by use of the lucky electron model[2], as illustrated by the energy band diagram in Figure 3. In the lucky electron model, an electron crosses the channel without collision thereby gaining 5.5-6.0eV of kinetic energy, more than sufficient to surmount the 3.2eV Si-SiO_2 energy barrier. However, the electron is traveling in the wrong direction. Its momentum is directed towards the drain. Prior to entering the drain and being swept away, this lucky electron experiences a collision with the silicon lattice and is re-directed towards the Si-SiO_2 interface, with the aid of the gate field. It has sufficient energy to surmount the barrier. However, an electron does not have to be completely lucky. It can be "somewhat lucky" or "barely lucky," making the process of programming efficient. We can observe from this model that the lateral field, determined by bias voltage, junction profiles, electrical channel length, and channel doping are important to the effectiveness of generating energetic electrons and are therefore key to the M.L.C. placement operation. Hence the abrupt drain junction and short channel length of the cell structure. After programming is completed, electrons are added to the floating gate, increasing the cell's threshold voltage. Programming is a selective operation, uniquely occurring on each individual cell.

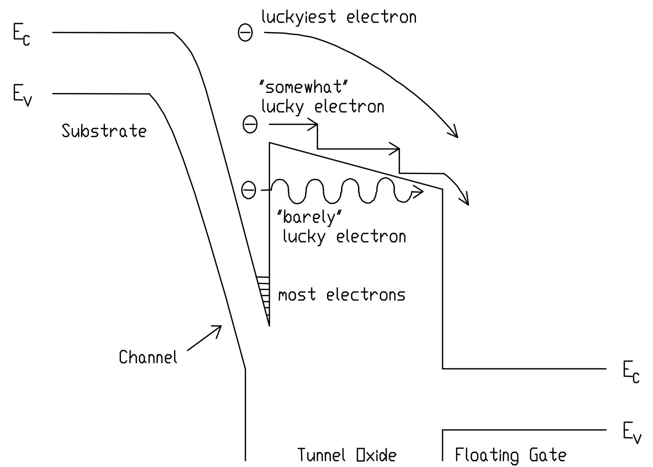

Figure 3: Energy band diagram of programming

Cell Operation: Erase

The distinguishing feature between EPROM and flash memory is the erase operation. EPROM removes electrons from the floating gate by exposure to ultra-violet

light. A photon of this light source has high enough energy that if transferred to an electron on the floating gate, that electron will have enough energy to surmount the Si-SiO_2 energy barrier and be removed from the floating gate. This is a rather cumbersome operation requiring a UV-transmissive package and a light source. It is also rather slow and costly, often requiring the removal of the memory from the system. In flash, the contents of the memory, or charge, are removed by means of applying electrical voltages, hence to be erased in a *flash*, with the memory remaining in the system. The electrical erase of flash is achieved by the quantum-mechanical effect of Fowler-Nordheim Tunneling[3], for which the bias conditions are shown in Figure 4. Under these conditions, a high field (8-10MV/cm) is present between the floating gate and the source. The source junction experiences a gated-diode condition during erase, hence the graded source junction of the cell structure. As evidenced by the energy band diagram of Figure 5, electrons tunneling through the first ~30**Å** of the SiO_2 are then swept into the source. After erase has been completed, electrons have been removed from the floating gate, reducing the cell threshold. While programming is selective to each individual cell, erase is not, with many cells (typically 64k-Bytes) being erased simultaneously.

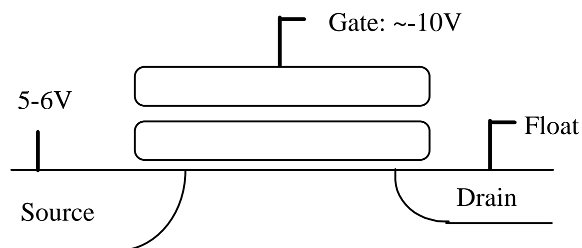

Figure 4: Cell bias conditions during erase

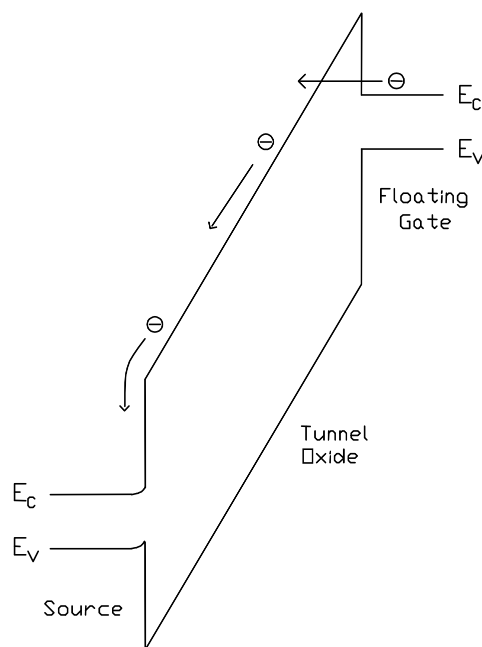

Figure 5: Cell energy band diagram during erase

Cell Operation: Read

The read operation of the cell should now be apparent. Storing electrons (programming) on the floating gate (Q_{FG} < 0), increases the cell Vt. By applying a control gate voltage and monitoring the drain current, the difference between a cell with charge and a cell without charge on their floating gates can be determined (Figure 6). A sense amplifier compares the cell drain current with that of a reference cell (typically a flash cell which is programmed to the reference level during manufacturing test). An erased cell has more cell current than the reference cell and therefore is a logical "1," while a programmed cell draws less current than the reference cell and is a logical "0." The floating-gate charge difference between these two states is roughly 30,000 electrons.

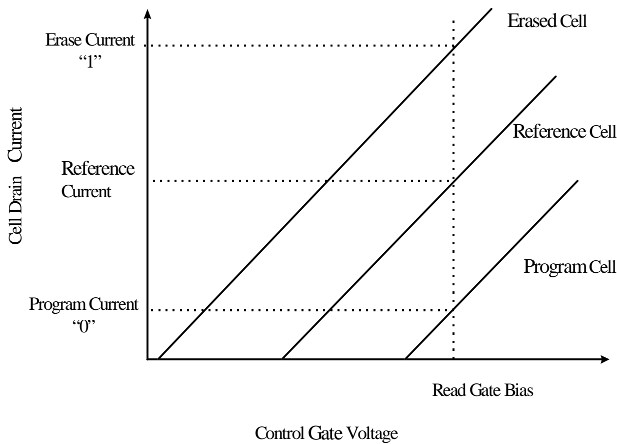

Figure 6: Erase, program and reference cell I-V

Array Configuration

Figure 7 shows a schematic drawing of the flash memory cells in a NOR array configuration. In this configuration, cells on the same wordline, or row, share common control gates. Cells on common bitlines, or columns, share common drains, which are connected via low resistance metalization, providing direct access to each cell's drain junction. The sources for cells in the array are common. They are connected locally via common degenerately doped silicon and globally via low resistance metalization. Decoders are linked to the control gate wordlines and drain bitlines to uniquely select cells at the cross point location. The direct access to the cell in this configuration versus alternative array architectures that have parasitic resistance or devices, ensures that accurate voltages can be applied to the cell and IR drops are minimized. This is a key aspect to achieving M.L.C.

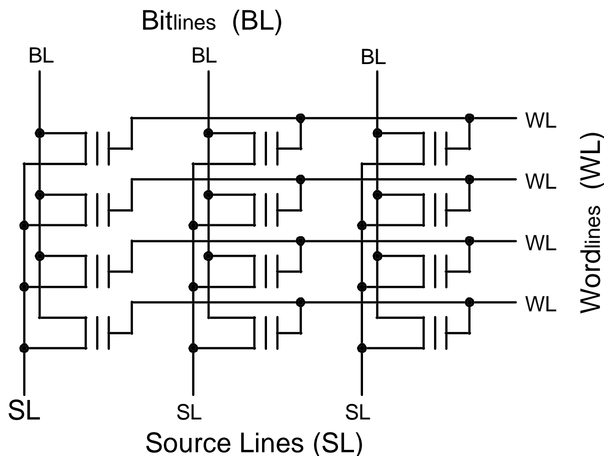

Figure 7: Array configuration

M.L.C. Key Features

We have reviewed thus far how a one bit per cell (1B/C) flash memory operates. As can be inferred from the previous discussion, M.L.C. is simply a means by which charge on the floating gate is modulated and detected to levels lower than the 30,000 electrons described above, such that intermediate charge levels, or states, can be extracted from the cell. These states can now represent not just the simple 1B/C "1" and "0," but rather an M.L.C. representation with four distinct charge states: "11," "10," "01" and "00," or 2 bits in one cell. These four distinct levels are illustrated in the I-V curve of Figure 8. The key aspects of achieving these intermediate states, or levels, are precise charge placement, precise charge sensing, and precise charge retention.

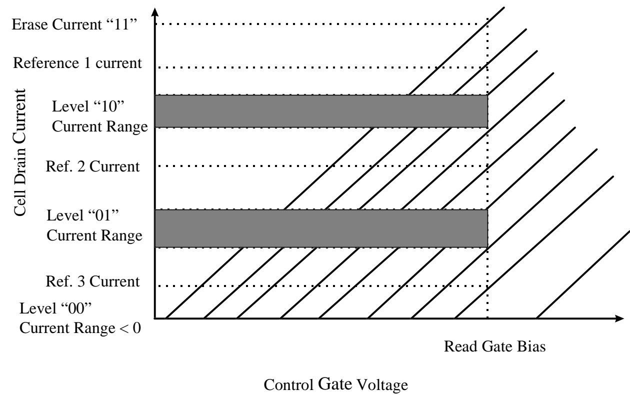

Figure 8: Cell and reference I-V curves of 4-level 2B/C

Precise Charge Placement

A comparison of Figures 6 and 8 shows that M.L.C. requires a means to control how much programming occurs within a cell. For a 1B/C product, all that is necessary is to have enough programming to change a "1" into a "0." Over-programming a cell to much higher Vt's (adding more floating gate charge) would be fine. This is not the case for M.L.C., where too much programming would cause an intermediate level to overshoot onto the next level. For instance, if a "10" was desired, but a cell was over programmed, a "01" might occur, leading to erroneous data. Therefore, a method of controlling precisely how much charge is transferred to the floating gate is required. Enough charge is needed to reach a state level without overshooting the desired level.

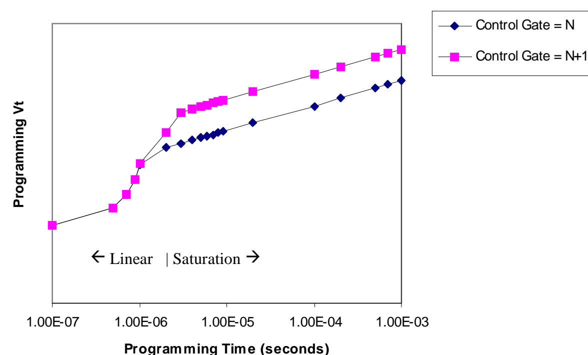

Figure 9: Programming threshold vs. time curve

To gain insight into how such precise control can be obtained, let's take a deeper look into the flash cell's programming characteristics. Figure 9 shows how the flash cell's Vt changes as a function of log-time under two different bias conditions. Two regions of operation are shown: linear and saturation, so-called because the linear region is linear when plotted in linear time, and the saturation region is where the cell Vt changes little with time, analogous to a MOS transistor I-V curve. Note also that in the linear region, the control gate voltage has little influence on the rate of programming, while in the saturation region, the control gate voltage has a strong dependence upon the saturated Vt. A characteristic of Figure 9 is that the flash cell programming slows as more charge is added to the floating gate. The reason for this behavior is that when in the linear region, energetic electrons, near the drain, are attracted to the floating gate. As programming progresses, the floating gate (which is coupled to the control gate and drain biases as governed by Equation 1) becomes charged more negatively, until it eventually reaches the same bias potential as the drain voltage. At this point, the energetic electrons become repelled by the floating gate charge. Programming slows, as near-drain electrons must tunnel through the SiO_2 barrier, or less energetic mid-channel electrons "jump" over the barrier. The strong gate dependence results from the vertical field limitation in this region. One can also see from Figure 9 that the saturated Vt increases in a one for one fashion with an increase in the programming control gate voltage. This is a simple result of the coupling Equation (1).

Given this characteristic curve, one could devise several possible methods of controlling the charge transfer to the floating gate. These methods would have to pass the criteria of being reliable (no overshoot), controllable (simple to implement), and fast (to ensure compatibility with standard flash memory product features). Programming in the linear region while being fast is not controllable. In this region, programming Vt is exponentially dependent upon time and the electron energy distribution (as determined by drain bias, channel length, doping profiles, etc.). Small variations will lead to large changes in the cell threshold and therefore overshoot of the desired state, thereby having a high likelihood of being unreliable. Minimization of these variations would be also difficult to implement. In the saturated region, the cell Vt simply depends upon the applied control gate voltage. Control in this region is more achievable. With ease of control, design optimization practices can be employed to achieve fast programming. This will be shown later. Therefore, to achieve speed and control, a placement algorithm that employed programming in the saturated region was developed.

This leaves us with reliability. Unlike Fowler-Nordheim Tunneling, used for programming in addition to erase in some versions of flash memories and subject to erratic programming due to the presence or absence of as few as one or two holes trapped in the oxide[4], channel hot-electron programming has no erratic programming mechanism. The programming threshold in saturation is simply a linear function of the applied control gate voltage. Programming in this region can be forced into an unstable operating point, known as impact-ionization induced latch-up[5]. This is the point where an excess of holes in the silicon substrate, created by the collisions of the energetic electrons with the silicon lattice, build up to the point where the parasitic NPN transistor in the silicon substrate turns on. Proper architectural design of the silicon process flow (i.e., use of EPI silicon) can easily prevent this from happening.

Therefore, the three success criteria are satisfied by exploiting stable device operation regions, namely programming in saturation. Now, a simple placement algorithm was chosen for implementation (outlined in the flowchart in Figure 10). The algorithm consists of the simple loop of programming in saturation, checking the cell Vt to determine if the desired state has been reached, stopping if the desired state is reached, or if not, incrementing the control gate voltage and providing an additional programming pulse and continuing in this fashion until the desired Vt has been achieved. In the Intel StrataFlash™ memory two bit per cell device, each programming pulse within the placement algorithm will transfer roughly 3,000 electrons of charge to the floating gate.

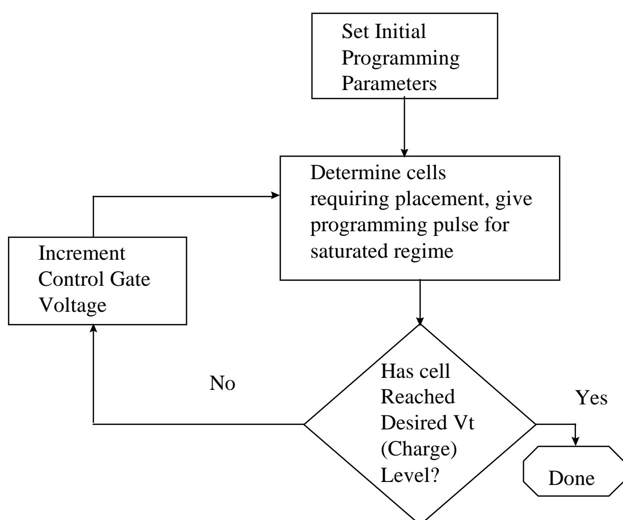

Figure 10: Placement algorithm flowchart

Precise Charge Sensing

As can be seen from the flowchart in Figure 10, integral to the placement algorithm is a means of detecting whether or not the desired cell Vt has been achieved. Without a precise means of sensing the floating gate charge, precise charge placement would not be possible. A look back at Equation 2, the cell drain current-voltage relationship, gives some insight into what is required to achieve precise charge sensing. Control gate and drain voltage control and process Leff, Zeff, mobility and oxide capacitance control are important aspects of precise charge sensing. Drain voltage control is facilitated by direct access to the cell drain junction (bypassing any resistive IR drops) allowable in the ETOX NOR flash memory array architecture; and by applying a high enough drain voltage to operate in the saturated mode (normal MOS device saturated I-V, not programming saturation as previously discussed) where drain bias variations have minimal current impact. Process control is important, since the 33,554,432 memory cells contained within a single Intel StrataFlash memory 64Mbit device represent a >10 sigma variation, and is achieved by proper process architecture and manufacturing process control, derived from the ten years of manufacturing experience with flash memories. Control gate voltage control is achieved by an on-chip read regulation circuit, which is fully explained in a later section.

Flash memory has a unique feature associated with its non-volatility: the data write (placement) can occur under one condition of ambient temperature and system power supply, while the read out of data (sensing) can occur at a later date, at different ambient temperature and system power supply. Being fundamentally a MOS transistor, the flash cell's drain current is a function of these ambient conditions. As such, the precise charge sensing is required to span wide ranges of operation. To facilitate this needed precision, the reference levels that separate charge state levels are generated by reference flash cells contained on-chip. These reference cells, whose Vt levels are precisely placed at manufacturing test under a controlled environment, will have the same tracking with temperature and power supply as the array flash cells. This contrasts to reference levels generated by other transistor types (i.e., NMOS or PMOS), which have different temperature, voltage, and process tracking than the flash memory cell. This lessens the necessary constraints on the read regulation circuitry.

Precise Charge Retention

Due to the non-volatility requirement of flash memory, it is important that any charge placed on the floating gate remain intact for extended periods of time, typically ten years. This translates to a requirement of not losing more than one electron per day from the floating gate. If electron loss occurs from even one memory cell in an array of millions, the data will be corrupted. The inherent storage capability exists due to the Si-SiO_2 energy barrier which traps electrons on the floating gate. The inter-poly-silicon oxide (ONO film mentioned in the cell structure) is processed to maximize charge storage capabilities[6]. Under normal circumstances, the energy barrier allows charge storage for hundreds of years. There are conditions of trapped oxide charge, known as intrinsic charge loss[7], which can cause one-time shifts in threshold. These shifts are rather small and are compensated for during manufacturing test. Random defects in the insulating oxides that can lead to charge loss are less of an issue with low-defect, high-yielding process technologies, but if still present, are screened out by the manufacturing tests. These defects are driven to low enough levels on ETOX flash memories where error-correcting-codes (ECC) are not needed. The remaining concern for charge retention is any degradation to the insulating oxides that occurs as a result of the stresses of device operation.

During normal operation, high fields are applied to the flash cell. The presence of the high fields over time can degrade the charge storage capabilities of the device, by effectively lowering the energy barrier, or by providing traps sites in the oxide that can act as intermediate tunneling locations. The benefit of channel hot-electron programming, compared to tunneling for programming, is that fast programming can occur at lower internal fields thereby lessening the probability of oxide damage. Nevertheless, occurrence of damage needed to be understood to ensure the stability of the M.L.C. charge.

Consequently, the charge retention ability of the insulating oxides under various process and bias field conditions were studied in great detail. Over the course of the four year M.L.C. development period, in excess of 200 billion ($2e10^{11}$) flash cells were studied for charge retention, each to a resolution of floating gate charge of ~100 electrons. This exhaustive study provided more physical insight into the oxide damage mechanisms and has enabled us to build large scale empirical models for charge retention. The net result of this study was the ability to optimize process recipes and operating bias fields to maximize charge retention. This allows Intel StrataFlash memory to maintain high reliability performance, without the use of any ECC.

Mixed Signal Design Implementation

The implementation of the described charge-placement algorithm and charge-sensing operation required a mixed signal circuit design of both digital and precision analog voltage generation, regulation and control circuits. The placement algorithm is executed by utilizing an on-board control engine, or the Flash Algorithmic Control Engine (FACE). FACE runs the placement algorithm by sequencing through the programming and sensing loops. During a read operation (sensing of data at a later time), the user has random access to the memory array. A read operation performs a precision-sensing operation and invokes circuitry controlling the precise cell bias voltages.

Placement Algorithm Implementation

The placement algorithm executed by FACE is stored in a small on-chip programmable flash array. The programmable microcode allows for flexibility in algorithm changes. FACE, illustrated in Figure 11, consists of the microcode storage array, program counter (PC), arithmetic logic unit (ALU), instruction decoder, clock generator, register files, and input/output circuitry. FACE uses 6,000 transistors for logic and 32k bits of flash memory for algorithm storage.

To describe the implementation of the placement algorithm, let us assume that a group of cells (i.e., a double-word, or 32-logical bits, 16-physical cells) is to be placed and is initially in the erased state (lowest floating gate charge state). Any cells not to remain in the erased state (representing logical data "11") will receive a programming pulse. FACE will look up the drain and initial control gate voltage stored in a permanent read-only register located on-chip. FACE will then set the control gate voltage through the digital to analog converter (DAC). The DAC circuit receives the FACE digital input and divides the on-chip generated 12-volt power supply (VP12) to achieve the desired control gate voltage for that particular programming pulse. The drain voltage, used during the programming pulse, is generated from a regulation circuit that sets the gate voltage on a source follower. FACE will continue to supply the programming voltages for the pre-determined amount of time sufficient to reach the saturation region. When the programming pulse is complete, FACE will reconfigure the circuits to perform the sensing portion of the algorithm, an operation called verification. The drain and control gate voltages are now set to the same values as used in a user read access to ensure common mode between verification and read. FACE will take the result of the verification and determine which cells have reached their destination charge level and which have not. Those that have not will require an additional programming pulse with an increased control-gate voltage. A cell that no longer requires additional programming pulses will have the drain voltage disabled by the program pulse selector circuit. This sequence of events continues until all cells in the double-word have completed programming.

Analog Circuit Blocks for Precise Charge Placement

Placement requires precision voltages covering a range of 4-12 volts, while the chip Vcc (user supplied voltage) is kept at a typical value of 5 volts. The voltages applied to the memory array need to be internally generated and precisely regulated. On-chip voltage generation is achieved by use of charge pumps, in which switched capacitors boost the user-supplied Vcc to higher values. Voltages are controlled using a precision voltage reference circuit and voltage regulation circuits (Figure 11).

During a programming pulse, two charge pumps are used. One charge pump generates the internal 12V supply (VP12). This is used to supply a precision control gate voltage to the flash cells, through the DAC circuit.

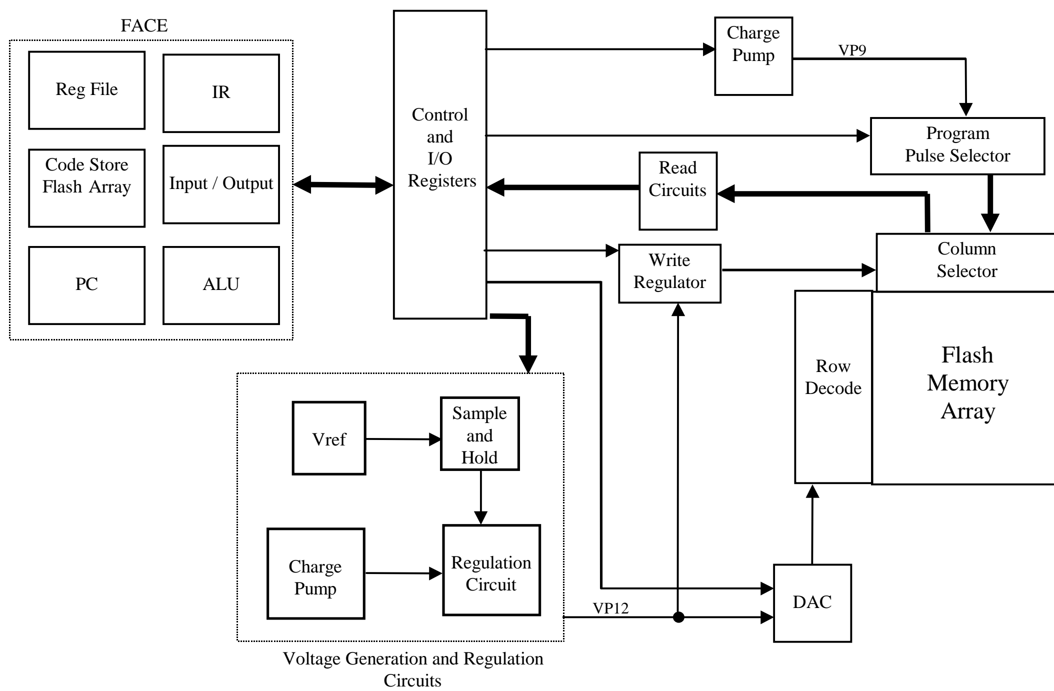

Figure 11: FACE and placement operation block diagram

VP12 also serves to generate the precision flash drain voltage through the write regulation circuit (WRC). The WRC generates a voltage that is applied to an NMOS transistor configured as a source follower. This transistor is in the bitline (or drain) path of the flash cell. The flash cell drain current is supplied through a second pump that generates the signal VP9. This pump is required to supply the programming current for up to 32 flash cells at a time.

During the placement algorithm, voltage stability is critical to precise charge storage. Any variations in the reference circuit voltages will be seen as variations in the flash control gate voltage, to which the programming saturated Vt is directly related. To achieve this absolute stability in the voltage reference circuit, a sample and hold circuit is employed. At the start of the placement algorithm, the sample and hold circuit samples the reference voltage and holds the value on a capacitor during the running of the entire algorithm. This guarantees the control gate voltage varies from pulse to pulse by only the desired step value and not by any additional components.

Circuit Blocks for Precise Charge Sensing

When the device is in the read mode of operation, FACE is disabled and the user has control to access the memory array. A read operation consists of sensing 16 bits worth of data from a random location in the memory array. With M.L.C., 8 flash cells are used to obtain 16 bits of data. During the read operation (Figure 12), the flash cell control gate voltage is controlled through a read regulator circuit (RRC). Minimizing this voltage variation will minimize the variations in cell current (Equation 2). This allows for more precise measurement of the charge level stored on the floating gate. Drain voltage stability is also important to ensure that the flash cell being sensed has a high enough drain voltage to keep the memory transistor operating in the saturated region of the MOS I-V.

Due to fluctuations in user supplied Vcc and a lower value than may be needed during read, an internal voltage charge pump is used during a read operation to generate the internal voltage to supply the flash cell control gate. The RRC uses the same voltage reference circuit that is used for voltage regulation during a placement operation, as mentioned above. However, in the case of a read operation, not as much voltage stability is required so the sample and hold circuitry is not used.

Parallel Charge Sensing

High speed random access and precise charge sensing are accomplished through a parallel charge-sensing scheme. Through direct connections to each memory cell, the data read operation determines the level of each memory cell quickly, accurately, and reliably. The data read operation senses which of the four levels the memory cell falls within based on the threshold voltages of three reference cells. This is done simultaneously with three sense amplifiers (Figure 13), where each sense amplifier compares the flash cell current being sensed to

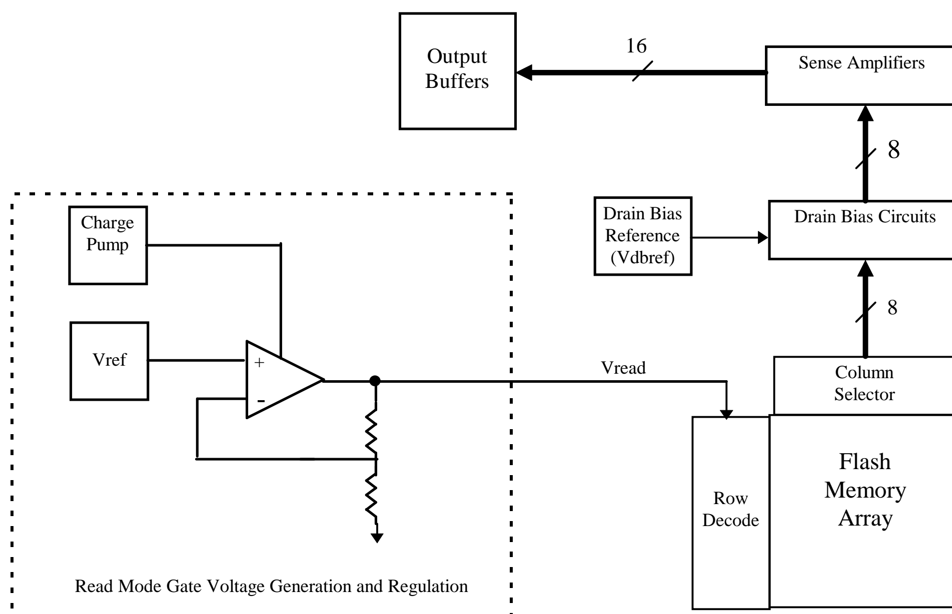

Figure 12: Read operation block diagram

the current of the flash reference cells.

The memory cell and the reference cells are biased in such a way that each conducts a current (Icell and Iref) proportional to their respective threshold voltage (Vt and VtRef). During a read operation, Vread is placed on the control gates of the memory and reference cells, the source terminals are grounded, and the drain voltages are set through a bias circuit that utilizes a precision voltage reference circuit.

The current for the memory cell being sensed is compared to the current of the three reference cells. The memory cell and reference cell current is converted to a voltage through an active load transistor. The resultant voltages are compared by the three sense amplifiers. A sense amplifier is associated with each of the three reference cells. Each sense amplifier also has an input from the flash cell being sensed. If the current of the cell being sensed is greater than the current of the reference cell (Icell > Iref or Vt < Vtref,), the sense amplifier output is a logic "1." If the current of the cell being sensed is less than the current of the reference cell, the sense amplifier output is a logic "0." The outputs of the three sense amplifiers are connected to a logic circuit that interprets the two data bits in parallel.

Cell Vt	Output Sense Amp 1	Output Sense Amp 2	Output Sense Amp 3	D1	D0
Vt < VtR1	1	1	1	1	1
VtR1 < Vt <VtR2	0	1	1	1	0
VtR2 < Vt <VtR3	0	0	1	0	1
Vt >VtR3	0	0	0	0	0

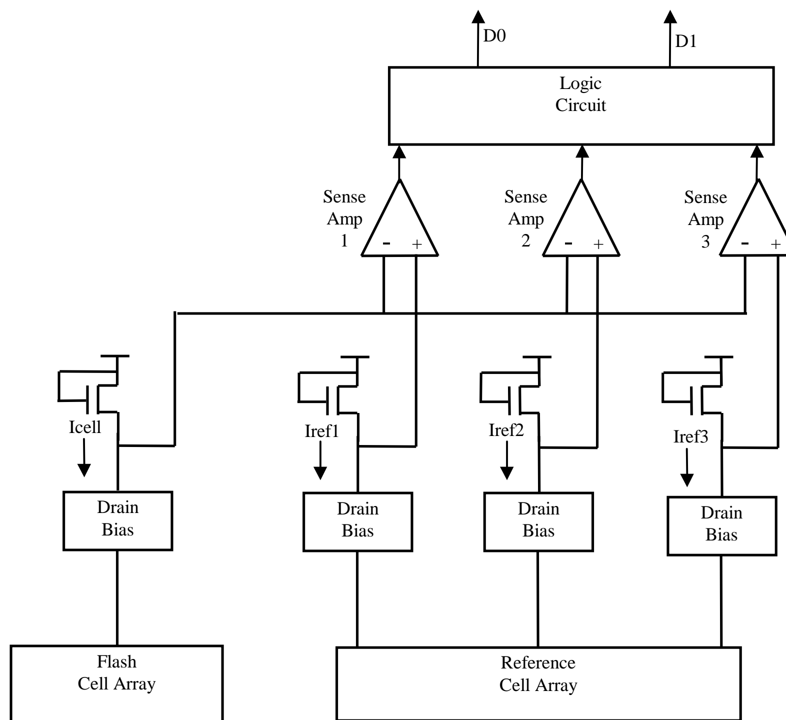

Figure 13: Parallel charge sensing

Low-Cost Design Implementation

Traditionally, a storage element in a memory corresponds to one bit of information. To double the amount of memory, the memory array or memory storage elements would need to be doubled. In addition to doubling the number of memory elements in the array, certain memory interface circuits must also be doubled. In particular, the memory array needs to be decoded requiring wordline and bitline decoders. In a typical single transistor non-volatile memory device (flash, EPROM), approximately 20% of the silicon area used is due to these interface circuits required to access the array. These interface circuits typically do not scale with process technology at the same rate as the memory array because they have high voltage and analog requirements.

Intel StrataFlash memory doubles the storage capacity of a memory device without doubling the memory array and the associated interface decoding circuitry. Additional circuitry is required to achieve the multiple bits per cell, but takes up a relatively small additional area. The additional overhead for circuitry is due mostly to the additional sense amplifiers, reference circuitry, and circuitry for voltage generation or charge pumps. The additional silicon area required for this circuitry represents only an additional 5% over what is necessary for a one bit per cell device. Implementations of M.L.C., which require externally supplied components (i.e., microcontroller, ECC, and voltage regulators), have the cost savings of M.L.C. diminished by these peripheral overheads. Intel StrataFlash memories achieve 2x the density at very close to 1x the area.

Low-Cost Process Manufacturing

ETOX flash memory has a long manufacturing history. As such, it was necessary that any implementation of M.L.C. not disrupt that history by having unique process requirements, which would cause a slow yield learning period or poor manufacturing throughput. First and foremost for M.L.C. to be successful, it must be able to ride on a technology that produces error-free one bit per cell flash memory. This requirement throughout Intel's ETOX NOR flash memory's history has resulted in tight manufacturing margins and the learning necessary for achieving such margins. Memories that rely on ECC for even one bit per cell operation have little margin built into the basic technology. Throughout the previous discussions, mention has been made of process manufacturing attributes for M.L.C. These attributes have been achieved by utilizing the same process flow as the standard one bit per cell flash memory. This approach has maintained shared learning and has lead to lower costs. In other words, low-cost process manufacturing was achieved through an understanding of M.L.C. requirements up-front in the design of the basic process architecture at the generation where M.L.C. is introduced. The tight manufacturing margins required for M.L.C. are a natural extension of the learning from manufacturing of error-free one bit per cell flash memory and are well within the manufacturing, equipment, and process module capability.

Standard Product Feature Set

One of the main challenges in implementing M.L.C. is maintaining product performance, usability, and reliability at the same levels as standard flash memories. If the implementation of M.L.C. resulted in a product that did not satisfy these goals, it would be relegated to a niche in the marketplace. Key features for a non-volatile memory are programming speed, read speed, power supply requirements, and reliability. This paper shows how our implementation of M.L.C. achieves these features. Before finishing, however, let us briefly discuss each one of them.

Programming Speed

Programming speed is achieved by choosing a placement algorithm that exploits stable device operating points to enable circuit performance optimization to occur, with little limitations of flash device operation. Parallel cell programming (32 cells, or 64 bits) at a time also amortizes the placement algorithm run time. The choice of charge sensing approaches also affects programming speed as it is integral to the placement algorithm. Sensing approaches other than those described in this paper can be used. An example would be a sensing scheme that varies control gate voltage to detect the threshold voltage directly. Such a scheme, while a more direct measure of floating gate charge, does not exploit the current drive capability of the flash cell, the drive used for sensing speed performance. To sum up, the choices of algorithms, optimizations, and architecture are what allow M.L.C. programming to be as good or better than one bit per cell flash memories.

Read Speed

As mentioned above, the choice of fixed control gate sensing and utilization of the flash cell's current drive capability allows fast read operation. In addition, parallel charge sensing allows for fast decode of the logic level, with little circuit overhead. As such, the read speed of Intel's StrataFlash memory is consistent with that of one bit per cell flash memories of comparable bit density.

Power Supply

As also discussed, the on-chip voltage generation and regulation is key to the implementation of M.L.C. One could specify an M.L.C. product that uses externally supplied precision voltages, but such a product would be more costly to the user, who would have to pay for the power supply, memory, and board space. Having the voltages generated and regulated on-chip allows for the Intel StrataFlash memory to plug directly into existing flash memory applications.

Reliability

Starting with high-yielding, low-defect memory, exhaustive cell studies and process and bias optimizations allow for an implementation of M.L.C. that achieves non-volatility and high reliability without requiring on-chip or system ECC. Thus the user can interface to the device with random memory location access, without latency for correction. Additionally, ECC requires overhead bits, which would diminish the cost advantages of M.L.C.

These standard flash memory product features, coupled with low-cost circuit design and manufacturing process implementation allow users to benefit from the low cost of M.L.C. without having to sacrifice needed features or performance.

Conclusion

It has been shown how Intel StrataFlash memory achieves multiple bits per cell, coupled with traditional process scaling, to provide an advance in memory cost reduction. The M.L.C requirements of precise charge placement, precise charge sensing and precise charge retention are achieved by exploiting stable device-operating points and direct access to the memory cell, employing mixed signal digital and analog design. Non-cell-related costs are held low by riding on the tight manufacturing margins developed for error-free one bit per cell flash memories. A standard product feature set ensures that the cost advantages of M.L.C. are available to the mainstream flash memory market.

Acknowledgments

The authors would like to thank the members of the M.L.C. development groups, whose dedicated work helped turn a few ideas into a product reality.

References

[1] Kynett, V.N., et. al., "An In-System Reprogrammable 256K CMOS Flash Memory," Technical Digest IEEE International Solid State Circuits Conference, 1988, pp. 132-133.

[2] Tam, S., Ko, P.K., and Hu, C., "Lucky-Electron Model of Channel Hot Electron Injection in MOSFET's," IEEE Transactions Electron Devices, September 1984.

[3] Lenzlinger, M. and Snow, E.H., "Fowler-Nordheim Tunneling into Thermally Grown SiO_2," Journal of Applied Physics, vol. 40, No. 1, January 1967, pp. 278-283.

[4] Ong, T.C., et. al., "Erratic Erase in ETOX™ Flash Memory Array," IEEE VLSI Symposium, 1993, p. 145.

[5] Eitan, B. and Frohman-Bentchkowsky, D., "Surface Conduction in Short-Channel MOS Devices as a Limitation to VLSI Scaling," IEEE Transactions on Electron Devices, vol. ED-29, No. 2, February 1982, pp. 254-266.

[6] Wu, K., et. al., "A Model for EPROM Intrinsic Charge Loss Through Oxide-Nitride-Oxide (ONO) Interpoly Dielectric," 28th Annual Proceedings IEEE International Reliability Physics Symposium, 1990, p. 145.

[7] Mielke, N., "New EPROM Data-Loss Mechanisms," 21st Annual Proceedings IEEE International Reliability Physics Symposium, 1983, p. 106.

Authors' Biographies

Al Fazio is a Principal Engineer in Flash Technology Development. He received a B.Sc. in Physics from the State University of New York at Stony Brook in 1982 and joined Intel the same year. He has been involved in development programs including SRAM, EPROM, E^2PROM, NVRAM, and Flash Memories. He was responsible for the Technology Development of the Intel StrataFlash™ memory. He holds more than a dozen patents and has authored or co-authored several technical papers, two of which have received Outstanding Paper Awards at the IEEE International Reliability Physics Symposium and at the IEEE International Solid State Circuits Conference. He is presently responsible for Intel's Multi-Level-Cell and Advanced Flash Memory Cell Development and currently serves as General Chairman of the IEEE Non-Volatile Semiconductor Memory Workshop. His e-mail address is al_fazio@ccm.sc.intel.com

Mark Bauer is a Senior Staff Engineer in Flash Circuit Design. He received his B.S.E.E. from the University of California, Davis in 1985. He joined Intel's Memory Components Division that same year, working on EPROM design. He was responsible for Circuit Design Development of the Intel StrataFlash^{TM} memory. He holds more than a dozen patents in the field of non-volatile memories and has authored two technical papers, one of which received an Outstanding Paper Award at the IEEE International Solid State Circuits Conference. He is presently responsible for Intel's next generation Multi-Level-Cell Circuit Design. His e-mail address is mark_bauer@ccm.fm.intel.com