

**INTEGRATED CIRCUIT FOR STORAGE AND
RETRIEVAL OF MULTIPLE DIGITAL BITS
PER NONVOLATILE MEMORY CELL**

5

FIELD OF THE INVENTION

10 This invention relates in general to semiconductor memories and, in particular, to nonvolatile semiconductor memories with the ability to store multiple digital bits per memory cell.

BACKGROUND OF THE INVENTION

15 Nonvolatile semiconductor memories, such as EEPROM, EPROM and FLASH integrated circuits, have traditionally been used to store a single digital bit per memory cell. This has been done by changing the threshold voltage (conduction) characteristics of the
20 cell by retaining a certain amount of charge on the floating gate of the memory cell. The threshold voltage range is normally partitioned into two levels (conducting versus nonconducting) to represent the storage of one digital bit per memory cell.

25 A wide range of charge can be reliably stored on the floating gate to represent a range of threshold voltages. Charge retention on the floating gate can be partitioned to represent multiple number of threshold voltage ranges and the threshold range can be partitioned
30 into multiple ranges to represent storage of more than one bit of digital data per memory cell. For example, four threshold partitions can be used to represent storage of two digital bits per memory location and sixteen partitions to represent storage of four digital
35 bits per memory location. Furthermore, the threshold voltage range can be partitioned to appropriately finer

resolution to represent the direct storage of analog information per memory cell.

5 The ability to store multiple digital bits per memory cell increases the effective storage density per unit area and reduces the cost of storage per digital bit. In addition to this, in the field of semiconductor memories, the costs of a modern fabrication facility often exceeds a billion dollars. Application of multibit storage per cell techniques to existing memory
10 fabrication processes and facilities allows the production of the next generation of higher density storage devices in the same manufacturing facilities, thereby increasing profitability and the return on investment.

15 Nonetheless, the problem of operational speed, i.e., the reading and writing operations, have yet to be satisfactorily addressed for devices having multiple bits per memory cell. A related problem is power dissipation. As more power is used to increase operational speeds,
20 power consumption is also undesirably increased. Still another problem is reliability. While charges can be stored in the floating gates of memory cells for very long periods, erasing and rewriting charges causes long term problems as to the certainty of the bits stored in a
25 memory cell. And, of course, any integrated circuit has problems of space. In an integrated circuit having multiple bits per cell, additional circuits must be added to handle the new requirements. This partially negates the advantages of the increased bits per memory cell.

30 The present invention solves or substantially mitigates these problems. The present invention speeds up the reading and writing operations of multibit memory cells. Power dissipation is lowered for reading operations. The present invention also permits the
35 reliable determination of the bits in the memory cells

over the long term and also conserves space on the integrated circuit.

SUMMARY OF THE INVENTION

5 The present invention provides for an integrated circuit having an array of memory cells, each memory cell storing multiple bits of information, and at least one data terminal. The integrated circuit also has a plurality of latches connected to the array of memory
10 cells with the latches organized into a first bank and a second bank. For reading and writing operations from and into the memory cell array, the latches and memory cell array are controlled so the first bank is coupled to the array of memory cells while the second bank is coupled to
15 the data terminal. Alternately the second bank to the array of memory cells while first bank is coupled to said one data terminal. This alternate coupling permits data to be simultaneously transferred between one bank of latches and the array of memory cells and transferred
20 between another bank of latches and the data terminal for faster read and write operations.

 To lower power dissipation, the memory cells of the array are read by voltage-mode operation. Furthermore, during writing operations, a voltage
25 corresponding to the amount of charge stored in the selected memory cell is compared to a reference voltage to determine whether high voltage programming of the memory cell should continue. Programming of the memory cells is terminated when the corresponding voltage
30 matches the reference voltages.

 For reading operations, the voltage corresponding to the amount of charge stored in a selected memory cell is compared to a sequence of reference voltages in a binary search pattern to

determine the plurality of bits stored in the memory cell.

BRIEF DESCRIPTION OF THE DRAWINGS

5

Fig. 1 is a block diagram illustrating the major circuit blocks implemented on a single integrated circuit chip according to the present invention;

10

Fig. 2A shows a circuit generally illustrating current-mode reading of the memory cells in Fig. 1; likewise,

Fig. 2B shows a circuit generally illustrating voltage-mode reading of the memory cells in Fig. 1;

15

Fig. 3 shows the organization of the reference cells and array cells within a block and the connection of the threshold partition voltage reference generation blocks to their respective arrays;

Fig. 4 is a block diagram of the multilevel dual mode shift registers in Fig. 1;

20

Fig. 5 illustrates the general organization of two Y-drivers in Fig. 1;

Fig. 6 shows details of the multilevel dual shift registers in Fig. 4 and circuitry that allow the dual shift registers to be used both during writing and reading operations;

25

Fig. 7 illustrates the reference multiplexer circuit in Fig. 5 for each Y-driver;

30

Fig. 8A shows the circuit details of the voltage comparator, the latch, the program and read control block and the high voltage switch, which are common to each Y-driver; Fig. 8B shows the circuit level detail of the voltage comparator, the latch, the program and read control block and the high voltage switch and the read mode path for the reference Y-drivers with additional circuitry which allow all reference cells in a block to be read in parallel; and Fig. 8C shows the

35

details of the Y-multiplexer circuit of a reference Y-driver and Y-multiplexers;

Fig. 9A shows details of the Y-multiplexer common to all Y-drivers, the X-Decoder block, the X-multiplexer common to each X-decoder and memory cells common to one Y-driver and one X-decoder with connections to the reference Y-multiplexer and reference cell array; and Fig. 9B shows the circuit of a single transistor memory cell according to one embodiment of the present invention;

Fig. 10 is a scale from 0 Volts to V_{max} Volts of the various program threshold partition voltages for the reference memory cells and the data storage memory cells;

Fig. 11 details the threshold partition voltage reference generation blocks; and

Fig. 12A represents the tree decoding in the binary search algorithm in a read operation to determine the digital bits stored a selected memory cell; and Fig. 12B is a flow chart for the binary search algorithm for a read operation.

DESCRIPTION OF THE PREFERRED EMBODIMENTS

It should be noted that the drawings have the elements with same reference numeral. This emphasizes the similar structure or operation of the elements. Furthermore, the symbol for a MOS transistor has been somewhat modified to a straight line representing the source and drain of the transistor and a short line parallel to the source/drain line to represent the gate of the transistor.

General Description of the Integrated Circuit

The major blocks of a preferred embodiment of the present invention are shown in Fig. 1. A nonvolatile

memory array 1 and a reference memory array 2 has memory cells connected in a two dimensional array of rows and columns. The memory cells can be any of the existing device architectures, such as, for example, EPROM, EEPROM, FLASH, or existing cell structures, such as single transistor, two transistor, split-gate, NAND, AND, and DINOR cell structures, or ground array architecture, including standard and virtual ground, known in prior art. Depending on which device architecture, cell structure or ground array architecture, is chosen, specific programming, erase and read algorithms can be easily developed, including the specific voltages required at each of the electrical terminals of the cell to facilitate the storage of more than one digital bit per nonvolatile memory cell. A cell can hold more than one nonvolatile device, for example, a NAND, DINOR or AND cell structure, already known in prior art. The specifics of the device, array architecture or cell structure and algorithms are not part of the present invention.

Each of the memory arrays 1 and 2 are further organized into blocks, having single or multiple rows. Each block consists of all of the columns or part of the columns of the arrays 1 and 2. In Fig. 1 a memory block is shown with all the columns in a single row. Each memory block consists of the cells from the reference array 2 and the cells from the memory array 1.

An error correction array 3 has nonvolatile memory cells similar to those used in the memory 1 and the reference array 2. In one embodiment, the error correction array 3 contains additional coding information required for an on-chip Error Correcting Code (ECC) mechanism, as is known in prior art for ECC implementation. In another embodiment, the error correction array 3 contains the full address of defective

cells which should be avoided during a write or a read operation. The size of the error correction array 3 depends on the maximum number of defective cells which may be corrected. During a production verification phase, the memory array 1 is tested to identify any defective cells. The address of these defective cells are programmed into the error correction array 3 before the chip is shipped from the factory. The error correction array 3 may be programmed using more than one bit per memory cell or may be programmed using a single bit per cell. If ECC correction is implemented, the error correction array 3 is automatically loaded with coding bits with on-chip ECC circuitry. An error correction control and logic block 16 contains all the necessary addressing, decoding and sequencing circuitry necessary to implement either one of the error corrections embodiments mentioned above.

A memory management array 4 contains address information for the blocks that are available for further writing at a certain time and also physical address information for blocks during sequential writing or reading of multiple blocks which are not necessarily physically contiguous in the memory array but are logically contiguous. Memory management of the array improves the long term reliability of the product and also allows for more efficient use of the memory in environments where serial data of variable length are frequently erased and rewritten. In such operations, only the beginning and ending block addresses are provided and the data is accessed through clocking. Instead of providing the ending block address, a stop signal can also be used to signify the end of the variable block serial data. The mode is called the "serial write and read access" mode and is generally used for digital audio record and playback systems, and also

for semiconductor memory systems which replace mechanical disks. The serial write and read access mode with error correction and memory management allows the present invention to substitute integrated circuit memory for digital audio record and playback systems and also for general digital data storage systems. A memory management logic block 24 contains the necessary sequencing circuitry to perform the memory management function in conjunction with the memory management array 4. The data in the memory management array 4 may be simply a single bit per memory cell or more than one digital bit per cell as in the arrays 1 and 2.

A redundancy block 5 has additional blocks of memory cells that can be used to repair whole blocks of cells which cannot be used. This sort of block redundancy is known to designers of memory integrated circuits. The number of blocks in the redundancy block 5 defines the maximum number of blocks that can be repaired either during the production verification phase or in the field during an embedded repair phase.

Addressing of the memory cells of the array 1 is provided by an address decoder 13 which is coupled to a serial interface block 14 which is connected to the external world. The decoded addresses are passed to a Y-counter block 12 and an X-counter block 11 from the decoder 13. The output from the Y-counter block 12 is passed to a Y-multiplexer block 8 which selects the desired block of memory cells in the array 1. The output of the X-counter block 11 is decoded in the X-decoder block 7 and the X-multiplexer block 6, to select the desired row in the selected block in the memory array 1.

The address decode block 13 generates the starting address of a selected row. The decoded address is set into the X-counter 11 and Y-counter 12 at the beginning of each new access operation of a certain

length of data stream. After the starting address is provided, data is serially accessed by a clock input to the chip. The serial interface block 14 contains the circuitry required to perform the appropriate serial protocol with other external chips. The serial protocol can be any of the industry standard serial protocols or a proprietary protocol. Generic serial interface signals are shown in Fig. 1 going in and out of the serial interface block 14.

The X-counter block 11 contains digital counters which increment their count by a clock signal YOUT, the output of the Y-counter block 12, on a line 27. The Y-counter block 12 is clocked by a signal CLCK on an input line 28 and generates a clock signal SHFT CLK on a line 29 to the various sections of the Y-drivers. The Y-counter block 12, in turn, provides the clock signal YOUT on the line 27 to the X-counter block 11.

The X-multiplexer block 6 provides the output of one X-decoder stage in an X-decoder block 7 on a selective basis to multiple rows of the array. This accommodates the circuitry of an X-decoder without letting the aspect ratio of the integrated circuit layout of the X-decoder from becoming inordinately large. X-multiplexers and their use are known in the prior art. The X-decoder block 7 contains the X-decoders which are used to select the rows of the memory array 1 and 2. Details about the X-decoder block 7 and the X-multiplexer block 6 are provided below and are also shown in Fig. 9A. The Y-multiplexer block 8, similar to the X-multiplexer block 6, selects the output of one of the Y-drivers, described in more detail below, and provides it on a selective basis to any one of a number of columns of the memory array. This is again done to fit the pitch of the Y-drivers and the memory array in the column direction.

A read-write circuit block 9 contains the necessary circuitry to perform high-voltage write and low-voltage read operations of data to and from the array 1. Details about the read-write block 9 are provided below.

A multilevel dual-shift register block 10 which has serially connected latches lies between the data input and output terminals and the memory array 1 and 2. Data to be written into the memory array 1 is moved serially through a DATA IN 25 line to the block 10 to the memory array 1. Data to be read from the memory array 1 is moved from the memory array 1 to the block 10 and then transferred serially from the block 10 through a DATA OUT 26 line. A detailed description is provided below. A system control logic block 15 contains the necessary control and sequencing circuitry to allow proper system operation. A test mode control and logic 17 block contains circuitry that allow full functional testing of the chip. Through the use of test modes, the chip is reconfigured into various alternative test configurations that allow faster and more efficient verification of the chip. These test modes are normally accessed in the verification phase at the factory but certain test modes may also be accessed in the field, such as for array repair test modes using the redundancy block 5.

A program/erase/read algorithm block 18 provides all the control and sequencing signals to perform the intelligent programming, erasing and reading of digital data from the memory array 1.

An oscillator block 19 generates clock signals for the high voltage generation and also provides clock signals for the program/erase/read algorithm block 18 and also for other system clocking and synchronization purposes. Alternatively, if the oscillator block 19 is

not placed on-chip, then its output signals must be supplied externally to the integrated circuit.

5 A charge pump 20 block generates high voltages on-chip. A high voltage shaping and control block 21 receives the output signal of the charge pump block 20 and properly shapes the high voltage pulses with predetermined rise and fall times. High voltage pulse shaping is critical for long term reliability of the operation of the integrated circuit. High voltages
10 shaped pulses can also be provided externally. Or unshaped high voltage can be provided from an external source, which can be then be pulse shaped with the proper rise and fall times with on-chip circuitry.

15 A nonvolatile scratch pad memory and registers block 22 has memory cells similar to those in the nonvolatile memory array 1. These memory cells are suitably organized and are normally used for external system house keeping and feature requirements. In an audio record and playback system, for example, the
20 nonvolatile scratch pad memory and registers block 22 contains the information for the number of messages and the time at which these messages were recorded. Data in the scratch pad memory and registers can be stored as single or multibit per memory cell.

25 An on-chip bandgap reference block 23 generates the necessary analog voltage and current references required for operation of the integrated circuit. These voltage and current references are used to provide reference voltages and currents which are compensated for
30 temperature and power supply variations. System performance is stabilized over wide temperature and power supply ranges.

General Reading Operation of a Memory Cell

Heretofore, current-mode operation is typically discussed for the reading of multiple bits per memory cell. Current-mode reading has an advantage of fast access times. Fig. 2A illustrates the general circuit arrangements for reading in the current-mode, using a single transistor memory cell. This general topology is applicable to other cell structures too.

A nonvolatile memory cell 30 is typically connected in an inverter mode. The voltage V_s at the source 31 of the transistor, which forms the cell 30, is connected to ground. The control gate 36 of the memory cell 30 is connected to a suitable voltage, V_g , or switched to the power source voltage. The drain of the memory cell 30, which also forms part of a column line 32 of the memory array of which the memory cell 30 is a part, normally is connected to a current sense amplifier 33. The nonvolatile memory cell 30 is connected to the column line 32 through some selection circuitry (not shown here for simplicity's sake). The current sense amplifier 33 typically is also connected to a reference current input line 34 for comparison purposes. The result of the comparison between the column line 32 current through the nonvolatile memory cell 30 and the reference current line 34 is generated as a logic level at the logic output line 35.

For a single bit per cell, the simple absence or presence of current through the memory cell 30 is determined. For multibits per memory cell, the amount of current passing through the cell 35 is compared against a set number of currents by changing the reference current at the input line 34. The signal at the logic output 35 is then decoded to determine the stored bits. For example, U.S. Patent No. 5,172,338 by Mehrotra et al. teaches multibit reading schemes using current-mode

reading and also shows various alternative embodiments. However, while current-mode reading could be used in the present invention, reading of memory cells in the voltage-mode is preferred. This lowers power consumption compared to the current-mode technique and renders the multibit per cell memory integrated circuit more appropriate for low power, relatively slower access applications, such as audio record and playback systems, and mechanical magnetic disk replacement systems.

In voltage-mode reading, the nonvolatile memory cell 30 is connected in the source follower mode, as illustrated in Fig. 2B, using a single transistor memory cell. The general voltage-mode topology is also applicable to other cell structures. The source 31 of the transistor forming the cell 30 is connected to a regulated supply at voltage V_s from a stable voltage reference, such as a bandgap reference. The control gate 36 is also connected to the same supply voltage as the source 31 or a voltage that is high enough to allow the accurate reading of the highest expected voltage V_d at the drain of the cell 30. A stable fixed bias current circuit 37 is connected between ground and the transistor's drain, which also forms part of the column line 32 of the memory array, as in Fig. 2A. The amount of fixed bias current is small, in the range of 0.5 microampere to 5.0 microampere. This small current prevents undue cumulative trapping of electrons during multiple read cycles, thereby preventing false readings of the memory cell 30. The voltage at the drain, which is also connected to the column line 32 through selection circuitry (not shown here), is equal to $V_g - V_{gd}$, where V_{gd} is the gate-to-drain voltage of the memory cell 30 required to source the current drawn by the bias current circuit 37. The drain of the transistor, part of the column line 32, is connected to an input terminal of a

voltage sense amplifier 38. The voltage sense amplifier 38 also has a reference voltage input line 39 and a logic out put line 40. Voltages at the transistor drain, the column line 32, and the reference voltage line 39 are compared and the resulting logic output signals are provided at the logic output line 40. The current required for reading in the voltage-mode is much less than for the current-mode. Thus reading in the voltage-mode has lower power dissipation.

The voltage read out at the line 32 depends on the amount of negative charge (electrons) on the floating gate 36 of the non-volatile memory cell 30. A large amount of charge on the floating gate increases the threshold voltage of the cell 30. The higher threshold voltage increases the gate-to-drain voltage V_{gd} of the cell 30. The voltage at the line 32 is then lower with respect to ground. Conversely, when the amount of charge in the floating gate is low; the threshold voltage of the cell 30 is lowered and the V_{gd} is decreased. The voltage at the line 32 is then higher with respect to ground. By controlling the amount of charge on the floating gate, suitable read back voltages are generated at the line 32. The process of injecting negative charge (electrons) into the floating gate is referred to as "erasing" and the process of removing charge from floating gate is referred to as "programming" the floating gate or memory cell.

During a multibit reading from a single memory cell, the voltage at the transistor's drain is compared to various voltages at the reference voltage line 39. The logic output at the line 40 is then decoded to provide the appropriate bits. With the source follower connection of the memory cell 30, data access is slowed because the whole column line 32 must be pulled up through the small memory cell. For certain applications, this slower access rate is acceptable. As described

below, the multilevel dual shift registers effectively improve the read access times.

Organization of the Memory Arrays

5 Fig. 3 illustrates the organization of the nonvolatile memory array 1 and the nonvolatile reference array 2. The memory cells in the reference memory array 2 are used to generate the comparison reference voltages for a voltage sense amplifier to determine the bits
10 stored in the memory cells selected in the array 1. In the preferred embodiment described here, four bits are stored per memory cell of each array 1 and 2. As mentioned previously, each block in the preferred embodiment consists of a row. Each row consists of
15 reference memory cells and array memory cells. All the cells in a row are erased simultaneously, and depending on the Y-multiplexer multiplexing scheme only part of the row is programmed and read from simultaneously. Since four bits are stored per memory cell, there are sixteen
20 reference memory cells per row. In this embodiment, each Y-driver drives eight memory cells so there are two Y-drivers 42 for a row of sixteen cells in the reference array 2. These Y-drivers 42 are labeled REFY-DRIVERS. In Fig. 3 only three Y-drivers 41 for the memory array 1
25 are illustrated. There are M Y-drivers 41. The three memory array Y-drivers shown are labeled Y-DRIVER0 to Y-DRIVER2. A reference threshold partition voltage generation block 44, part of the bandgap reference block 23 of Fig. 1, drives sixteen reference lines, each with
30 one of the reference voltages REFB0-REFB15, into the REFY-DRIVERS 42 and an array threshold partition voltage generation block 43, also part of the block 23 of Fig. 1, drives the sixteen reference lines, each with one of the reference voltages REFA0-REFA15, into the array Y-drivers

41. The voltage relationships between the REFA0-15 and REFB0-15 signals is shown in Fig. 10.

During a writing operation, a WRITE signal on WR line 46 is high, which turns on a set of N-channel transistors 45 (outlined by a dashed rectangle). The sixteen REFA015 reference voltages of the block 43 are passed to the Y-driver reference voltage lines, RFL015. These reference level voltages, REFA0-REFA15, from the block 43 are selectively programmed into the memory array 1 cells. Likewise, the reference voltages, REFB0-REFB15, from the block 44 are selectively programmed into the reference cells of the array 2.

During a reading operation, the WRITE signal on the WR line 46 is driven low to turn off the transistors 45. Instead, a set of transistors 47 (also outlined by a dashed rectangle) are turned on to pass the reference REFB0-15 output voltages stored in the reference cells of the array 2 to the Y-driver 41 reference voltage lines, RFL015. The REFB0-REFB15 voltages stored and read back from the cells of the reference array 2 are used as reference voltages to ascertain the digital bits stored in the cells of the memory array 1 through a binary search technique described below. The use of reference cells per block, or row as in the preferred embodiment, cancels power supply and temperature variations by placing such variations in the common mode. The memory cells in both array 1 and 2 are subject to the same variations. The reference cells in array 2 are also subject to the same number of program and erase cycles as that of the memory cells in the array 1, thereby placing the long term aging effects of the cells in a block or row in the common mode. This reference mechanism has the advantage of lower current read back mode and allows for longer and better long term reliability and accurate read back of digital bits, compared to previously described

techniques. The on-chip threshold voltage generation (temperature and power supply compensated) blocks 44 and 43 also create higher reliability compared to prior efforts in this field. The blocks 44 and 43 do not use nonvolatile memory cells to generate threshold partition voltages, but rather depend on much more reliable and stable components, such as resistors, operational amplifiers and bandgap voltage sources. Thus the present invention has improved long term reliability and accuracy, and stability over temperature and power supply variations.

In another embodiment of the present invention, the cells of the reference array 2 are first programmed. Then the output of the programmed reference cells from the array 2 are used to selectively program the cells of the memory array 1, with an offset to place the programmed levels midway between the programmed reference levels, as indicated in Fig. 10. This method does not require the block 43 but requires additional time to program the reference cells first.

Dual Shift Registers for Data

Fig. 4 is a block level representation of the multilevel dual shift registers block 10, shown in Fig. 1 and part of each of the Y-drivers 41 of Fig. 3. The multilevel dual shift register block 10 has latches which are organized into two banks, A and B. Each bank of latches is connected serially to form a large shift register. Each bank has four latches for each Y-driver 41. In Fig. 5, for each Y-driver 41, during a writing operation the data enters serially through the dual shift registers of block 10 and during reading operations the data exits serially through the dual shift registers of block 10. The data information travels from top to bottom within each Y-driver 41 during writing operations

and from bottom to top during reading operations. In general, signals common to all Y-drivers 41 travel horizontally.

Of course, the depth of the Y-driver latches depends on the number of bits stored in one memory cell. In the preferred embodiment four bits are stored in each cell. Therefore, four latches exist per each Y-driver 41. For example, in Fig. 4, the Y-driver 0 has four serially connected latches 60-63 and the Y-driver 1 has four latches 65-67. Continuing further, Y-driver M-1 has the last four latches connected serially. M is the number of Y-drivers and therefore, the total number of latches is $4 \times M$. It is important to note that all the latches are connected across all of the Y-drivers 41 of a bank in a long serial link to form a shift register. True and complementary outputs of every latch are parallel, as described below with respect to Fig. 6.

The two shift registers, bank A and bank B, are connected through transmission switches 145 and 146 to the DATA IN line 25 and DATA OUT line 26, respectively. When a REGSEL control line 147 is high, the DATA IN line 25 and the DATA OUT line 26 are connected to the bank A shift register through the switches 145. When the REGSEL line 147 is low, the DATA IN line 25 and the DATA OUT line 26 are connected to the bank B shift register through the switches 146. The SHFT CLK signal on the line 29 clocks the shift registers. With every cycle of the SHFT CLK signal, the data bits move to the next latch. For example, the bit in latch 60 moves to latch 61 and the bit previously in latch 61 moves on to latch 62 and so on. In the normal operation of the dual shift registers, one bank always operates in the serial mode and the other bank in the parallel mode. The bank which is in the serial mode, receives data from, or reads data out of, the data terminals connected to the DATA IN and

DATA OUT lines 25 and 26 serially. At the same time, the other bank in the parallel mode receives data from, or loads data into, the memory cells of the array 1 in parallel. As the bank in the serial mode completes its serial operations on the data, the other bank simultaneously completes its parallel operations with the data to and from the array 1. Thereafter, the serial bank is switched to the parallel mode and the parallel bank is switched to the serial mode by changing the state of the REGSEL line 147. This synchronous switching from serial to parallel and vice versa occurs continuously during writing into and reading from the memory array 1. Since there are M Y-drivers, M memory cells are written in parallel. Since four bits are written per cell, a total of $4 \times M$ bits are written in parallel. This essentially provides a $4 \times M$ faster write rate compared to a single bit operation. Similarly, $4 \times M$ bits are read in parallel and then shifted out providing $4 \times M$ faster read rates. In fact, the read rate can be performed even faster by clocking the shift registers at a higher clock rate. The maximum clock rate is limited by the time required for the parallel data to be loaded into the latches for a serial shifting operations. Hence, as described above, the multilevel dual shift registers block 10 allows for faster read and write access times of the memory cell array 1.

The switching between bank A and bank B during both reading and writing operations can also be non-synchronous. For example, during writing operations, if the latches of the bank in the serial mode are loaded before the latches of the other bank in the parallel mode can program the memory cells with multiple bits, then the switch of serial and parallel modes between the two shift registers must wait until for the bank in the parallel mode part completes its programming operation.

Conversely, if the parallel mode programming operation is completed before the serial operations of the first bank are completed, then the parallel mode bank must wait until the serial mode bank is loaded with data. The same is true for read operations. Thus both synchronous and non synchronous operations of the dual shift register operation are possible through the implementation of the appropriate circuitry in the system control logic block 15 (shown in Fig. 1). Details of the latches 60-63 of the Y-driver0 and latches 64-67 of the Y-driver1 are shown in Fig. 6.

Data Between Dual Shift Registers and Memory Array

Fig. 5 illustrates the organization of the Y-drivers 41 with the multilevel dual shift registers block 10, the read-write block 9 and the Y-multiplexer block 8. The individual Y-drivers 41 are each the same in terms of operations and circuit detail. Only Y-driver 0 and Y-driver 1 are shown. The other Y-drivers up to Y-driver M-1 are represented by dashed lines.

Fig. 7 illustrates the circuit details of a reference multiplexer 50 of each read-write block 9 in a Y-driver 41. The true and complementary output signals of each of the latches within a Y-driver 41 are passed to a reference multiplexer 50. Depending on the particular bits in the four latches within a Y-driver 41 (in this case, Y-driver 0), the reference multiplexer 50 connects one of the reference voltage lines, RFL0-RFL15, to the RFLOUT output terminal of the multiplexer 50. Signals on the lines 60A, 61A, 62A, 63A and 60B, 61B, 62B, 63B carry the true and complementary output signals, AA, AB, BA, BB, CA, CB, DA and DB, respectively from the four latches of each Y-driver 41, as shown in Fig. 6.

The reference multiplexer 50 is essentially a 16-to-1 multiplexer, commonly known in prior art. As

apparent in Fig. 7, only one of the RFL0-15 signals appears as the output signal RFLOUT, depending on the signals, 60A through 63B, from the output terminals 60A-63B of the latches. Transistors T11 through T164 are N type transistors and the operation of the multiplexer 50 should be understood. The size of the multiplexer depends on the number of bits that are being stored in one memory cell. For example, a 6 bit per memory cell storage system requires a 64-to-1 multiplexer.

Fig. 8A shows the details of the Voltage Comparator 51, the Latch 52, the Program/Read control circuit 53 and the High Voltage Switch 54 of the read-write block 9. The circuitry in Fig. 8A is common to each of the Y-drivers 41. The Voltage Comparator 51 has transistors 70-76. Transistors 70 and 71 are P-channel transistors and the rest are N-channel transistors. A VBIAS voltage on a line 198 from the block 23 in Fig. 1 provides proper current biasing for the Voltage Comparator 51. The circuit of the Voltage Comparator 51 is known in prior art. Whenever the voltage on a signal line 200 to the gate of the transistor 73 is higher than RFLOUT voltage on the signal line to the gate of the transistor 72 by even a very small amount, then the SET output on the Voltage Comparator output line 199 is also high, and vice versa. The gate of the transistor 73 is normally called the non-inverting input and the gate of the transistor 72 is called the inverting input. The signal line 200 and the signal line 206 described below connect the non-inverting input to the Y-multiplexer 55. The two lines 200 and 206 form a path to read the multiple bits stored in the cells of the array 1. The inverting input receives the RFLOUT signal, the output of the reference multiplexer 50, as previously described. The SET output line 199 of the

Voltage Comparator 51 is connected to an input terminal, the gate of the transistor 80, of the Latch 52.

5 The Latch 52 has transistors 80 through 85. Transistors 82 and 83 are P- channel transistors and the rest are N-channel transistors. The Latch 52 is a classic cross coupled inverter type with an input node, the gate of the transistor 80, connected to the SET output line 199 and another input node, the gate of the transistor 85, connected to the RESET input line 202.

10 This latch circuit and its operations is well known to integrated circuit designers. The transistors 81 and 82 form one inverter and the transistors 83 and 84 form the other inverter. The output node of the Latch 52 is connected by a signal line 201 to the Program Read

15 Control circuit 53. When the signal on the SET line 199 is high or pulsed high, the Latch output on the output line 201 is high. When the RESET line 202 is high or pulsed high, the signal on the Latch output line 201 is low. The signals on the SET line 199 and the RESET line

20 202 are never high at the same time.

 The Program/Read Control circuit 53 has two AND gates 88 and 89 and two inverters 86 and 87. A PROG (program) line 204 is an input to this circuit. The signal on the PROG line 204 is high when the write mode

25 is active, i.e., a writing operation, and is low when the read mode is active, i.e., a reading operation. When PROG is high (write mode active), the output of the AND gate 88 depends on the state of the output line 201 from the latch 52. If Latch output line 201 is low, then the

30 output of the AND gate 88 on the line 205 is high if the PROG signal on the line 204 is high, and vice versa. When the signal on the PROG line 204 is high (write mode active), then the output of the AND gate 89 is low. The output line 203 of the AND gate 89 is connected to the

35 gate of a transistor 100. During writing operations, the

transistor 100 is turned off and does not allow signals to pass from the line 206, which is connected to the Y-multiplexer 55, to the line 200. Lines 200 and 206 form part of the read path.

5 The High Voltage Switch 54 has an inverter 90, two N-channel transistors 91 and 94, a capacitor 92 and a high voltage transistor 93. The High Voltage Switch 54 operates as a transmission gate which allows high
10 voltages on an HV line 209 from the high voltage shaping and control block 21 (Fig. 1) to pass to the line 206 when the line 205 is high, or blocks high voltages from the HV line 209 from passing to the line 206 when the line 205 is low.

15 Connected to the read path formed by the signal lines 200 and 206 are the transistors 101 and 102 which provide the current load to a selected nonvolatile memory cell during reading operations. A VB line 208 is a current bias line generated from the Bandgap Reference block 23 (Fig. 1) to the gate of the transistor 102. The
20 transistor 102 operates as a source of the load current during the read mode. The transistor 101 with its control gate connected to a VCTL line 207 acts as a switch to turn the load current on or off. Inverters 103 and 104 buffer the SET output on the line 199 from the
25 Voltage Comparator 51 and provides an output signal on a READ DATA line 210 during reading operations only. The line 210 is connected to its corresponding latches (see Fig. 6) and the line 206 to its corresponding
30 Y-multiplexer 55. Thus the transistors 101 and 102 act as the bias current circuit 37 and the Voltage Comparator 51 acts as the voltage sense amplifier 38 of Fig. 2B for reading operations in the voltage-mode.

35 Fig. 8B shows the read-write block 9 of the reference Y-drivers 42. The Voltage Comparator 51, Latch 52, Program Read Control 53 and High Voltage Switch 54

are same as that of the Y-drivers 41 for the memory array 1, but there are modifications to read eight reference memory cells at a time. During a reading operation, a reference Y-driver 42 reads all the reference cells connected to it. Since there are eight reference cells for each reference Y-driver 42 in the present embodiment, there are eight current loads formed by the transistors 111 and 112, each set of transistor output by dashed boxes. The eight VCTL0-VCTL7 lines are forced high to connect the current loads to their respective read lines 220-227.

During writing operations, only one of the reference cells is written to, as selected by the REF Y-multiplexer 56, shown in Fig. 8C, in reference Y-driver 42. Whenever any one of the control lines MCTL0-MCTL7 is high, the bit line side RVD (Fig. 9A) is connected to the read path lines 260-267 in Fig. 8B.

During a reading operation, all the VCTL0-VCTL7 and MCTL0-MCTL7 control lines are high; this allows all the reference cells to be read in parallel. All VCTL0-VCTL7 control lines high also places the current loads on the respective read paths of the reference cells. In a reading operation the READ signal 219 is also high to allow the read voltage from the reference cells to be passed to the RFL lines. Eight reference voltages read back from the reference cells 0-7 are passed to the RFL0-7 signal lines respectively through reference Y-driver0 and eight reference voltages read back in parallel from the reference cells 8-15 are passed to the RFL8-15 signal lines through reference Y-driver1. In the present embodiment it is assumed that the voltages REFB015 (Fig. 10) are programmed into the reference cells 0-15 respectively. With the READ signal on the line 219 high, the transistors 211 are OFF and thus the read back voltage signals do not pass to the

Comparator input line 200. Notice that transistors 110 and 93 have been placed similarly on all the lines to allow same functionality during a writing operation mode for all the reference cells as occurs to the memory cells in array 1 through the Y-drivers 41.

In the reference Y-multiplexer 56 shown in Fig. 8C, each MCTL signal drives three series transistors M1, M2, M3. This arrangement provides same impedance on the line as provided by the Y-multiplexer 55 for the array 1, since there are three transistors in series whenever a memory array 1 cell is selected by the Y-multiplexer 55. This achieve better write and, more importantly, read mode matching characteristics between the cells of the reference array 2 and the memory array 1. The inverters 103 and 104 in Fig. 8A have been removed in the Fig. 8B. This is because in read operations digital bits are read out from the cells of the memory array 1, whereas reference voltage levels are read out from the cells in the reference array 2.

Fig. 9A shows a Y-multiplexer 55 for the Y-driver 41 for the memory array 1. The Y-multiplexer 55 is similar to the reference multiplexer 50. In the present embodiment the Y-multiplexer 55 is 8-to-1. The type of the multiplexer varies (N to 1) depending on the cell size and also on the amount of circuitry in the Y-drivers. For the described Y-multiplexer, a single transmission path is connected between the line 206 and one of the lines, VD0 through VD7, depending on the Y-address signals M0A-M2A and M0B-M2B from the Y-counters. VD0 through VD7 are the column lines in the memory array 1. During program and erase operations, the signals pass from the line 206 to the VD0-7 lines. During a read operation, signals pass from the VD0-7 lines to the line 206.

Fig. 9A also shows connections to a certain number of the nonvolatile memory cells of the array 1. In this embodiment, one Y-driver drives eight columns and one X-decoder drives four rows of the array 1. Each row is considered to be a block in the present embodiment. In other embodiments, multiple rows may form one single block. The selection of the rows by a single X-decoder is performed by the X-multiplexer 58 receiving four X-address signals, PA through PD, from the X-counters, as described previously. This basic topology can be extended in both the X-direction to increase the number of rows in the array and in the Y-direction to increase the number of columns, in order to increase the size of the array.

Fig. 9A also shows the reference array 2 and the reference multiplexers 56. There are sixteen reference cells from the reference array per block. Whenever a block is selected through the X-multiplexers 58, both reference and array cells are selected. The MCTLO-MCTL7 lines drive the reference Y-multiplexer 56.

For the embodiment described here, there are eight times more cells within one row than the number being programmed at one time. The Y-drivers 42 and 41 program every eighth cell in a row. A total of eight programming cycles are required to program all the cells in a row. Thus cells 0, 8, 16... are programmed in the first programming cycle. Cells 1, 9, 17... are programmed in the second programming cycle and so on. Eight programming cycles program one row. At the same time, the reference cells 0 and 8 are programmed in the first programming cycle. The reference cells 1 and 9 are programmed in the second programming cycle and so on until eight programming cycles complete the programming of all sixteen reference cells.

The latches of the REF Y-DRIVER0 and REF Y-DRIVER1 are set to output 0 and 8 respectively during the first programming cycle, to 1 and 9 respectively during the second programming cycle, and so to set the reference multiplexer of the reference Y-drivers 42 to select the proper RFLOUT voltage at the multiplexer's 50 output terminal from the REFB0-15 voltages provided by the reference generation block 44 shown in Fig. 3. During this writing operation, the latches of the reference Y-driver 42 are internally set to program the appropriate voltages into the reference cells at the selected locations in the array 2. At the same time, the latches of the Y-driver 41 are set externally by the data which is to be stored in the memory array 1. Of course, the number of programming cycles for a row is dependent upon the ratio of the Y-multiplexer. An 8:1 Y-multiplexer requires eight programming cycles, while a 16:1 multiplexer requires 16 programming cycles.

20 Reading Operations from the Memory Array

To further appreciate the voltage mode reading method in the details of the circuitry, reference should be made to Fig. 9A. The source line which is common for the both array 1 and 2 in the preferred embodiment is connected to a regulated supply voltage Vs. The connection to the transistor 35 of the cell of the arrays 1 and 2 are shown in Fig. 9B. Assuming that the cell circled and marked XX in the array 1 is being read. The X-multiplexer 58 selects block 2 through line VG2, also called the word line. A word line is connected to the control gate of each of the memory cells in the block. The selected word line is connected to the same supply as is connected to the source, i.e., Vs, or to a voltage that is high enough to allow the accurate reading of the highest expected voltage at the column line VD4 with

respect to ground. The Y-multiplexer 55 connects the column line VD4 to the line 206. Referring to Fig. 8A now, the line 206 is connected to the line 200 through the turned ON transistor 100. During a read operation, the PROG line 204 is low and RESET line 202 is high. This forces the gate 203 of the transistor 100 to be high to turn the transistor 100 on. The combination of transistors 101 and 102 form a current source (represented as the bias current circuit 37 in Fig. 2B) between the line 200 and ground. The line 200 also is connected to the non-inverting input of the Voltage Comparator 51 (represented as the voltage sense amplifier 38 in Fig. 2B). The transistor 101 acts as a switch for the current source. The transistor 101 is only turned on for a short period of time to accomplish proper voltage comparison by the Voltage Comparator 51. Power dissipation and also the potential for charges to be trapped in the oxide layer of memory cell transistor is minimized. The RFLOUT input (represented as the reference voltage 39 in Fig. 2B) connected to the inverting input terminal of the Voltage Comparator 51 is the voltage read back from the appropriate reference cell as selected through the reference multiplexer 50 from one of the lines RFLO-15 as shown in Fig. 5 and Fig. 7. The result of the comparison at the Voltage Comparator 51 is placed on the read data line 210 (represented as the logic output 40 in Fig. 2B). During read operations, the high voltage switch 54 is turned OFF and the high voltage line 209 is disconnected from line 206 by the high voltage transistor 93.

The dual shift registers, described previously, of the block 10 are used both in the write and the read operations in order to reduce the number of devices in the integrated circuit. The operation of the dual shift registers during a writing operation, has been described

previously. In a read operation (refer to Fig. 6), the four latches in a Y-driver 41 are preset through the operation of the binary search algorithm. The signals BIT3, BIT2, BIT1, BIT0 are forced high sequentially according to the binary search algorithm shown in Figs. 12A and 12B. The operation begins with a RESET pulse on the RESETB line of one bank of the multilevel dual shift registers. The RESET pulse resets all the latches on one bank of the dual shift registers. According to the binary search algorithm, the BIT3 signal is forced high. This sets the line 63A high and the line 63B low for all the latches connected to the BIT3 signal line, latches 0, 4, 8 and so forth), in all the Y-drivers 41. The voltage on the RFL8 line of the reference multiplexers 50 is thus selected for the RFLOUT terminal of each Y-driver 41.

At the same time during this read operation, the RFL0-15 lines are driven in parallel by the voltages read back from the cells of the reference array 2, as described previously. According to the binary search algorithm, if voltage read back from the memory cell is higher than the selected voltage on the RFLOUT within each driver, then the data output on the READ DATA line 210 in each Y-driver 41 is high. This forces the output terminal 601 of the NAND gate 600 low (see Fig. 6), which sets the latch connected to the BIT3 line. The signal at the output terminal 602 of the latch remains high even when the data on the line 210 is removed. Once the latch is set, the signals at the output terminals 63A and 63B remain high and low respectively, even when the line BIT3 is forced low. If the voltage read back from the memory cell is lower than the voltage at the RFLOUT terminal, then the signal on the READ DATA line 210 is low. This forces the signal at the output terminal 601 of NAND gate 600 to stay high and the latch to remain reset. Thus when the signal BIT3 is forced low, the signals at the

latch output terminals 63A and 63B would be low and high respectively, the reset state of the latch. The binary search algorithm continues by forcing the BIT2, BIT1 and BIT0 lines high respectively. A compare operation of the voltages on the READ DATA line 210 and on the RFL0UT line within each Y-driver 41. The connected latches are set if READ DATA 210 is high or left reset if RESET DATA line 210 is low. Depending on the set or reset states of the latches within each Y-driver 41, a different voltage from the RFL0-15 lines is selected on the RFL0UT terminal through the reference multiplexer 50 inputs 63A,B to 60A,B (the output of the latches).

Sequentially four bits from a single memory cell are read into the four latches within each Y-driver 41. If N bits were stored per memory cell, then there would be N latches per Y-driver 41 and N bits per y-driver would be read in N cycles of the binary search algorithm. All the M Y-drivers 41 are simultaneously loading their respective latches. After the latches on one bank of the dual shift registers are loaded, the bank is placed in the shift mode and the latched data is then serially clocked out from this bank. While the data is being shifted out, the other bank of the dual shift registers is placed in the parallel read mode and the data of another M cells are read into the latches of this bank. As this bank completes loading its latches, the previous bank simultaneously completes its shifting operation. This alternating operation of parallel loading of data from the memory cells and serial shifting of data provides very fast read access times.

During a read operation, the state of the four latches within each reference Y-driver 42 is not used. The RFL0UT lines are not used within the reference Y-drivers 42. Instead, the voltages read from all the

reference cells are placed on the RFL0-15 lines as shown in Fig. 8B and described previously.

Writing Operations into the Memory Array

5 For a write operation, the programming and erase algorithms, as is known in prior art, typically use a repetitive high voltage pulsed program cycle, followed by a normal read cycle, to set the threshold voltages of nonvolatile memory cells with a high degree of accuracy. 10 Prior to initiation of the programming algorithm, an erase pulse of sufficient amplitude and duration is normally applied to completely erase the memory cells. Instead of one erase pulse, some algorithms also repeat the high voltage erase pulse followed by a read operation 15 as necessary for the erase function. In the present invention a single erase pulse is used and then a repetitive high voltage pulse programming algorithm is applied to accurately set the threshold voltages. Also in the present embodiment the erase programming and 20 reading occurs on a block basis for faster write and read access times. Thus M memory cells, representing 4 times M digital bits are simultaneously written to or read from.

 After an erase cycle has erased all the memory 25 cells in a block, the programming cycle is performed. Initially the Latch 52 (detailed in Fig. 8A, 8B) in each Y-driver is reset by pulsing the RESET 202 line. Thereafter, on a repetitive basis after the application of each programming pulse a read cycle is performed. 30 Within each Y-driver 41 and reference Y-driver 42, the read cycle is performed to determine whether the memory cell has reached the desired voltage level set at the RFL0UT output of the reference multiplexer 50. If the voltage level read back on the line 200 (Fig. 8A, 8B) has 35 not reached the RFL0UT level, then the Latch 52 remains

reset and additional high voltage pulses are impressed upon the memory cell. The source of the high voltage pulses is the high voltage shaping and control block 21 described in Fig. 1.

5 During any of the repetitions, if the read back voltage on the line 200 is higher than the voltage on the RFLOUT line, the Latch 52 is set and the high voltage switch of the respective Y-driver 41 (and reference Y-driver 42) is turned off. This stops further
10 transmission of high voltage pulses to the memory cell connected to that particular Y-driver. It should be understood that while certain Y-drivers may stop transmission of high voltage pulses to their respectively
15 connected memory cells, other Y-drivers may still be passing high voltage pulses to their respective memory cells in order to program the appropriate read back voltage levels. The read back mode during programming is exactly the same as during the normal read mode, except that the output signals on the READ DATA line 210 (Fig.
20 6) is not stored by the latches of the block 10. The use of same reading circuitry during programming and reading modes provides more accurate and reliable data storage and retrieval.

Fig. 10 shows the relationship between the
25 threshold partition reference levels for the nonvolatile reference array 2 memory cells and for the nonvolatile memory array 1 memory cells. The threshold voltage range to be partitioned is shown to be from 0V to Vmax. REFA0 to REFA15 are the threshold partition voltages for the
30 nonvolatile memory array 1 cells and REFB0 to REFB15 are the threshold partition voltages for the nonvolatile reference array 2 cells. The REFA0-15 levels are midway between the REFB0-15 levels. This ensures accurate and reliable long-term read comparison of the threshold
35 levels.

Fig. 11 shows the circuit details for threshold partition voltage generation blocks for both the nonvolatile memory array 1 and the nonvolatile reference array 2. A bandgap voltage reference unit 300 is an on-chip temperature and power supply voltage source. The operational amplifier OPAMP 301 is a high gain, unconditionally compensated amplifier. The circuitry for both the reference unit 300 and the operational amplifier 301 are known to integrated circuit designers. Resistors 302 through 318 are equal value resistors connected as shown.

The threshold partition generation block for the nonvolatile memory array 1 is formed when a resistor 318 is not connected in parallel to a resistor 317. The outputs are called REFA0 to REFA15. When the resistor 318 is connected in parallel with the resistor 317, then the nonvolatile reference array 2 threshold partition generation block is formed and the outputs are called REFB0 to REFB15. Through the programming algorithm, the sixteen reference cells per block in the preferred embodiment are programmed to each of the threshold partition voltages REFB0 through REFB15. The nonvolatile memory array 1 cells are programmed to any one of the REFA0 through REFA15 threshold partition voltage levels as defined by the bits in the latches within each Y-driver 41.

For the embodiment described here, there are eight times more cells within one row than the number of cells which are programmed at a time. The Y-multiplexers 55 program every eighth cell in a row. A total of eight program cycles are required to complete programming all the cells in a row. Thus, cells 0, 8, 16 and so forth are programmed during the first program cycle. Cells 1, 9, 17 and so forth are programmed during the second

program cycle and so on for eight program cycles to complete the programming of one row.

At the same time, the reference cells 0 and 8 are programmed through the two reference drivers 42 during the first cycle; cells 1 and 9 during the second cycle, and so on, as selected through the reference Y-multiplexers 56. The latches of the REFY-driver0 and REFY-driver1 are set to binary values "0" and "8" respectively during first cycle, and binary values "1" and "9" respectively during second cycle, and so on to set the reference multiplexer 50 of the reference Y-drivers 42. The multiplexer 50 selects the proper voltage from the REFBO-15 voltages provided by the reference threshold partition voltage generation block 44 for the RFLOUT output voltage. In other words, during the writing operation, the latches of the block 10 of each reference Y-driver 42 are internally set to program the appropriate voltages into the reference cells at the selected cell locations, while the latches of the block 10 of the Y-drivers 41 for the memory array 1 are set externally from the data to be stored in the memory array 1. The number of program cycles per row depends on the depth of the Y-multiplexers 55 and 56. For example, as described, an 8:1 multiplexer for the Y-multiplexer 55 requires 8 program cycles, while a 16:1 multiplexer would require 16 program cycles to finish programming a full row.

While various preferred and alternate embodiments of the present invention have been disclosed and described in detail, it should be evident that the present invention is equally applicable by making appropriate modifications to the embodiment described above. Therefore, the above description should not be taken as limiting the scope of invention which is defined by the metes and bounds of the appended claims.



What is claimed is:

- 1 1. An integrated circuit having an array of
2 memory cells, each memory cell capable of storing
3 multiple bits of information, and at least one data
4 terminal, said integrated circuit comprising:
5 a plurality of latches connected to said array
6 of memory cells, said latches organized into a first bank
7 and a second bank; and
8 control means alternately coupling said first
9 bank to said array of memory cells and said second bank
10 to said one data terminal, and said second bank to said
11 array of memory cells and said first bank to said one
12 data terminal whereby data is simultaneously transferred
13 between one bank of latches and said array of memory
14 cells and transferred between another bank of latches and
15 said data terminal for faster read and write operations,
16 said control means coupling each bank of latches to a
17 block of memory cells in said array, said block of memory
18 cells having M memory cells, and each bank of latches
19 having N x M memory cells, N latches coupled to each
20 memory cell.
- 1 2. The integrated circuit of claim 1 wherein
2 said control means couples a bank of latches to said data
3 terminal serially;
4 wherein during a writing operation, said
5 control means alternately couples one bank of latches to
6 said block of memory cells for a parallel transfer of
7 data from said bank of latches to said block of memory
8 cells, and couples another bank of latches to said data
9 terminal for a serial transfer of data from said data
10 terminal to said another bank of latches; and
11 wherein during a reading operation, said
12 control means alternately couples one bank of latches to



13 said block of memory cells for a parallel transfer of
14 data from said block of memory cells to said bank of
15 latches, and couples another bank of latches to said data
16 terminal for a serial transfer of data from said another
17 bank of latches to said data terminal.

1 3. The integrated circuit of claim 1 further
2 comprising first and second data terminals,
3 wherein during a writing operation, said
4 control means alternatively couples one bank of latches
5 to said block of memory cells for a parallel transfer of
6 data from said bank of latches to said block of memory
7 cells, and couples another bank of latches to said first
8 data terminal for a serial transfer of data from said
9 first data terminal to said another bank of latches; and
10 wherein during a reading operation, said
11 control means alternatively couples one bank of latches
12 to said block of memory cells for a parallel transfer of
13 data from said block of memory cells to said bank of
14 latches, and couples another bank of latches to said
15 second data terminal for a serial transfer of data from
16 said another bank of latches to said second data
17 terminal.

1 4. In an integrated circuit having an array of
2 memory cells, each memory cell storing a plurality of
3 bits, each memory cell having first and second terminals
4 and a control terminal, a circuit for reading said
5 plurality of bits comprising:
6 a bias current reference generating a bias
7 current through a selected memory cell independent of a
8 plurality of bits stored in said selected memory cell;
9 multiplexer circuits responsive to address
10 signals for connecting said selected memory cell in said
11 array to said bias current reference; and



12 a voltage comparator connected to a node
13 between said second terminal of said selected memory cell
14 and said bias current reference, said voltage comparator
15 further connected to a reference voltage for comparing
16 said reference voltage and a voltage at said node
17 corresponding uniquely to said plurality of bits stored
18 in said selected memory cell for said bias current to
19 determine a bit stored in said memory cell.

1 5. The reading circuit of claim 4 further
2 comprising:
3 means for sequentially changing said reference
4 voltage in an ordered sequence to determine a plurality
5 of bits stored in said selected memory circuit; and
6 a second memory array having a plurality of
7 memory cells storing a plurality of reference voltages,
8 said sequentially changing means connecting selected
9 memory cells in said second memory array to said voltage
10 comparator in said ordered sequence.

1 6. In an integrated circuit having an array of
2 memory cells, a circuit for programming an amount of
3 charge in a selected memory cell corresponding to a
4 plurality of bits, said circuit comprising
5 a high voltage circuit for generating a high
6 voltage for programming memory cells;
7 a bias current reference;
8 multiplexer circuits connecting said selected
9 memory cell to said bias current reference, said bias
10 current reference generating a bias current through said
11 selected memory cell independent of a plurality of bits
12 stored in said selected memory cell;
13 a voltage comparator connected to a node
14 between said selected memory cell and said bias current
15 reference to determine a voltage at said node



16 corresponding uniquely to said amount of charge stored in
17 said selected memory cell for said bias current;
18 a programming circuit connected to said high
19 voltage circuit and said voltage comparator, said
20 programming circuit in response to said voltage
21 comparator engaging said high voltage circuit to program
22 said selected memory cell until said voltage
23 corresponding to said amount of charge stored in said
24 selected memory cell matches said reference voltage.

1 7. In an integrated circuit having an array of
2 memory cells, each memory cell capable of storing
3 multiple bits of information, a plurality of latches
4 connected to said array of memory cells, said latches
5 organized into a first bank and a second bank, and at
6 least one data terminal, a method of operating said
7 integrated circuit comprising:
8 alternately coupling said first bank to said
9 array of memory cells and said second bank to said one
10 data terminal, and said second bank to said array of
11 memory cells and said first bank to said one data
12 terminal, each bank of latches coupled to a block of
13 memory cells in said array, said block of memory cells
14 having M memory cells, and each bank of latches having N
15 x M memory cells, N latches coupled to each memory cell;
16 and
17 simultaneously transferring data between one
18 bank of latches and said array of memory cells, and
19 transferring data between another bank of latches and
20 said data terminal for faster read and write operations.

1 8. The method of claim 7 wherein in said
2 coupling step
3 a bank of latches is coupled to said data
4 terminal serially;



5 during a writing operation, one bank of latches
6 is alternately coupled to said block of memory cells for
7 a parallel transfer of data from said bank of latches to
8 said block of memory cells, and another bank of latches
9 is coupled to said data terminal for a serial transfer of
10 data from said data terminal to said another bank of
11 latches; and

12 during a reading operation, one bank of latches
13 is alternately coupled to said block of memory cells for
14 a parallel transfer of data from said block of memory
15 cells to said bank of latches, and another bank of
16 latches is alternately coupled to said data terminal for
17 a serial transfer of data from said another bank of
18 latches to said data terminal.

1 9. In an integrated circuit having an array of
2 memory cells, each memory cell having first and second
3 terminals and a control terminal, each memory cell
4 storing a plurality of bits, a method for reading said
5 plurality of bits in memory cell, comprising:

6 connecting said first terminal of said memory
7 cell to a first voltage source;

8 connecting said control terminal of said memory
9 cell to a voltage source to electrically couple said
10 second terminal to said first terminal;

11 connecting said second terminal to a bias
12 current circuit, said bias current circuit generating
13 through said second terminal and said selected memory
14 cell a fixed bias current independent of a plurality of
15 bits stored in said selected memory cell;

16 connecting said second terminal to a circuit
17 for sensing a voltage at said second terminal relative to
18 a predetermined reference voltage corresponding uniquely
19 to said plurality of bits stored in said selected memory
20 cell for said bias current and generating a digital



21 output corresponding to said sensed voltage at said
22 second terminal.

1 10. In an integrated circuit having an array
2 of memory cells, each memory cell having a floating gate
3 capable of holding an amount of charge indicative of a
4 logic state stored in said memory cell, a method for
5 reading a memory cell selected from said array,
6 comprising:
7 generating a memory cell voltage responsive to
8 an amount of charge on a floating gate of said memory
9 cell by connecting said memory cell by a terminal of said
10 memory cell to a bias current reference so that a bias
11 current is generated through said memory cell independent
12 of said amount of charge on said memory cell floating
13 gate and so that voltage at said memory cell terminal
14 corresponds uniquely to said amount of charge on said
15 memory cell floating gate for said bias current; and
16 sequentially comparing said memory cell voltage
17 with respect to one of a plurality of reference voltages,
18 each sequentially comparing step determining a bit to
19 determine a plurality of bits responsive to said amount
20 of charge stored in said memory cell.

1 11. The method of claim 10 wherein said
2 sequentially comparing step comprises sequentially
3 comparing said memory cell voltage to four reference
4 voltages to determine four bits for said memory cell.

1 12. In an integrated circuit having an array
2 of memory cells, each memory cell having a floating gate
3 capable of holding an amount of charge indicative of
4 multiple bits stored in said memory cell, a method for
5 writing to a memory cell selected from said array,
6 comprising:



7 receiving a plurality of bits representative of
8 multiple bits to be stored in said selected memory cell;
9 generating a memory cell voltage responsive to
10 an amount of charge on a floating gate of said memory
11 cell;
12 generating a plurality of reference voltages in
13 said integrated circuit;
14 programming a floating gate of said memory cell
15 so that said memory cell voltage matches one of a
16 plurality of reference voltages, said one reference
17 voltage corresponding to said plurality of bits,
18 including
19 erasing any charge from said floating
20 gate;
21 subjecting said floating gate to high
22 voltage pulses;
23 determining said memory cell voltage by
24 generating a bias current through said selected
25 memory cell independent of the amount of charge
26 on the floating gate of said selected memory
27 cell, said memory cell voltage corresponding
28 uniquely to the amount of charge on the
29 floating gate of said selected memory cell for
30 said bias current;
31 comparing said memory cell voltage with
32 said one reference voltage; and
33 repeating said subjecting, determining and
34 comparing steps until said memory cell voltage
35 matches said one reference voltage.

1 13. An integrated circuit comprising
2 an array of memory cells, each memory cell
3 capable of holding a voltage corresponding to a plurality
4 of bits, said array organized into blocks, each block
5 having reference memory cells and data memory cells;



6 voltage generation circuits generating first
7 and second sets of reference voltage levels;
8 programming circuits simultaneously setting
9 voltages in said data memory cells with respect to said
10 first set of reference voltage levels, said voltages
11 corresponding to data bits, and setting said second set
12 of reference voltage levels in said reference memory
13 cells; and
14 reading circuits comparing voltages set in said
15 data memory cells with respect to said second set of
16 reference voltage levels in said reference memory cells
17 to determine data bits corresponding to said set voltages
18 in said data memory cells.

1 14. An integrated circuit comprising
2 an array of memory cells, each memory cell
3 capable of holding a voltage corresponding to a plurality
4 of bits, said array organized into blocks, each block
5 having reference memory cells and data memory cells;
6 voltage generation circuits generating a set of
7 reference voltage levels;
8 programming circuits setting said set of
9 reference voltage levels in said reference memory cells
10 and setting voltages in said data memory cells with
11 respect to said set of reference voltage levels, said
12 voltages corresponding to data bits; and
13 reading circuits comparing voltages set in said
14 data memory cells with respect to said set of reference
15 voltage levels in said reference memory cells to
16 determine data bits corresponding to said set voltages in
17 said data memory cells.

1 15. A method of operation in an integrated
2 circuit having an array of memory cells, each memory cell



3 capable of holding a voltage corresponding to a plurality
4 of bits, said array having reference memory cells and
5 data memory cells, said method comprising
6 receiving pluralities of data bits;
7 programming voltages in said data memory cells
8 with reference to a first set of reference voltage
9 levels, each voltage corresponding to a plurality of data
10 bits;
11 simultaneously programming a second set of
12 reference voltage levels in said reference memory cells;
13 and
14 comparing voltages programmed in said data
15 memory cells with respect to said second set of reference
16 voltage levels in said reference memory cells to
17 determine data bits corresponding to said voltages
18 programmed in said data memory cells.

1 16. The method of claim 15 wherein said
2 programming voltages step comprises simultaneously
3 programming voltages in all of said data memory cells in
4 a predetermined unit of said array and said set of
5 reference voltage levels in all of said reference memory
6 cells in said unit; and
7 said comparing step comprises simultaneously
8 comparing voltages in all of said data memory cells in
9 said unit with respect to said set of reference voltage
10 levels in all of said reference memory cells in said
11 unit.

1 17. A method of operation in an integrated
2 circuit having an array of memory cells, each memory cell
3 capable of holding a voltage corresponding to a plurality
4 of bits, said array having reference memory cells and
5 data memory cells, said method comprising
6 receiving pluralities of data bits;



7 programming a set of reference voltage levels
8 in said reference memory cells;
9 programming voltages in said data memory cells
10 with reference to said set of reference voltage levels,
11 each voltage corresponding to a plurality of data bits;
12 and
13 comparing voltages programmed in said data
14 memory cells with respect to said second set of reference
15 voltage levels in said reference memory cells to
16 determine data bits corresponding to said voltages
17 programmed in said data memory cells.

1 18. The method of claim 17 wherein said
2 programming voltages step comprises simultaneously
3 programming voltages in all of said data memory cells in
4 a predetermined unit of said array and said set of
5 reference voltage levels in all of said reference memory
6 cells in said unit; and
7 said comparing step comprises simultaneously
8 comparing voltages in all of said data memory cells in
9 said unit with respect to said set of reference voltage
10 levels in all of said reference memory cells in said
11 unit.

1 19. An integrated circuit comprising
2 an array of memory cells, each memory cell
3 capable of holding a voltage corresponding to a plurality
4 of bits, said array organized into blocks, each block
5 having reference memory cells and data memory cells;
6 voltage generation circuits generating first
7 and second sets of reference voltage levels;
8 programming circuits simultaneously setting
9 voltages with respect to said first set of reference
10 voltage levels in said data memory cells, said voltages
11 corresponding to data bits, and setting a second set of



12 reference voltage levels in said reference memory cells;
13 and
14 reading circuits selectively comparing voltages
15 set in said data memory cells with respect to said second
16 set of reference voltage levels in said reference memory
17 cells to determine data bits corresponding to said set
18 voltages in said data memory cells, and with respect to
19 said first set of reference voltage levels to operate
20 with said programming circuits to set voltages with
21 respect to said first set of reference voltage levels in
22 said data memory cells.

1 20. An integrated circuit comprising
2 an array of memory cells, each memory cell
3 capable of holding a voltage corresponding to a plurality
4 of bits, said array organized into blocks, each block
5 having reference memory cells and data memory cells;
6 voltage generation circuits generating a set of
7 reference voltage levels;
8 programming circuits setting said set of
9 reference voltage levels in said reference memory cells
10 and setting voltages with respect to said set of
11 reference voltage levels in said data memory cells, said
12 voltages corresponding to data bits; and
13 reading circuits selectively comparing voltages
14 set in said data memory cells with respect to said set of
15 reference voltage levels in said reference memory cells
16 to determine data bits corresponding to said set voltages
17 in said data memory cells and to operate with said
18 programming circuits to set voltages with respect to said
19 set of reference voltage levels in said data memory
20 cells.