



(43) International Publication Date  
6 December 2012 (06.12.2012)

- (51) International Patent Classification:  
G06F 12/08 (2006.01) G11C 16/06 (2006.01)
- (21) International Application Number:  
PCT/US2012/036923
- (22) International Filing Date:  
8 May 2012 (08.05.2012)
- (25) Filing Language: English
- (26) Publication Language: English
- (30) Priority Data:  
13/118,721 31 May 2011 (31.05.2011) US
- (71) Applicant (for all designated States except US): **MI-CRON TECHNOLOGY, INC.** [US/US]; 8000 South Federal Way, MS 525, Boise, Idaho 83716 (US).
- (72) Inventors; and
- (75) Inventors/Applicants (for US only): **NEMAZIE, Siamack** [US/US]; 27872 Via Cortia Way, Los Altos, California 94022 (US). **TABRIZI, Farshid** [US/US]; 2173 Shadow Ridge Way, San Jose, California 95138 (US). **IMAN, Berhanu** [US/US]; 1020 Lupine Drive, Sunnyvale, California 94086 (US). **SHAH, Ruchir** [IN/US]; 1303 Fairway En-

trance Drive, San Jose, California 95131 (US). **BENSON, William E.** [US/US]; 603 North Claremont, San Mateo, California 94401 (US). **GEORGE, Michael** [US/US]; 6941 Corte Madrid, Pleasanton, California 94566 (US).

- (74) Agent: **BOLVIN, Kenneth W.**; Leffert Jay & Polglaze, P.A., P.O. Box 2230, Minneapolis, Minnesota 55402-0230 (US).
- (81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.
- (84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ,

[Continued on next page]

(54) Title: DYNAMIC MEMORY CACHE SIZE ADJUSTMENT IN A MEMORY DEVICE

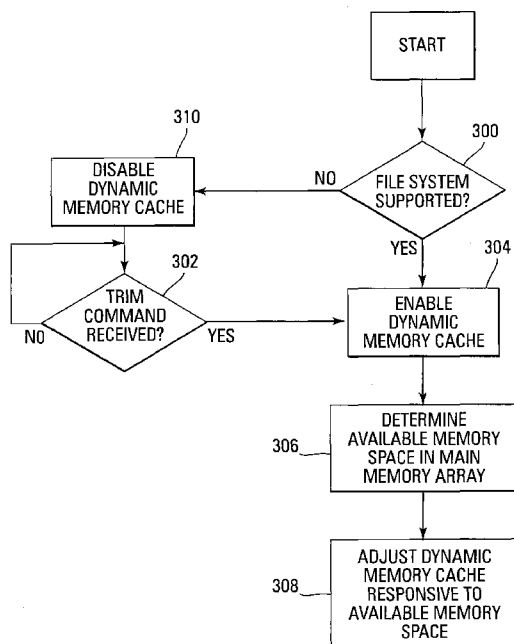


FIG. 3

(57) Abstract: Methods for dynamic memory cache size adjustment, enabling dynamic memory cache size adjustment, memory devices, and memory systems are disclosed. One such method for dynamic memory cache size adjustment determines available memory space in a memory array and adjusts a size of a memory cache in the memory array responsive to the available memory space.



TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

**Published:**

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

## DYNAMIC MEMORY CACHE SIZE ADJUSTMENT IN A MEMORY DEVICE

## TECHNICAL FIELD

[0001] The present embodiments relate generally to memory and a particular embodiment relates to dynamic SLC cache in an MLC memory.

## BACKGROUND

[0002] Flash memory devices have developed into a popular source of non-volatile memory for a wide range of electronic applications. Flash memory devices typically use a one-transistor memory cell that allows for high memory densities, high reliability, and low power consumption. Common uses for flash memory include personal computers, flash drives, digital cameras, and cellular telephones. Program code and system data such as a basic input/output system (BIOS) are typically stored in flash memory devices for use in personal computer systems.

[0003] A typical flash memory device is a type of memory in which the array of memory cells is typically organized into memory blocks that can be erased on a block-by-block basis (and reprogrammed on a page-by-page basis). Changes in a threshold voltage of each of the memory cells, through erasing or programming of a charge storage structure (e.g., floating gate or charge trap) or other physical phenomena (e.g., phase change or polarization), determine the data value programmed into each cell. The data in a cell of this type is determined by the presence or absence of the charge in the charge storage structure.

[0004] A programming operation typically comprises a series of incrementally increasing programming pulses that are applied to a control gate of a memory cell being programmed in order to increase that particular memory cell's threshold voltage. Each memory cell can be programmed as single level cell (SLC) memory or multiple level cell (MLC) memory where the cell's threshold voltage ( $V_t$ ) is indicative of the data value programmed into that cell. For example, in an SLC memory, a  $V_t$  of 2.5V might indicate a programmed cell while a  $V_t$  of -0.5V might indicate an erased cell. An MLC memory uses multiple  $V_t$  ranges that each indicates a different state. Multilevel cells can take advantage of the analog nature of a traditional flash cell by assigning a bit pattern to a specific  $V_t$  range. This technology permits the storage of data values representing two or more bits per cell, depending on the quantity of  $V_t$  ranges assigned to the cell.

[0005] During programming of a block of memory cells, a fixed cache of memory cells is typically used to temporarily store data that is to be programmed into the block of memory cells. For example, in an MLC memory device, a fixed size SLC cache can be used to store data for programming into an MLC block of memory cells. This can improve memory reliability. The memory performance is also improved prior to the cache becoming near full, at that point part of the cache must be moved to MLC blocks to create more room in the cache, and the performance advantage of the cache will be diminished. The performance improvement is function of the size of the cache but has the drawback of reducing user capacity since a fixed portion of the memory is used as a SLC cache and cannot be used to store user data with the same efficiency as MLC.

[0006] For the reasons stated above and for other reasons that will become apparent to those skilled in the art upon reading and understanding the present specification, there is a need in the art for a more efficient way to temporarily store data during programming.

#### BRIEF DESCRIPTION OF THE DRAWINGS

[0007] Figure 1 shows a block diagram of one embodiment of a memory array in a memory device that incorporates a memory cache.

[0008] Figure 2 shows a schematic diagram of one embodiment of a portion of a memory array in accordance with the block diagram of Figure 1.

[0009] Figure 3 shows flow chart of one embodiment of a method for dynamic cache size adjustment.

[0010] Figure 4 shows a plot of one embodiment of one function for determining when to adjust the cache size.

[0011] Figure 5 shows a plot of another embodiment of a function for determining when to adjust the cache size.

[0012] Figure 6 shows a block diagram of one embodiment of a memory system.

#### DETAILED DESCRIPTION

[0013] In the following detailed description, reference is made to the accompanying drawings that form a part hereof and in which is shown, by way of illustration, specific

embodiments. In the drawings, like numerals describe substantially similar components throughout the several views. Other embodiments may be utilized and structural, logical, and electrical changes may be made without departing from the scope of the present disclosure. The following detailed description is, therefore, not to be taken in a limiting sense.

[0014] Figure 1 illustrates a block diagram of one embodiment of memory array 104 of a memory device that incorporated dynamic memory cache 102. The memory is 104 is partitioned and includes a first partition; herein after referred to as main MLC memory wherein cells in that partition are programmed in MLC mode and a second partition 102 herein after referred to as dynamic SLC memory cache wherein cell in that partition are programmed in SLC mode. In one embodiment the blocks allocated to main memory in the memory array 100 are used as MLC and the blocks allocated to dynamic memory cache 102 are used as SLC. In a dynamic memory cache the allocation of blocks changes dynamically and is not fixed, a block may remain MLC or SLC or switch between being used as MLC or SLC. Without loss of generality, the blocks used for the cache 102 are referred to as SLC cache blocks and the blocks used for storing user data are referred to as MLC main memory blocks.

[0015] Typical prior art memory caches are fixed in size and are always enabled so that a portion of the memory array is always dedicated to a temporary data cache, reducing the amount of memory available for storing user data. The dynamic data cache 102 of Figure 1 is adjusted in size dynamically, depending on the free space available, and may not always be enabled. Thus, the dynamic memory cache 102 can be adjusted such that it does not take up more memory than is necessary to accomplish the cache function wherein the cache size is dynamically adjusted. Another feature of present invention is that partition of memory to MLC and SLC can be at array or block level. Yet another feature of present invention is that the arrays or blocks in a partition are not required to be continuous. In some systems the data is always first written to cache. In other systems only some type of data is written to cache, for example data less than a page size. In systems with fixed or dynamic memory cache some events will trigger moving the valid data from the cache to the main memory blocks (in some systems both valid and invalid data in the cache may be moved). Such events include the number of free blocks in cache falling below a threshold. When a block in cache is moved to main memory the block is erased and it is reclaimed. Similarly, blocks in main memory

containing old and new data can be reclaimed by moving the new data to another block in main memory and then erasing the old block.

[0016] The dynamic SLC memory cache 102 uses a variable number of blocks of memory of the memory array 104 to temporarily store data that is to be programmed into the main memory array 100. For example, the dynamic SLC memory cache 102 can store all pages (lower and upper) of data until the upper page of data is successfully programmed in the main MLC memory array 100. This can reduce corruption of a previously programmed lower page of the main MLC memory array 100 if a power failure occurs during upper page programming. In one embodiment, data is stored in the dynamic memory cache 102 blocks in SLC and when all pages required for programming a page in an MLC block are available (in dynamic memory cache 102) then the data is moved to a block in main memory array 100, so that the cache 102 can store all pages of data required until one page of the MLC main memory array 100 is programmed. For example, the SLC dynamic memory cache 102 can store a lower page of data until an upper page of the MLC main memory array 100 is successfully programmed. This can reduce corruption of a previously programmed lower page of the MLC main memory array 100 if a power failure occurs during upper page programming.

[0017] Figure 2 illustrates a schematic diagram of one embodiment of a portion of the NAND architecture memory array 201, as illustrated in Figure 1, comprising series strings of non-volatile memory cells. The present embodiments of the memory array are not limited to the illustrated NAND architecture. Alternate embodiments can use NOR or other architectures as well.

[0018] The memory array 201 comprises an array of non-volatile memory cells (e.g., floating gate) arranged in columns such as series strings 204, 205. Each of the cells is coupled drain to source in each series string 204, 205. An access line (e.g. word line) WL0 – WL31 that spans across multiple series strings 204, 205 is coupled to the control gates of each memory cell in a row in order to bias the control gates of the memory cells in the row. Data lines, such as even/odd bit lines BL\_E, BL\_O, are coupled to the series strings and eventually coupled to sense circuitry that detect the state of each cell by sensing current or voltage on a selected bit line.

[0019] Each series string 204, 205 of memory cells is coupled to a source line 206 by a source select gate 216, 217 (e.g., transistor) and to an individual bit line BL\_E, BL\_O by a drain select gate 212, 213 (e.g., transistor). The source select gates 216, 217 are controlled by

a source select gate control line SG(S) 218 coupled to their control gates. The drain select gates 212, 213 are controlled by a drain select gate control line SG(D) 214.

[0020] Figure 3 illustrates a flow chart of one embodiment of a method for dynamically adjusting the size of the dynamic memory cache illustrated in Figure 1. Since some file systems might not be recognized, the method determines if the file system implemented (e.g., installed, executed) on the memory device is one that is supported 300. As is well known in the art, a file system (e.g., File Allocation Table (FAT), New Technology File System (NTFS)) is a method for storing and organizing computer files and their data. It organizes these files into a database for the storage, organization, manipulation, and retrieval by a computer's operating system.

[0021] If the file system does not support dynamic cache size adjustment 300, the dynamic memory cache adjustment is disabled 310. Dynamic memory cache size adjustment might still be implemented if the memory device supports any command protocol that allows deletion of ranges of logical block addresses in order to implement the dynamic changing of the memory cache size. One such protocol known in the art is typically generically referred to as a TRIM protocol. It is thus determined if a TRIM command has been received 302. This step 302 is repeated until the TRIM command is received. Once the TRIM command is received 302, the dynamic memory cache is enabled 304.

[0022] Since the size of the dynamic memory cache is adjusted in response to the available memory space in the main memory array, the available memory space in the main memory array is determined 306. The amount of available memory space can change constantly. In one embodiment, the amount of available memory space in the main memory array can be determined after every write or erase operation. In other embodiments, the amount of available memory space can be determined periodically or at random times. Available memory space can include both erased memory that is not targeted for immediate use as well as memory that has not yet been erased but the data stored in the memory is old and no longer valid. In yet another embodiment, the amount of available memory is adjusted in response to a received TRIM command (if TRIM is supported) or, in case of known file systems, when clusters are deallocated or written.

[0023] In yet another embodiment, the amount of available memory is adjusted in response to the TRIM command (if TRIM is supported) or, in case of known file systems, when clusters are deallocated or written. In systems with fixed or dynamic memory cache, some

events will trigger moving the valid data (or, optionally, valid and invalid data) from the dynamic SLC cache memory to the main MLC memory blocks. Such events include the number of free blocks in cache falling below a threshold. When a block in dynamic SLC cache is moved to main MLC memory the block is erased and it is reclaimed and added to the pool of free blocks. Similarly, blocks in main MLC memory containing old and new data can be reclaimed by moving only the valid data to another block in main MLC memory and then erasing the old block.

[0024] The size of the dynamic memory cache is then adjusted in response to the available memory space 308. The size of the dynamic memory cache can be a percentage of the available memory space, all of the available memory space, or a certain number of blocks of the available memory space. In one embodiment, the percentage of the available memory space used can also be dynamically changed. For example, only 50% of the available memory space might be allocated to the dynamic memory cache at one time and, at a later time, 90% of the available memory space might be allocated to the dynamic memory cache. Such dynamic allocation of the percentage of available memory space can be performed as often as desired.

[0025] Figure 4 illustrates a plot of one embodiment of a function that can be used to determine the amount of available memory space to allocate to the dynamic memory cache. This function is a linear ramped function in which a fixed percentage (e.g., 50%) of the available memory space is allocated to the dynamic memory cache.

[0026] This plot includes the available memory space along the x-axis and the size of the dynamic memory cache along the y-axis. The memory size values on both the x and y-axes, in this and the following embodiments, are for purposes of illustration only as the present embodiments are not limited to any certain values. The slope of the line determines the percentage of available memory space that is allocated to the dynamic memory cache. The illustrated example shows a 50% embodiment.

[0027] Figure 5 illustrates a plot of another embodiment of a function that can be used to determine the amount of available memory space to allocate to the dynamic memory cache. This function is a staircase function that can be used to reduce the frequency of dynamic memory cache size adjustment.

[0028] Changing dynamic memory cache sizes uses a particular amount of time to perform various clean-up tasks. For example, when a memory block that has been used as part of the dynamic memory cache is to be returned to the main memory array, the data in that memory block needs to be moved and the memory block erased. In one embodiment, the SLC memory block of the dynamic memory cache is reallocated as an MLC memory block. Additionally, memory address pointers and other memory housekeeping tasks should also be performed to reallocate a dynamic memory cache block. Thus, it is typically desirable to reduce the frequency of dynamic memory cache size changes.

[0029] The staircase function is one way of reducing the frequency of dynamic memory cache size changes. This can be seen in reference to the plot of Figure 5. Unlike the embodiment of Figure 4 where the size change is performed in response to the slope of the line, the change of cache size does not occur in the stair step function until the present step reaches another particular threshold of available memory space.

[0030] For example, the initial dynamic memory cache size of 1 MB does not change until the available memory space reaches the 2 MB threshold. Then the dynamic memory cache size is increased to 2 MB. The illustrated available memory space and dynamic memory cache sizes are for purposes of illustration only as different amounts of available memory space can trigger reallocating different amounts of memory to the dynamic memory cache. Similarly, alternate embodiments of the step function illustrated in Figure 5 can remain at the same dynamic memory cache size longer and/or allocated greater amounts of memory to the dynamic memory cache each time a threshold is reached.

[0031] In yet another embodiment of a function for determining the amount of available memory space to allocate to the dynamic memory cache, the cache size can be a function of if the amount of unused available memory is increasing or decreasing over a particular time period. For example, instead of a fixed percentage as illustrated in Figure 4, if the available memory space size is increasing over a particular time period, the function might allocate a greater percentage of memory to the dynamic memory cache size than if the available memory space size was decreasing over the particular time period. In another embodiment, if the available memory space size is decreasing over a particular time period, the function might allocate a smaller percentage of memory to the dynamic memory cache size.

[0032] In yet another embodiment, to minimize overhead and dynamically allocating most of free space to cache a common pool of free (erased) blocks is used for both dynamic SLC

cache and main MLC memory. As free blocks are required (that is the number of free block fall below a first “start” threshold) dynamic SLC cache blocks are reclaimed and added to free pool until the number of free blocks is equal or more than a second “stop” threshold. The start and stop threshold may be adjusted dynamically. For example during foreground that is execution of command from host where we want to minimize reclaiming dynamic SLC cache blocks the stop threshold will not be high, but in background where there is no host commands the stop threshold will be set higher.

**[0033]** In one embodiment, the memory blocks that are allocated as dynamic memory cache blocks are contiguous memory blocks. In another embodiment, the dynamic memory cache blocks are not contiguous. In such an embodiment, a bit map can be used to indicate and track which memory blocks are used as dynamic memory cache blocks. In another embodiment, the bit indicating SLC or MLC may be combined with other information in the table for Logical to Physical mapping. In another embodiment, a list can be maintained for both SLC and MLC blocks. Such a list is typically a linked list for ease of implementation. In another embodiment the bit indicating SLC or MLC may be combined with other information in the table for Logical to Physical mapping. In another embodiment, a list is additionally maintained for both SLC and MLC blocks. Such a list is typically a linked list for ease of implementation.

**[0034]** Figure 6 illustrates a functional block diagram of a memory device 600. The memory device 600 is coupled to an external controller 610. The controller 610 may be a microprocessor or some other type of controller. The memory device 600 and the controller 610 form part of a memory system 620. The controller 610 can be coupled to a host and the controller 610 can be responsive to commands from the host.

**[0035]** The memory device 600 includes an array 630 of memory cells (e.g., non-volatile memory cells). The memory array 630 is arranged in banks of word line rows and bit line columns. In one embodiment, the columns of the memory array 630 comprise series strings of memory cells.

**[0036]** Address buffer circuitry 640 is provided to latch address signals provided through I/O circuitry 660. Address signals are received and decoded by a row decoder 644 and a column decoder 646 to access the memory array 630.

[0037] The memory device 600 reads data in the memory array 630 by sensing voltage or current changes in the memory array columns using sense amplifier circuitry 650. The sense amplifier circuitry 650, in one embodiment, is coupled to read and latch a row of data from the memory array 630. Data input and output buffer circuitry 660 is included for bidirectional data communication as well as the address communication over a plurality of data connections 662 with the controller 610. Write circuitry 655 is provided to write data to the memory array.

[0038] Memory control circuitry 670 decodes signals provided on control connections 672 from the controller 610. These signals are used to control the operations on the memory array 630, including data read, data write (program), and erase operations. The memory control circuitry 670 may be a state machine, a sequencer, or some other type of controller to generate the memory control signals. In one embodiment, the memory control circuitry 670 and/or the external controller 610 are configured to control execution of the dynamic memory cache size adjustment.

[0039] The memory device illustrated in Figure 6 has been simplified to facilitate a basic understanding of the features of the memory. A more detailed understanding of internal circuitry and functions of flash memories are known to those skilled in the art.

#### Conclusion

[0040] In summary, one or more embodiments of the method for dynamic memory cache size adjustment can provide increased capacity, performance (read or write) and/or reliability for user data in a memory device by dynamically adjusting the amount of memory allocated to a memory cache (e.g., used as SLC).

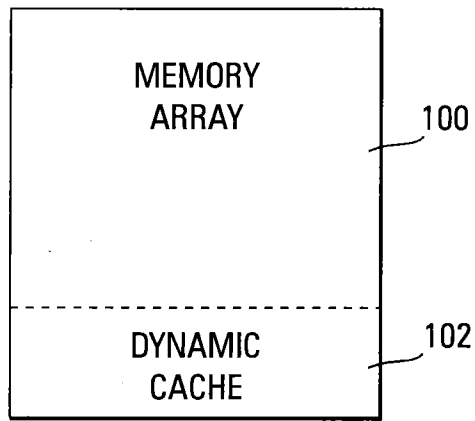
[0041] Although specific embodiments have been illustrated and described herein, it will be appreciated by those of ordinary skill in the art that any arrangement that is calculated to achieve the same purpose may be substituted for the specific embodiments shown. Many adaptations of the invention will be apparent to those of ordinary skill in the art. Accordingly, this application is intended to cover any adaptations or variations of the invention.

What is claimed is:

1. A method for dynamic memory cache size adjustment, the method comprising:  
determining available memory space in a memory array; and  
adjusting a size of a memory cache in the memory array responsive to the available  
memory space.
2. The method of claim 1 and further including:  
determining if a file system of a memory device that includes the memory array is  
supported; and  
enabling dynamic memory cache size adjustment if the particular file system is  
supported.
3. The method of claim 1 wherein the available memory space is one of erased memory  
that is not targeted for use or unerased memory that stores data that is not valid.
4. The method of claim 1 wherein the size of the memory cache is a percentage of the  
available memory space.
5. The method of claim 4 and further including dynamically adjusting the percentage.
6. The method of claim 4 wherein the percentage of the available memory space is a  
variable percentage of the available memory space.
7. The method of claim 1 wherein adjusting a size comprises allocating a first number of  
blocks of the memory array for use as the memory cache and allocating a second  
number of blocks of the memory array for use as main memory.
8. The method of claim 7 wherein allocating a first number of blocks includes allocating  
a block previously allocated for use as part of the main memory for use as part of the  
memory cache.

9. The method of claim 7 wherein memory cells in a block allocated for use as part of the main memory are programmed as MLC cells, and wherein memory cells in a block allocated for use as part of the memory cache are programmed as SLC cells.
10. The method of claim 7 wherein allocating a block previously allocated for use as part of the main memory for use as part of the memory cache comprises switching between programming memory cells of the block as MLC cells to programming the memory cells of the block as SLC cells.
11. The method of claim 2 wherein determining if a file system of the memory device is supported comprises determining if the file system of the memory device supports dynamic cache size adjustment and enabling dynamic memory cache size adjustment.
12. The method of claim 2 wherein determining if a file system of the memory device is supported comprises determining if a command is received that supports a protocol that allows deletion of a range of logical block addresses.
13. The method of claim 2 wherein dynamic memory cache size adjustment is disabled if either the particular file system does not support dynamic memory cache size adjustment or a command is not received that supports deletion of a range of logical block addresses of the memory device.
14. A memory device comprising:  
an array of memory cells comprising a memory cache; and  
memory control circuitry coupled to the array of memory cells and configured to  
determine available memory in the array of memory cells and adjust a size of  
the memory cache responsive to the available memory.
15. The memory device of claim 14 wherein adjusting the size of the memory cache comprises adjusting a number of blocks of the array whose memory cells are programmed as SLC cells.

16. The memory device of claim 14 wherein the memory control circuitry is further configured to enable or disable dynamic memory cache size adjustments responsive to a file system used.
17. The memory device of claim 14 wherein the memory control circuitry is configured to increase the size of the memory cache responsive to available memory reaching one of a plurality of available memory thresholds.
18. The memory system of claim 17 wherein the plurality of available memory thresholds forms a step function of available memory.
19. The memory device of claim 14 wherein the memory control circuitry is configured to adjust the size of the memory cache dynamically as a function of whether the available memory is increasing or decreasing over a particular time period.



**FIG. 1**

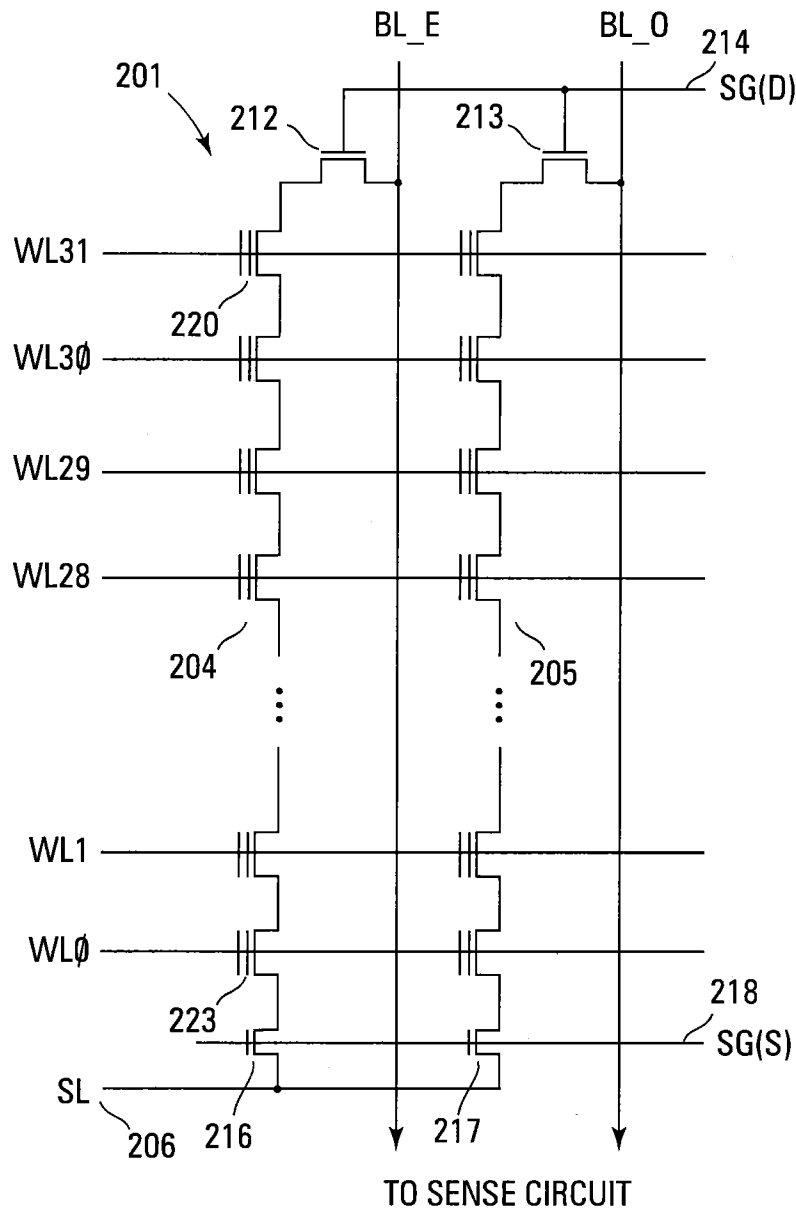


FIG. 2

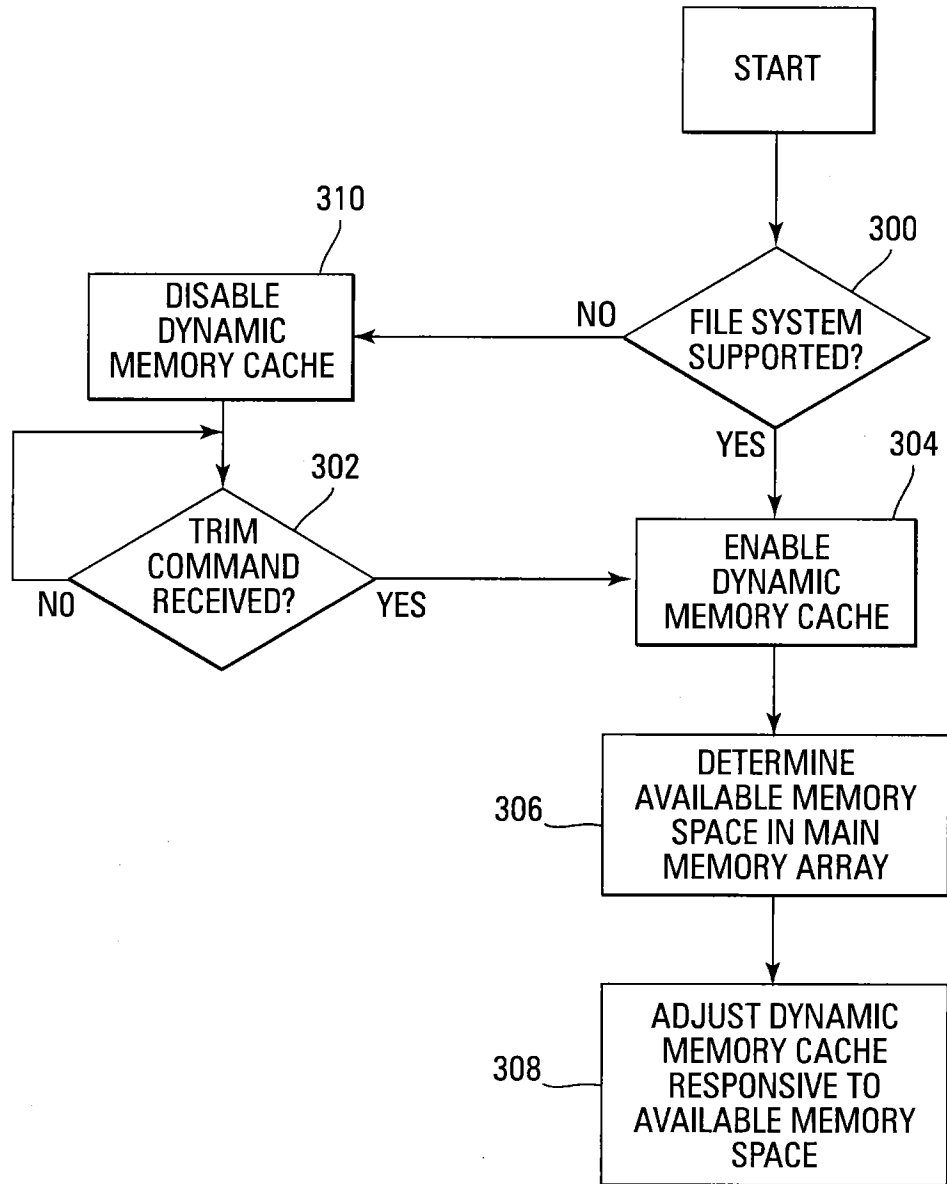


FIG. 3

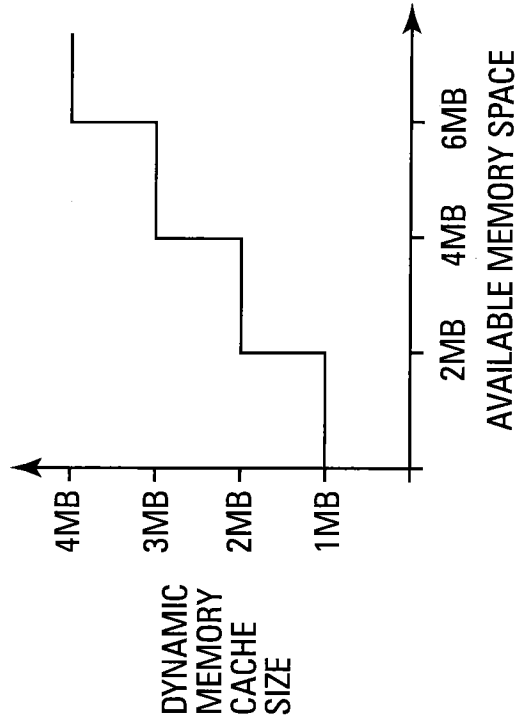


FIG. 5

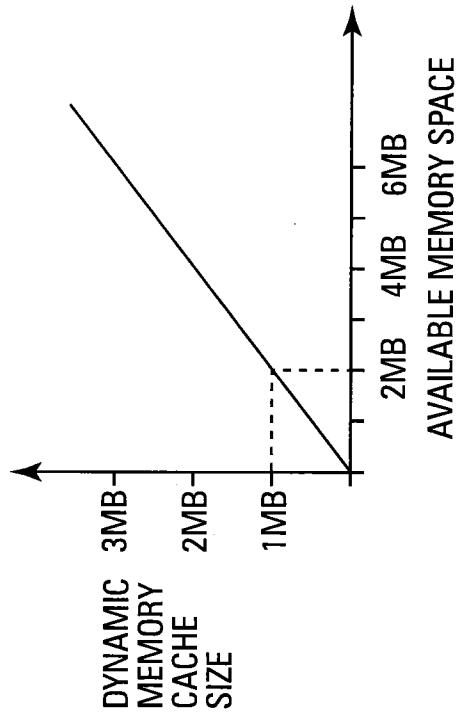


FIG. 4

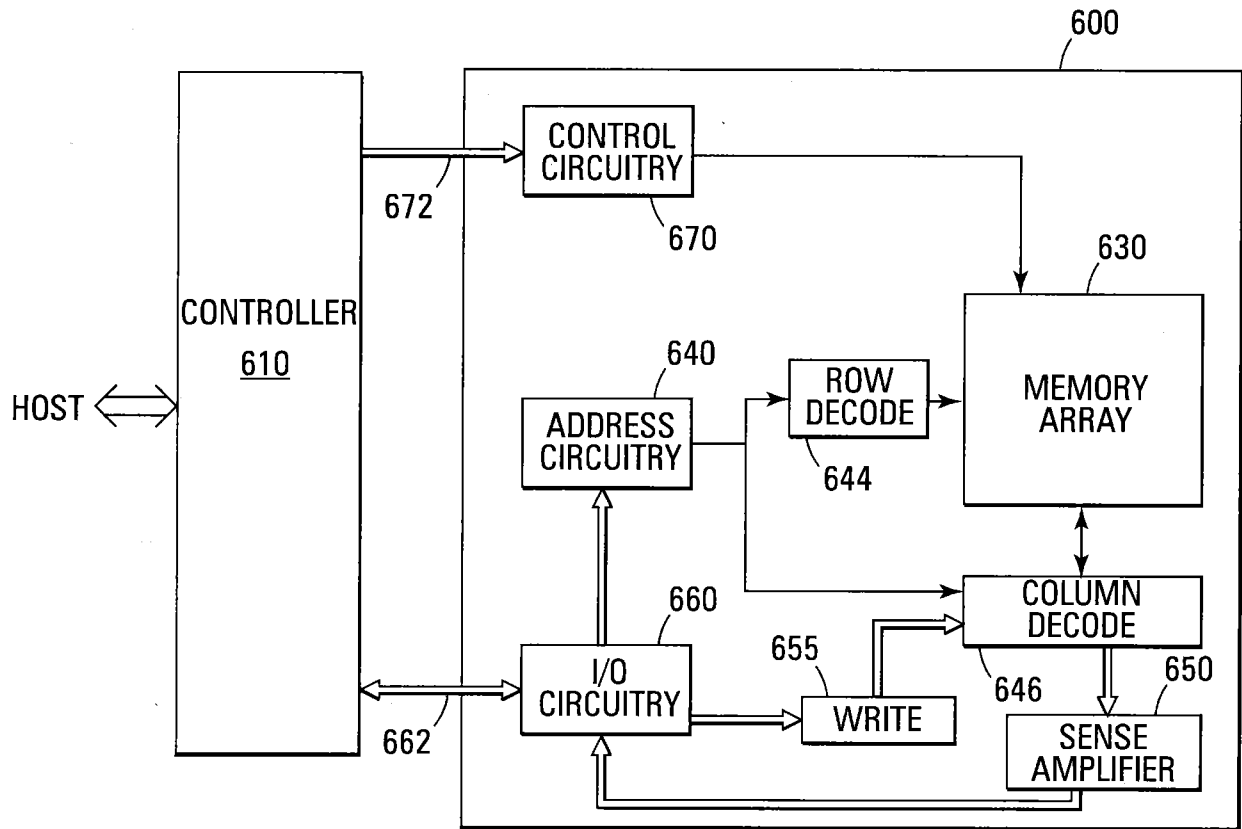


FIG. 6