

RTMP

YO994-007

RAID LEVEL 5 WITH FREE BLOCKS PARITY CACHE

Background of the InventionField of the Invention

5 This invention relates generally to permanent storage systems for digital information, especially those of the disk type, and more particularly to disk array systems which create and store parity blocks in order to facilitate recovery from a disk failure.

Related Art

10 A Redundant Array of Inexpensive Disks (RAID) has been proposed as a low cost alternative to a Single Large Expensive Disk (SLED) for providing large storage of digital information with high throughput. The theory of RAID is to use relatively inexpensive disks, which may individually have a higher chance of failure than more expensive disks, and compensating for this higher failure rate
15 by adding redundancy by creating and storing parity blocks to facilitate recovery from a disk failure.

FIG. 1A shows a disk array subsystem architecture on which a RAID organization can be implemented. A disk controller 30 connected to a host system 10, and having a cache 31 manages an
20 array of inexpensive disks 40-43. In a RAID organization with a total of $N+1$ disks, one parity block is created for each N data blocks, and each of these $N+1$ blocks (N data blocks plus one parity block) is stored on a different disk. In one implementation, a parity block is computed from the N data blocks by computing a
25 bitwise "Exclusive Or" (XOR) of the N data blocks. The parity block along with the N data blocks from which that parity block was



THIS SHEET HAS BEEN
REPLACED

THIS SHEET HAS BEEN
REPLACED

Y0994-007

computed are called a parity group. Any block in a parity group can be computed from the other blocks of that parity group.

In "A Case for Redundant Arrays of Inexpensive Disks (RAID)", *Proc. of ACM SIGMOD International Conference on Management of Data*, pp. 109-116, 1988, incorporated herein by reference, D. A. Patterson, G. Gibson and R. H. Katz describe five types of disk arrays classified as RAID levels 1 through 5. Of particular interest are disk arrays with an organization of RAID level 5, because the parity blocks in such a RAID type are distributed evenly across all disks, and therefore cause no bottleneck problems.

One shortcoming of the RAID environment is that a disk write operation is far more expensive than on a SLED, because a data write on RAID requires as many as four disk access operations as compared with two disk access operations on a SLED. Whenever the disk controller in a RAID organization receives a request to write a data block, it must not only update (i.e., read and write) the data block, but it also must update (i.e., read and write) the corresponding parity block to maintain consistency. For instance, if data block *D1* in FIG. 2A is to be written, the new value of *P0* is calculated as:

$$\text{new } P0 = (\text{old } D1 \text{ XOR new } D1 \text{ XOR old } P0)$$

Therefore, the following four disk access operations are required:

- (1) read the old data block *D1*;
 - (2) read the old parity block *P0*;
 - (3) write the new data block *D1*; and
 - (4) write the new parity block *P0*.
- The reads must be completed before the writes can be started.

In "Performance of Disk Arrays in Transaction Processing Environments", *Proc. of International Conference on Distributed Computing Systems*, pp. 302-309, 1992, J. Menon and D. Mattson teach



YO994-007

that caching or buffering storage blocks at the disk controller can improve the performance of a RAID disk array subsystem. If there is a disk cache, the pre-reading from the disk array of a block to be replaced can be avoided if the block is in the cache. Furthermore, if the parity block for each parity group is also stored in the cache, then both reads from the disk array can be avoided if the parity block is in the cache.

Commonly assigned and co-pending United States Patent Application 07/017,920, filed February 16, 1993 (IBM Docket YO993-013), describes a system wherein parity blocks from high write activity parity groups are stored in a cache buffer in order to reduce the number of disk accesses during updating.

A drawback of prior schemes is that at any given time, various data blocks within a given parity group may not contain useful data. For example, some of the data blocks may not have been used, while others, although previously used, may have been freed by the operating system. Thus, while each parity block is assigned to a group of data blocks, each parity block bit depends on all corresponding data block bits, regardless of whether the data block is in use.

Summary of the Invention

The present invention eliminates the drawbacks discussed above. The present invention is a system for writing data to a disk array wherein a parity block can provide parity information either for all data blocks within its parity group or for any arbitrary subset of these data blocks. This allows parity information to be stored only for data blocks that are currently in use. The system includes a cache memory coupled to the disk array for storing parity blocks associated with parity groups comprising



Y0994-007

5 respective locations. In accordance with the present invention,
each of the parity blocks includes an identifier indicative of data
blocks within a particular parity group which are protected by the
parity data. Write logic, coupled to the cache memory and the disk
10 array, includes means for reading the identifier and determining
whether a destination block is not protected by the parity data.
The write logic also includes write means for writing to the
unprotected location and update means for updating the parity data
and the identifier associated with the parity block including the
10 location, to indicate that the location is protected.

Brief Description of the Drawings

The present invention will be better understood by reference
to the drawings, in which:

15 FIG. 1A is a block diagram illustrating the architecture of a disk
array subsystem;

FIG. 1B is a block diagram illustrating the architecture of a disk
array subsystem in which the present invention is implemented;

FIG. 2A is a conventional RAID level 5 data mapping showing the
placement of data and parity blocks;

20 FIG. 2 is a partitioned parity/data disk cache;

FIG. 3 is a flow chart of a single block write according to an
embodiment of the present invention;

FIG. 4 is a flow chart of a free blocks command according to an
embodiment of the present invention;



YO994-007

FIG. 5 is a block diagram describing the parity and identifier blocks of the present invention.

Detailed Description of the Preferred Embodiment

5 The present invention may be practiced on the disk array subsystem shown in FIG. 1B, which includes a disk controller 30 having a cache memory 31 and an array of disks 40, 41, 42, and 43. The cache memory 31 is divided into two partitions: a data partition 31a for buffering data blocks and a parity partition 31b for buffering parity blocks. The data partition 31a and the parity partition 31b are independently accessible by the cache controller. In accordance with an embodiment of the present invention, data block writes are performed under control of variable write protection logic 32, which is connected to both partitions 31a and 31b of the cache 31. The data mapping in the disk array 40-43 of the preferred embodiment to be described is a RAID level 5 (RAID V) mapping.

FIG. 2 represents the data structures used by the disk controller 30 for its caches 31a and 31b. When block 202 is given a stripe's ID, RAID V determines which disk has the parity and the numbering of data blocks on the other disks. The disk controller then works with real disk addresses. Cache directories 204 and 206 are implemented as hash tables, one for the parity cache and one for the data cache. Given a disk block address, the hash table is consulted to determine whether or not the data having the address is in the cache, and if it is in the cache, where it is. A hash table entry for a block in the parity cache points to a least recently used (LRU) stack 208. The address of the LRU stack entry determines where the actual data is, since there is a one-to-one correspondence between LRU stack entries and cache blocks. To compute the address of data, use (cache space base address) +



YO994-007

(block size) * (LRU stack entry address - LRU stack base address) / (LRU stack entry length). LRU stack 210 entries have a data block address field.

5 In operation, each cache is used in the following way: (1) each time a parity/data block is accessed, it is moved from its current position to the most recently used (MRU) position in the LRU stack; (2) when space is needed to read-in a parity/data block, the cache space for the LRU block is freed, the block is read into this location, placed at the MRU position, and the cache directory
10 is updated accordingly.

In a preferred embodiment, an LRU stack is implemented as a doubly-linked list, or chain, of fixed size entries. Each entry may have information about the object for which it is a LRU stack. In this case, a pointer to the cache block is unnecessary if
15 address computation techniques are used. Each entry may also have information about two pointers, one to the next MRU entry (or NIL if it is the MRU entry) and one to the next LRU entry (or NIL if it is the LRU entry). The LRU stack can be implemented in software, since entries can be moved anywhere out of the middle of the stack
20 and then put on the top of the stack with a few instructions.

FIG. 2A shows a RAID level 5 data mapping. In a RAID level 5 system with a total of $N+1$ disks, each parity group includes N data blocks and a parity block. The traditional approach is to divide data blocks into parity groups, such as that shown in FIG. 2A. In
25 FIG. 2A, for example, data blocks $D0$, $D1$, $D2$ and $D3$ and parity block $P0$ form a parity group. In FIG. 2A parity blocks are labelled 100, 101, 102, 103, and 104. Parity block $P0$ (labelled 100), therefore, is the XOR of data blocks $D0$ (labelled 110), $D1$ (labelled 111), $D2$ (labelled 112), and $D3$ (labelled 113). In
30 accordance with the present invention, for each data block update



Y0994-007

request from the host system 10 to the disk controller 30, if the corresponding parity block is already in the control unit cache 31, three physical disk access operations are required for a data update: 1) read the old data; and 2) write the new data; 3) write the new parity block.

In accordance with an embodiment of the present invention, each of the parity blocks includes an identifier, or "prefix", which indicates which of the data blocks in the parity group are protected. When the host system 10 sends a write command to the disk controller 30, variable write protection logic 32 reads the prefix from the relevant parity block in the cache memory 31. In instances where the parity block prefix associated with a parity group (or "stripe") to be written-to indicates that a particular data block is not protected by the parity, the old parity is XOR'ed with the new data block, the prefix is updated to indicate that the data block is protected, the new data is written to the data block on disk, and the new parity block is written to disk. Thus, the step of reading the old data from the disk is avoided completely.

A flow chart of the variable protection single block write logic 32 in accordance with the invention is illustrated in FIG. 3. This variable protection write logic 32 can be embodied as a microcode control program executing on the disk controller's microprocessor.

The logic starts when the disk controller receives a write command from the host processor 10. The write command includes the data block, and the write address. Also included in the write command is a flag, set by the host file system, which indicates whether the entire stripe consists of free (i.e., unused) space.



YO994-007

In step 302 the disk controller examines the free stripe flag to determine whether the entire stripe to which the data block belongs is free. If the entire stripe is free, in step 304 the disk controller sets the stripe's parity block P equal to the data block D ($P=D$). Also, data block D is marked as protected in the free blocks bit vector. It is noted that this operation requires no reads from the cache or disk. In step 306 the controller writes P and D to the disk.

If, in step 302, the disk controller determines that the entire stripe is not free, in step 308 the controller determines if the parity block P is in the cache 31 by means of hash table 304. If not, in step 310 the controller determines if the data block being replaced (D_{old}) is in the cache 31. If yes, in step 312 the controller reads P from the disk. If not, in step 314 the controller reads both P and D_{old} from the disk.

Next, in step 316 the controller examines the parity block prefix to determine whether D_{old} is protected. If yes, in step 318 the controller determines the new value for the parity P as $P = P \text{ XOR } D_{old} \text{ XOR } D$. Then, in step 306, P and D are written to the disk. If D_{old} is not protected, in step 320 the new value for P is determined as $P = P \text{ XOR } D$ and D is marked as protected in the block prefix. Then, in step 306, P and D are written to the disk.

If in step 308 it is determined that P is in the cache, in step 324 the controller examines the parity block prefix to determine whether D_{old} is protected. If yes, in step 326 the disk controller determines whether D_{old} is in the cache. If yes, steps 318, and 306 are performed. If not, D_{old} is read from the disk in step 328 and then steps 318 and 306 are performed.



YO994-007

If from step 324 it is determined that D_{old} is not protected, in step 320 the controller determines the new parity value as $P = P \text{ XOR } D$ and then marks D as protected. Then step 306 is performed. It is noted that in step 320 all values are taken from the cache and thus no reads are required from the disk.

In a preferred embodiment, disk space is managed by a component of the host operating system known as the file system. The file system is responsible for mapping logical entities known as files and directories to physical space on disk, and for managing the free (unused) disk space. For example, when it is necessary to create a new file, or to extend an existing file, the file system may allocate existing free space for this purpose, so that the free space becomes non-free (i.e., in use). Conversely, when files are deleted, or when the space necessary to store a file decreases, previously in-use space may become free. In order to optimize the performance of disk writes to free space, and in order to speed recovery after a disk failure, it is desirable that free disk blocks be unprotected. For this purpose, a "Free Blocks Command" is provided to the disk controller, the logic of which is shown in FIG. 4. This command can be issued by the file system, for example, either whenever previously in-use space is made free in any given stripe S , or else in batches (a batch consisting of one command for each stripe containing protected free space which it is desired to make unprotected), periodically, as determined by a given time interval, or by a threshold amount of protected free space having accumulated, etc.

The "Free Blocks Command" of FIG. 4 operates as follows: First, in step 402 the cache maintenance program determines if P is in the cache. If not, in step 404 P is read from disk into the cache. Next, in step 406, the program determines whether the number of blocks to be freed (k) is equal to the number of blocks



Y0994-007

in the stripe (N) (not counting P). If the entire stripe (except for P) is to be freed, in step 408 all of the data blocks are marked unprotected. In step 410, each of the blocks B_j to be freed which are stored in cache is marked as free cache space. Then in
5 step 412, the new parity block is written to the disk.

If less than the entire stripe of data blocks is to be freed, then in step 414 the number of in-use data blocks (d) which are not stored in cache is determined. d is the number of data blocks that have to be read from disk to compute parity by XORing all in-use
10 data blocks. In step 416, the number b of protected data blocks to be freed that are not in the cache is determined. b is the number of blocks that need to be read from the disk if parity is recomputed by subtracting out the protected blocks that are becoming unprotected.

Next, in step 415 it is determined whether d is less than or equal to b. The purpose of this step is to do the minimum number of I/O operations of the two alternatives. If d is less than or equal to b, in step 418 all of the in-use data blocks not in the
15 cache are read from disk and then in step 420 the parity is recomputed by XORing all of the data blocks D₁ ... D_n with one another. If, on the other hand, d is greater than b, in step 422 all of the protected data blocks not in the cache are read from disk and then in step 424 each data block B_j to be freed is XOR'ed with the parity block P. This operation subtracts the protected
20 data blocks B_j to be freed out of the parity block.
25

In step 426, all of the blocks B_j being freed are marked as unprotected by clearing their entries in the relevant block prefix. Then, steps 410 and 412 are performed as previously described.



Y0994-007

FIGS. 5A-5C show some examples of how parity is determined using a parity block prefix 502 in accordance with the present invention. An embodiment having four data blocks and one parity block per stripe (parity group) is shown. The prefix is mapped such that each data block Dn has a position in the prefix. A zero indicates that the respective data block is not included in the parity calculation, and a 1 indicates that the data block is included in the parity calculation.

In FIG. 5A, data blocks D9, D10 and D11 are free, thus only data block D8 is protected by the parity block P2. The prefix 502a is set to 1000 (binary) to indicate that the parity bits 504a of parity block P2 were generated only from block D8. The three zeros indicate that the remaining three data blocks are unprotected.

In FIG. 5B, it is assumed that the controller needs to perform a write to block D10. By examining the old parity block prefix 502a the controller determines that only block D8 is protected by the parity bits, and thus the new parity is calculated as $P = D8 \text{ XOR } D10$, or $P = D10 \text{ XOR old } P$, and is stored in parity block 504b.

In FIG. 5C, it is assumed that the controller needs to perform a write to block D9. The old parity 502b indicates that block D9 is not protected. Thus, the new parity is calculated as $P = D8 \text{ XOR } D9 \text{ XOR } D10$, or $P = D9 \text{ XOR old } P$. P is stored in parity block 504c, and prefix updated as shown in 502c.

With the RAID 5 with free blocks parity system of the present invention, if a disk in the disk array fails, the recovery scenario is the same as that of a standard RAID disk array. The missing block in each parity group having a block on the failed disk is derived from the other blocks of its parity group on other disks. An advantage during recovery, however, is that in the present



YO994-007

system the prefix provides a map of the data blocks described by the parity block. Reconstruction of unused data blocks, therefore, is unnecessary.

5 In the event of a single disk failure, the contents of the disk in prior RAID architecture disk systems can be reconstructed. Essentially, for each block of the failed disk, if it is a parity block, then the parity information is re-computed, and if it is a data block, then its contents can be reconstructed using all other data blocks in the stripe, together with the parity block for the
10 stripe.

Using the variable stripe protection scheme of the current invention, reconstruction takes place as follows. First, suppose the block on the failed disk is a data block. From the parity block for the stripe, it can be determined if the block is
15 protected or unprotected. If it is protected, then its contents are reconstructed as described above, except that only the other protected data blocks (together with the parity block) in the stripe are used. If only a few other blocks are protected, this results in I/O savings (as compared to prior art methods), and
20 hence faster disk recovery, with increasing savings for fewer protected blocks in each stripe. The other case is that the data block is unprotected. In this case, no further action is necessary, which clearly is a large improvement over prior art methods. In particular, note that if the failed disk contains
25 large amounts of unprotected (free) space, disk recovery will be extremely fast as compared to the prior disk reconstruction methods.

Next, suppose the block on the failed disk is a parity block. In this case there are two methods for reconstructing the block.
30 First, if it is desired to have the disk reconstruction proceed



YO994-007

independently of a host system, then it is necessary to make the worst-case assumption that all data blocks in the stripe are valid (in-use) blocks. In this case, the parity block is found by computing the parity of all data blocks in the stripe, and furthermore, all data blocks are marked as protected in the parity block prefix area.

A second method, which involves communication with a host system, is as follows. For each parity block on the failed disk, the disk controller queries a host file system for the status (free or in-use) of the data blocks in the stripe. The host file system returns a bit-vector indicating, for each block in the stripe, whether it is free. Each such free block is marked as unprotected, and the other data blocks in the stripe are marked as protected. Finally, parity is computed for the protected blocks in the stripe, thus reconstructing the parity block on the failed disk. If fewer than all of the data blocks are protected in a stripe, this also is an improvement over prior methods, since only the protected blocks need to be read from disk to recompute the parity block.

In summary, the present invention reduces the number of required reads and writes when fewer than all data blocks within a stripe are protected by the parity block. Specifically, the parity blocks include identifiers which classify each data block as being either: protected data, i.e., data for which parity information has been computed and stored in the parity block; or unprotected data, i.e., data for which there is no stored parity information. A data block in the latter category should always be a block that is not in use. Additionally, data blocks in the first (protected) category may be further classified as being in use or not in use. The case in which a protected data block contains data that is not in use may arise, for example, when information in the given block is updated and written to another disk location, or when the



YO994-007

information becomes obsolete and the disk space containing the data is freed by the operating system. Writes to unprotected data blocks are performed without first determining the current block contents, as these are not included in the current parity information. Further, data blocks that are protected, but not in use, can be converted to unprotected blocks by recomputing parity information for the in-use data blocks. In a preferred embodiment, the system utilizes a cache memory whose contents include a number of data and parity blocks, as well as write logic which reads and modifies prefixes in parity blocks so as to determine the actions required for reading and writing blocks of each type.

While the invention has been particularly shown and described with reference to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein departing from the spirit and scope of the invention.





YO994-007

CLAIMS

Having thus described our invention, what we claim as new, and desire to secure by Letters Patent is:

- 1 1. A system for writing data to a disk array, comprising:
 - 2 a cache memory coupled to the disk array, the cache memory
 - 3 storing one or more parity blocks associated with stripes of
 - 4 disk data blocks, each parity block including parity data
 - 5 for one or more data blocks in its stripe and an identifier
 - 6 identifying protected data blocks in its stripe; and
 - 7 write logic coupled to the cache memory and the disk array,
 - 8 the write logic including:
 - 9 means for determining whether or not a location is protected
 - 10 by the parity data;
 - 11 means for writing data to a selected location;
 - 12 means for computing new parity data and updating the
 - 13 identifier associated with the new parity data if necessary
 - 14 to indicate that the selected location is protected.
- 1 2. The system according to claim 1, further comprising:
 - 2 means for reading from the disk array a block of data after
 - 3 it has been freed; and
 - 4 means for recomputing parity data and updating the
 - 5 identifier associated with the recomputed parity data to
 - 6 reflect the freeing of the block of data.





Y0994-007

1 3. The system according to claim 2, including means for reading
2 from the disk array a block of data as it is freed.

3 4. The system according to claim 2, further comprising means
4 for determining an optimal set of blocks to read from disk so as
to recompute the parity data.

5. The system according to claim 2, further comprising means
for reading from the disk array a plurality of blocks of data
after they have been freed.

1 6. A method of writing new data to disk in a redundant array of
2 inexpensive disks, comprising:

3 storing the new data in a cache memory;

4 computing a parity block for in-use data blocks of a group
5 of data blocks in the disk array;

6 attaching to the parity block an identifier indicating which
7 data blocks of the group of data blocks are in use;

8 choosing a destination data block for the new data;

9 determining whether the destination block belongs to a group
10 of free data blocks, and if so, setting the parity block
11 using only the new data, and writing the set parity block
12 and the new data to disk; otherwise, determining from the
13 identifier whether the destination block is protected by the
14 parity block;

15 if so, using the new data, the destination block and
16 the parity block to compute a new parity block





YO994-007

- 17 reflecting the new data and writing the new parity
18 block to cache and disk;
- 19 if not, computing a new parity block reflecting the new
20 data, writing the new parity block to cache and disk
21 and updating the identifier to indicate that the
22 destination block is now protected by the parity block.
- 1 7. The method of claim 6, wherein if the destination block is
2 protected by the parity block and if the contents of the
3 destination block are not stored in cache, reading the
4 contents of the destination block from disk prior to
5 computing the new parity block.
- 1 8. The method of claim 7; wherein the parity block and each of
2 the blocks in the group of data blocks are stored on a
3 different disk in the disk array.
- 1 9. The method according to claim 6, wherein if the parity block
2 is not stored in cache, reading the parity block and the
3 identifier from disk prior to the step of determining
4 whether the destination block is protected.
- 1 10. The method of claim 6, wherein the new parity block is
2 computed by XORing the new data block and the parity block.
- 1 11. A method of accounting for unused data blocks in a redundant
2 array of inexpensive disks, comprising:
- 3 computing a parity block for data blocks in a group of data
4 blocks stored on the disks;





YO994-007

5 attaching to the parity block an identifier indicating which
6 blocks of the group are in use;

7 when a block is to be reclassified from in-use to unused,
8 computing a new parity block using the parity block and the
9 data in the block to be reclassified, and updating the
10 identifier to indicate that the block to be reclassified is
11 not protected by the new parity block.

1 12. The method according to claim 11, wherein the array has an
2 associated cache memory, and wherein if the parity block and
3 identifier are not stored in the cache memory, reading the
4 parity block and identifier from disk prior to the step of
5 computing a new parity block.

1 13. The method according to claim 11, wherein the array has an
2 associated cache memory, and if the data block to be
3 reclassified is not stored in the cache memory, reading the
4 data block to be reclassified from disk prior to the step of
5 computing a new parity block.

1 14. The method according to claim 11, wherein the identifier
2 comprises a number of bits equal to the number of data
3 blocks in the group, each bit corresponding to a different
4 one of the data blocks in the group, and wherein the step of
5 updating the identifier comprises changing the state of the
6 bit corresponding to the block to be reclassified.

1 15. The method according to claim 11, wherein the parity block
2 and each of the data blocks in the group are stored on a
3 different disk in the array.

IBM Confidential

THIS SHEET HAS BEEN
REPLACED

YO994-007

- 1 16. The method according to claim 11, wherein the step of
- 2 computing a new parity block comprises XORing the parity
- 3 block and the data in the block to be reclassified.

IBM Confidential

This Sheet Has Been Replaced



* C L M S *