



- (51) **International Patent Classification:**
H03M 13/15 (2006.01) *G06F 12/00* (2006.01)
- (21) **International Application Number:**
PCT/IB2013/001188
- (22) **International Filing Date:**
15 March 2013 (15.03.2013)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant:** EMC CORPORATION [US/US]; 176 South Street, Hopkinton, Massachusetts 01748 (US).
- (72) **Inventors:** ALIEV, Artem Alexandrovich; Botanicheskaya ul., d. 5, Petrodvorets, St. Petersburg, 198504 (RU). TRIFONOV, Peter Vladimirovich; Ul. Hlopina, d. 7, k. 4, kv. 9, St. Petersburg, 194021 (RU). MILOSLAVSKAYA, Vera Dmitriyevna; Pochtamtsky per., d. 2, kv. 9, St. Petersburg, 190000 (RU). ALEXEEV, Alexander Nickolaevich; Ul. Scherbakova, d. 21, kv. 13, St. Petersburg, 197375 (RU).
- (74) **Agent:** GORDON, Ronald, L., Ph.D.; Bainwood, Huang & Associates, LLC, 2 Connector Road, Westborough, MA 01581 (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, BN, 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, PA, 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, 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:
— with international search report (Art. 21(3))

(54) **Title:** LOAD BALANCING ON DISKS IN RAID BASED ON LINEAR BLOCK CODES

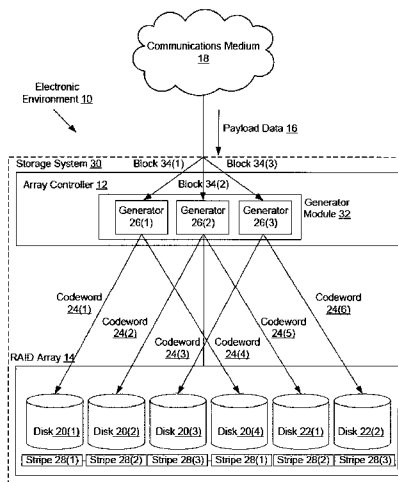


Figure 1

(57) **Abstract:** An improved technique involves assigning a different generator matrix to each data stripe of the redundant disk array such that all of the different generator matrices represent the same code. For example, when a $k \times n$ generator matrix G represents a linear code C , k being the block length and n the code length, then for any invertible $k \times k$ matrix P , the matrix $G' = PG$ is also a generator that represents C . When C is a systematic code, then G consists of a $k \times k$ identity matrix representing payload data concatenated with a $k \times (n-k)$ parity matrix representing parity data. Certain matrices P represent row operations on G , meaning that the matrix G' may have the columns of the identity matrix in G to different locations in G' .

LOAD BALANCING ON DISKS IN RAID BASED ON LINEAR BLOCK CODES

BACKGROUND

Redundant arrays of inexpensive disks (RAIDs) provide for reliable storage of payload data by storing additional data in some of the disks of the array. The design of a RAID involves ensuring that, if some number of disks fails, then the data stored in those disks may be recovered from other disks in the array.

Some RAIDs apply codes to segments of payload data to create codewords on the disks of an array. In this manner, in the event of a failure, a RAID controller is able to recover the payload data from those codewords. Some codes are maximum-distance separable (MDS) in that such codes allow for recovery of up to r erasures when r parity disks have been set aside for parity data. An example of such an MDS code is a Reed-Solomon code with an alphabet based on $GF(2^m)$, where m is the length of a payload block.

Other codes, however, are non-MDS codes based on alphabets such as $GF(2)$. In these codes, many mathematical operations on codewords may be carried out with XOR operations. In conventional RAID systems that use systematic, non-MDS codes, a storage processor stores parity data and payload data on separate disks and perform partial stripe updates of parity data independently from the payload data.

SUMMARY

Unfortunately, there are deficiencies with the above-described conventional RAID systems that use non-MDS codes. In most storage systems parity data are updated far more frequently than information symbols, causing parity disks to wear more quickly than payload disks. If one could afford the computational overhead of an MDS code, such as used in classical RAID-5 or RAID-6, then cycling mapping of codeword symbols onto disks can be used, which provides disk load balancing. Cyclic mapping, however, is not an option for non-MDS codes because there are some configurations of r erasures which are not recoverable with such codes.

In contrast to the conventional RAID system that use non-MDS codes in which unbalanced loads may cause premature wear on parity disks, an improved technique involves assigning a different generator matrix to each data stripe of the redundant disk array such that all of the different generator matrices represent the same code. For example, when a $k \times n$ generator matrix G

represents a linear code $\mathcal{C}$, k being the block length and n the code length, then for any invertible $k \times k$ matrix P , the matrix $G' = PG$ is also a generator that represents $\mathcal{C}$. When $\mathcal{C}$ is a systematic code, then G consists of a $k \times k$ identity matrix representing payload data concatenated with a $k \times (n - k)$ parity matrix representing parity data. Matrices P represent row operations on G .

5 There exist such matrices P , so that matrix G' may have the columns of the identity matrix in different locations than G .

Advantageously, the improved technique allows for load balancing in non-MDS codes when different generator matrices representing that non-MDS code are used for different data stripes. By ensuring that the columns of the parity matrix are distributed evenly across the stripes, then the parity updates affect the stripes, and therefore the disks, about the same. Further, one may design an optimization scheme that minimizes the load imbalance between the stripes for a given systematic code $\mathcal{C}$.

One embodiment of the improved technique is directed to a method of storing data in a redundant disk array in a storage system that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n . The method includes constructing a set of ℓ generator matrices, each generator matrix of the set of ℓ generator matrices being compatible with the code and corresponding to a data stripe of the redundant disk array, ℓ being at least 2. The method also includes generating a respective codeword for a given information vector by multiplying the given information vector by a particular generator matrix of the set of ℓ generator matrices. The method further includes storing the respective codeword in a data stripe of the redundant array to which the particular generator matrix corresponds.

Additionally, some embodiments of the improved technique are directed to a system constructed and arranged to store data in a redundant disk array in a storage system that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n . The system includes a network interface, memory, and a controller including controlling circuitry constructed and arranged to carry out the method of storing data in a redundant disk array in a storage system that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n .

Furthermore, some embodiments of the improved technique are directed to a computer program product having a non-transitory computer readable storage medium which stores code including a set of instructions which, when executed by a computer, cause the computer to carry out

the method of storing data in a redundant disk array in a storage system that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n .

5 BRIEF DESCRIPTION OF THE DRAWING

The foregoing and other objects, features and advantages will be apparent from the following description of particular embodiments of the invention, as illustrated in the accompanying figures in which like reference characters refer to the same parts throughout the different views.

Figure 1 is a block diagram illustrating an example electronic environment for carrying out
10 the improved technique.

Figure 2 is a block diagram illustrating an example array controller within the electronic environment shown in Figure 1.

Figure 3 is a flow chart illustrating an example optimization process within the electronic environment shown in Figure 1.

15 Figure 4 is a flow chart illustrating further details of the example optimization process shown in Figure 3.

Figure 5 is a flow chart illustrating an example method of carrying out the improved technique within the electronic environment shown in Figure 1.

20 DETAILED DESCRIPTION

An improved technique involves assigning a different generator matrix to each data stripe of the redundant disk array such that all of the different generator matrices represent the same code. For example, when a $k \times n$ generator matrix G represents a linear code $\mathcal{C}$, k being the block length and n the code length, then for any invertible $k \times k$ matrix P , the matrix $G' = PG$ is also a
25 generator that represents $\mathcal{C}$. When $\mathcal{C}$ is a systematic code, then G consists of a $k \times k$ identity matrix representing payload data concatenated with a $k \times (n - k)$ parity matrix representing parity data. Matrices P represent row operations on G , and there exist matrices P such that G' has the columns of the identity matrix in locations other than that in G .

Advantageously, the improved technique allows for load balancing in non-MDS codes when
30 different generator matrices representing that non-MDS code are used for different data stripes. By ensuring that the columns of the parity matrix are distributed evenly across the stripes, then the

parity updates affect the stripes, and therefore the disks, about the same. Further, one may design an optimization scheme that minimizes the load imbalance between the stripes for a given systematic code $\mathcal{C}$.

Figure 1 illustrates an example electronic environment 10 for carrying out the improved technique. Electronic environment 10 includes storage system 30 which in turn includes array controller 12 and RAID array 14, and communications medium 18.

Communication medium 18 provides network connections between array controller 12 and storage system 30. Communications medium 18 may implement a variety of protocols such as TCP/IP, UDP, ATM, Ethernet, Fibre Channel, SCSI combinations thereof, and the like. Furthermore, communications medium 18 may include various components (e.g., cables, switches/routers, gateways/bridges, NAS/SAN appliances/nodes, interfaces, etc.). Moreover, the communications medium 18 is capable of having a variety of topologies (e.g., queue manager-and-spoke, ring, backbone, multi drop, point to-point, irregular, combinations thereof, and so on).

Array controller 12 is constructed and arranged to convert blocks of payload data 16 into various codewords 24(1), 24(2),...,24(6) (codewords 24) according to generator matrices 26(1), 26(2), and 26(3) (generator matrices 26) in generator module 32. Array controller 12 is also constructed and arranged to send codewords to different disks 20 and 22 of RAID array 14 according to a predefined scheme. Array controller 12 is a server, although in some arrangements, array controller 12 may be a dedicated unit of a server, a personal computer, a laptop computer, or the like. Array controller 12 includes generator module 32.

Generator module 32 is constructed and arranged to generate codewords 24 from blocks of payload data 16 according to a predefined scheme, such as a round-robin scheme. Along these lines, generator module encodes a first block 34(1) of payload data 16 using generator matrix 26(1), a second block 34(2) of payload data 16 using generator matrix 26(2), a third block 34(3) of payload data 16 using generator matrix 26(3), and then subsequent blocks repeating through the generator matrices 26. Generator module 32 is typically software running on array controller 12, although in some arrangements, generator module 32 is a stand-alone piece of hardware.

RAID array 14 is constructed and arranged to store codewords 24 in payload disks 20 and parity disks 22 according to a prespecified scheme. RAID array 14 includes disks 20(1), 20(2), 20(3), and 20(4) (disks 20), and disks 22(1) and 22(2) (disks 22).

During operation, array controller 12 receives payload data 16 over communications medium 18. Payload data 16 is broken into blocks 34(1), 34(2), 34(3) (blocks 34) of length k ; in some

arrangements, array controller 16 breaks payload data into blocks 34. In turn, generator module 32 takes in block 34(1) and applies generator matrix 26(1) to it to create codeword 24(1) of length n . Analogously, generator module 32 applies generator matrix 26(2) to block 34(2) to create codeword 24(2) and generator matrix 26(3) to block 34(3) to create codeword 24(3). Generator module 32
 5 applies generator matrices 26 to subsequent blocks of payload data 16 (not shown) in a round-robin fashion.

It should be understood that the code $\mathcal{C}$ underlying generator module 32 is, in this example, a systematic code. This implies that the generator matrices 26 include columns of a $k \times k$ identity matrix representing block 34 and columns of a $k \times (n - k)$ parity matrix representing parity data in
 10 codeword 24. Array controller 12 assembles generator matrices 26 such that the parity data in codeword 24 occupy different positions according to which generator matrix 26 generator module encodes blocks 34.

As an example to illustrate the above discussion, consider a $(6,3,3)$ linear binary code with generator matrix 26(1):

$$15 \quad G^{(1)} = \begin{pmatrix} 1 & 0 & 0 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 & 1 \end{pmatrix}$$

Here disks 1,2,3 are payload ones, and 4,5,6 are parity ones. One may verify that the following matrices 26(2) and 26(3), respectively, are also generator matrices for code $\mathcal{C}$:

$$G^{(2)} = \begin{pmatrix} 1 & 1 & 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}, \text{ and}$$

$$G^{(3)} = \begin{pmatrix} 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 1 & 0 \\ 0 & 1 & 1 & 0 & 0 & 1 \end{pmatrix}.$$

20 In $G^{(2)}$, disks 4,5,6 are payload, and 1,2,3 are parity; in $G^{(3)}$, disks 1,4,6 are payload and 2,3,5 are parity. It should also be understood that arbitrary arrangements of information symbols, i.e., payload data, within codewords 24 are not necessarily compatible with the code $\mathcal{C}$. A constraint on the derived generator matrices is that, when the columns of the generator matrices 26 corresponding to parity data occupy a particular set of indices, the columns of nominal generator
 25 matrix 26(1) at that particular set of indices should be linearly independent. Such a set of indices in which linear independence of the columns in generator matrix 26(1) is achieved will be described below as an information set.

Array controller 12 then sends each codeword 24 to RAID array 14 to be stored in a

particular stripe 28 in disks 20 or 22. In this example, array controller 14 stores those codewords that had been encoded with matrix 26(1) in stripe 28(1), those encoded with matrix 26(2) in stripe 28(2), and those encoded with matrix 26(3) in stripe 28(3). Further details of array controller 12 are described below with respect to Figure 2.

5 Figure 2 illustrates details of an example array controller 12. Array controller 12 includes controller 40, which in turn includes processor 44 and memory 46, and data interface 42, such as SATA, SAS, Fiber Channel, etc.

Memory 46 is configured to store code 48 that contains instructions configured to cause processor 44 to carry out the improved technique. For example, for array controller 12, code 48
10 contains instructions for applying generator matrices 26 (see Figure 1) to blocks 34. Memory 46 generally takes the form of, e.g., random access memory, flash memory or a non-volatile memory.

Processor 44 takes the form of, but is not limited to, ARM, Intel or AMD-based MPUs, and can include a single or multi-cores each running single or multiple threads. In some arrangements, processor 44 is one of several processors working together. Processor 44 is configured to carry out
15 the improved technique by executing code 48. Processor 44 includes generator module 32, although in some arrangements, generator module 32 may be a stand-alone hardware module or software residing in memory.

It should be understood that the point of using the different generator matrices 26(1), 26(2), and 26(3) is to balance the load incurred through updating across stripes 28. Along these lines,
20 suppose that ρ is an update rate of information symbols. Suppose further that only the generator matrix 26(1) was being used over all stripes. Then a disk update rate, i.e., an update rate of the j th symbol of a codeword 24 is given by

$$\lambda_j = \rho w_{1j},$$

where $w_{1j} \geq 1$ is the number of nonzero elements in the j th column of $G^{(1)}$. That is, with
25 only a single generator matrix 26(1) for all blocks 34, the parity symbols in codewords 24 are updated more than the information symbols; because the parity and information symbols are stored in the same disks for every codeword, the disks 22 storing parity symbols are subject to much more wear than those 20 storing information symbols.

Taking into account the different generator matrices 26, however, the expression for the disk
30 update rate becomes

$$\lambda_j = \frac{\rho}{\ell} \sum_{i=1}^{\ell} w_{ij},$$

where $w_{ij} \geq 1$ is the number of nonzero elements in the j th column of $G^{(1)}$. Disk imbalance is now averaged over stripes 28. (In the case illustrated in Figure 1, $\ell = 3$.)

The effect of this averaging can be seen explicitly in the matrices illustrated above. For $G^{(1)}$ alone, the update rates for each codeword symbol is

$$(1 \ 1 \ 1 \ 3 \ 2 \ 1)\rho,$$

while for the different generator matrices 26, the update rates become

$$\left(\frac{4}{3} \ \frac{5}{3} \ 2 \ \frac{5}{3} \ \frac{5}{3} \ \frac{4}{3}\right)\rho.$$

The latter update rates are more evenly distributed than the former update rates. There are any number of figures of merit to describe how evenly or unevenly update rates are distributed across disks 20 and 22, or stripes 28.

For example, consider the following figure of merit:

$$F(Q) = \sum_{j=1}^n |\lambda_j - \beta|,$$

where β is the average update rate across the codeword symbols, and Q represents a particular family of collections of information sets that indicates the positions of the information symbols derived from each generator matrix. In the former case, the collection of information sets which is now called Q_1 has a figure of merit $F(Q_1) = 4\rho$. In the latter case, the collection of information sets which is now called Q_2 has a figure of merit $F(Q_2) = \frac{10}{9}\rho$. Clearly, the figure of merit defined above is smaller when the load is more evenly distributed across the codeword symbols and, consequently, across the stripes.

The fact that a more balanced load is tied to a smaller figure of merit raises the question of an optimal load balance through minimization of $F(Q)$ over a family of collections of information sets. Such an optimization scheme is described below with respect to Figure 3.

Figure 3 illustrates a process 60 for optimizing a load balance across codewords symbols, i.e. a family of information sets. Process 60 includes steps 62, 64, 66, 68, 70, 72, 74, and 76.

In step 62, processor 44 produces the nominal generator matrix, which is referred to as $G^{(1)}$ above. It is assumed that it is known how to generate this matrix, and nothing further need be discussed here.

In step 64, processor 44 generates a family of collections Q_i , $1 \leq i \leq T$, of information sets I_l , $1 \leq l \leq \ell$. In doing this, processor 44 builds each information set according to a process described in detail below with respect to Figure 4.

In step 66, processor 44 generates a set of generator matrices $G^{(l)}$ corresponding to the information sets I_l , $1 \leq l \leq \ell$, belonging to a collection Q_i . In some arrangements, processor 44 generates a generator matrix from an information set I_l using a gaussian elimination algorithm which is known in the those skilled in the art.

In step 68, processor 44 produces, for the collection Q_i , an update rate from the set of generator matrices $G^{(l)}$ according to
$$\lambda_j = \frac{\rho}{\ell} \sum_{l=1}^{\ell} w_{lj}$$
 . It should be understood that this update rate may have been computed as part of step 64, as will be discussed below with respect to Figure 4.

In step 70, processor 44 forms the figure of merit
$$F(Q_i) = \sum_{j=1}^n |\lambda_j - \beta|$$
 as described above.

In step 72, processor 44 steps through the family of collections Q_i , $1 \leq i \leq T$ and chooses the collection Q_{i_0} that minimizes F .

In step 74, processor 44 assigns the generator matrices of the optimal collection Q_{i_0} of information sets to generator module 32 for converting blocks 34 into codewords 24.

Figure 4 illustrates further detail of step 64 in which processor 44 creates a collection Q_i of information sets from the nominal generator matrix. Step 64 further includes steps 78, 80, 82, 84, 86, 88, and 90.

In step 78, processor 44 initializes Q_i and an array of update rates λ_j , $1 \leq j \leq n$. Processor 44 then loops over $1 \leq s \leq \ell$.

In step 80, processor 44 initializes each information set $I_s^{(l)}$ of the collection Q_i . Processor 44 then loops over $1 \leq i \leq k$.

In step 82, processor 44 runs a random number generation operation on $1 \leq j \leq n$ to produce an index $j_s^{(l)}$.

In step 84, processor 44 adds the index $j_s^{(l)}$ to information set $I_s^{(l)}$ such that two criteria are satisfied. The first criterion is that the column $G_{j_s^{(l)}}^{(l)}$ is linearly independent from the columns $G_{j_o^{(l)}}^{(l)}$ for each $j_o^{(l)} \in I_s^{(l)}$. The second criterion is that the update rate $\lambda_{j_s^{(l)}}$ is maximized.

In step 86, processor 44 appends the index $j_s^{(i)}$ to the information set $I_s^{(i)}$.

In step 88, processor 44 adds to the value of the update rates λ_j a constant weighting factor

c times the column $G_{j_s^{(i)}}^{(i)}$.

In step 90, once the information set $I_s^{(i)}$ is complete (i.e., the indices that define the information set are present), then processor 44 adds $I_s^{(i)}$ to collection Q_i of information sets. This step finishes once the entire family of collections has been built in this way.

Figure 5 illustrates a method 100 of storing data in a redundant disk array that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n , including steps 102, 104, and 106. In step 102, a set of ℓ generator matrices is constructed, each generator matrix of the set of ℓ generator matrices being compatible with the code and corresponding to a data stripe of the redundant disk array, ℓ being at least 2. In step 104, a respective codeword is generated for a given information vector by multiplying the given information vector by a particular generator matrix of the set of ℓ generator matrices. In step 106, the respective codeword is stored in a data stripe of the redundant array to which the particular generator matrix corresponds.

While various embodiments of the invention have been particularly shown and described, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention as defined by the appended claims.

For example, it should be understood that some embodiments are directed to array controller 12, which is constructed and arranged to update data stored on a redundant array of disks having an array controller in a storage system, each disk of the redundant array of disks including a disk controller apart from the array controller. Some embodiments are directed to a process of updating data stored on a redundant array of disks having an array controller in a storage system, each disk of the redundant array of disks including a disk controller apart from the array controller. Also, some embodiments are directed to a computer program product which enables computer logic to update data stored on a redundant array of disks having an array controller in a storage system, each disk of the redundant array of disks including a disk controller apart from the array controller.

It should also be understood that some embodiments are directed to array controller 2, which is constructed and arranged to store data in a redundant disk array that employs a code which transforms an information vector of information symbols of length k into a codeword of code

symbols of length n . Embodiments are directed to a process of storing data in a redundant disk array that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n .

In other arrangements, array controller 12 is implemented by a set of processors or other
5 types of control/processing circuitry running software. In such arrangements, the software instructions can be delivered, within array controller 12, either in the form of a computer program product 120 (see Figure 2) or simply instructions on disk or in pre-loaded in memory 46 of array controller 12, each computer program product having a computer readable storage medium which stores the instructions in a non-volatile manner. Alternative examples of suitable computer readable
10 storage media include tangible articles of manufacture and apparatus such as CD-ROM, flash memory, disk memory, tape memory, and the like.

CLAIMS

What is claimed is:

- 5 1. In a storage system, a method of storing data in a redundant disk array that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n , the method comprising:

constructing a set of ℓ generator matrices, each generator matrix of the set of ℓ generator matrices being compatible with the code and corresponding to a data stripe of the
10 redundant disk array, ℓ being at least 2;

generating a respective codeword for a given information vector by multiplying the given information vector by a particular generator matrix of the set of ℓ generator matrices; and

15 storing the respective codeword in a data stripe of the redundant array to which the particular generator matrix corresponds.

2. A method as in claim 1,

wherein the code is a systematic code in which the codeword of length n contains the information vector of length k and a check symbol vector of check symbols of length $n - k$;

20 wherein each generator matrix of the set of ℓ generator matrices includes columns of a $k \times k$ identity matrix and columns of a $k \times (n - k)$ check matrix;

wherein a nominal generator matrix is defined to be a generator matrix having k of its columns being columns of the $k \times k$ identity matrix; and

wherein constructing the set of ℓ generator matrices includes:

25 for each generator matrix of the set of ℓ generator matrices, distributing by elementary row operations the columns of the $k \times k$ identity matrix and the $k \times (n - k)$ check matrix within that generator matrix, columns of the nominal generator matrix which have positions of the columns of that generator matrix that are columns of the $k \times k$ identity matrix being linearly independent.

- 30 3. A method as in claim 2,

wherein the method further comprises:

prior to constructing the set of ℓ generator matrices, constructing a collection of ℓ information sets, each information set of the collection of ℓ information sets corresponding to a generator matrix of the set of ℓ generator matrices and being a set of positions of code symbols within a codeword.

4. A method as in claim 3,

wherein distributing the columns of the $k \times k$ identity matrix and the $k \times (n - k)$ check matrix includes:

performing a gaussian elimination operation on the nominal generator matrix, the gaussian elimination operation producing a generator matrix with columns of the $k \times k$ identity matrix having positions within the generator matrix being given by the information set of the collection of ℓ information sets corresponding to the generator matrix.

5. A method as in claim 3, further comprising:

prior to distributing the columns of the $k \times k$ identity matrix and the $k \times (n - k)$ check matrix, forming a figure of merit indicative of an imbalance of a rate of updating symbols across the data stripes of the redundant disk array, the figure of merit being a function of a collection of information sets, the rate of updating symbols being indicative of a number of times a code symbol is used as an information symbol in the collection of information sets; and

performing an optimization operation on the figure of merit over a family of collections of information sets, the optimization operation producing an optimal collection of information sets;

wherein distributing the columns of the $k \times k$ identity matrix and the $k \times (n - k)$ check matrix includes:

for each information set of the optimal set of information sets, producing a generator matrix of the set of ℓ generator matrices with columns of the $k \times k$ identity matrix having positions within the generator matrix being given by an information set of the optimal collection of information sets.

6. A method as in claim 5,

wherein forming the figure of merit includes:

forming a sum of an absolute value of a difference between the rate of updating symbols of a collection of information sets and a specified mean update rate over the collection of information sets, the sum being taken over each symbol of the codeword.

7. A method as in claim 5,

wherein performing the optimization operation on the figure of merit includes:

generating the family of collections of information sets;

finding the particular collection of information sets that minimizes the figure of merit over the family of collections of information sets, that particular collection of information steps being the optimal collection of information steps.

8. A method as in claim 7,

wherein generating the family of collections of information sets includes:

for each collection of the family of collections:

producing, with a random number generator, a collection of position indices, each position index of the collection of position indices corresponding to an information set of the collection of information sets;

for each position index of the collection of position indices:

producing a set of position indices with i) columns of the nominal generator matrix which have positions of the set of position indices being linearly independent and ii) the number of nonzero elements of those columns being maximized.

9. A method as in claim 8,

wherein the code has an alphabet that belongs to $GF(2)^m$; and

wherein producing the set of position indices includes:

summing the columns of the nominal generator matrix to produce the number of nonzero elements of those columns being maximized.

10. An apparatus constructed and arranged to store data in a redundant disk array that employs a code which transforms an information vector of information symbols of length k into a

codeword of code symbols of length n , the apparatus comprising:

a data interface;

memory; and

a controller including controlling circuitry, the controlling circuitry being constructed and arranged to:

construct a set of ℓ generator matrices, each generator matrix of the set of ℓ generator matrices being compatible with the code and corresponding to a data stripe of the redundant disk array, ℓ being at least 2;

generate a respective codeword for a given information vector by multiplying the given information vector by a particular generator matrix of the set of ℓ generator matrices; and

store the respective codeword in a data stripe of the redundant array to which the particular generator matrix corresponds.

11. An apparatus as in claim 10,

wherein the code is a systematic code in which the codeword of length n contains the information vector of length k and a check symbol vector of check symbols of length $n - k$;

wherein each generator matrix of the set of ℓ generator matrices includes columns of a $k \times k$ identity matrix and columns of a $k \times (n - k)$ check matrix;

wherein a nominal generator matrix is defined to be a generator matrix with k of its columns being columns of the $k \times k$ identity matrix.

12. An apparatus as in claim 11,

wherein the controlling circuitry is further constructed and arranged to:

prior to constructing the set of ℓ generator matrices, access a collection of ℓ information sets, each information set of the collection of ℓ information sets corresponding to a generator matrix of the set of ℓ generator matrices and being a set of positions of code symbols within a codeword and having been constructed offline.

13. An apparatus as in claim 12,

wherein distributing the columns of the $k \times k$ identity matrix and the $k \times (n - k)$

check matrix includes:

performing a gaussian elimination operation on the nominal generator matrix, the gaussian elimination operation producing a generator matrix with columns of the $k \times k$ identity matrix having positions within the generator matrix being given by the information set of the collection of ℓ information sets corresponding to the generator matrix.

14. An apparatus as in claim 12,

wherein the controlling circuitry is further constructed and arranged to:

prior to distributing the columns of the $k \times k$ identity matrix and the $k \times (n - k)$ check matrix, forming a figure of merit indicative of an imbalance of a rate of updating symbols across the data stripes of the redundant disk array, the figure of merit being a function of a collection of information sets, the rate of updating symbols being indicative of a number of times a code symbol is used as an information symbol in the collection of information sets; and

perform an optimization operation on the figure of merit over a family of collections of information sets, the optimization operation producing an optimal collection of information sets;

wherein distributing the columns of the $k \times k$ identity matrix and the $k \times (n - k)$ check matrix includes:

for each information set of the optimal set of information sets, producing a generator matrix of the set of ℓ generator matrices with columns of the $k \times k$ identity matrix having positions within the generator matrix being given by an information set of the optimal collection of information sets.

15. An apparatus as in claim 14,

wherein forming the figure of merit includes:

forming a sum of an absolute value of a difference between the rate of updating symbols of a collection of information sets and a specified mean update rate over the collection of information sets, the sum being taken over each symbol of the codeword.

16. An apparatus as in claim 14,

wherein performing the optimization operation on the figure of merit includes:

generating the family of collections of information sets;

finding the particular collection of information sets that minimizes the figure of merit over the family of collections of information sets, that particular collection of information steps being the optimal collection of information steps.

17. An apparatus as in claim 16,

wherein generating the family of collections of information sets includes:

for each collection of the family of collections:

producing, with a random number generator, a collection of position indices, each position index of the collection of position indices corresponding to an information set of the collection of information sets;

for each position index of the collection of position indices:

producing a set of position indices with i) columns of the nominal generator matrix which have positions of the set of position indices being linearly independent and ii) the number of nonzero elements of those columns being maximized.

18. An apparatus as in claim 17,

wherein the code has an alphabet that belongs to $GF(2)^{\frac{n}{k}}$; and

wherein producing the set of position indices includes:

summing the columns of the nominal generator matrix to produce the number of nonzero elements of those columns being maximized.

19. A computer program product having a non-transitory, computer-readable storage medium

which stores code for storing data in a redundant disk array that employs a code which transforms an information vector of information symbols of length k into a codeword of code symbols of length n , the code including instructions which, when executed by the disk controller, causes the disk controller to:

construct a set of ℓ generator matrices, each generator matrix of the set of ℓ generator matrices being compatible with the code and corresponding to a data stripe of the redundant disk array, ℓ being at least 2;

generate a respective codeword for a given information vector by multiplying the

given information vector by a particular generator matrix of the set of ℓ generator matrices;
and

store the respective codeword in a data stripe of the redundant array to which the
particular generator matrix corresponds.

5

20. A computer program product as in claim 19,

wherein the code is a systematic code in which the codeword of length n contains the
information vector of length k and a check symbol vector of check symbols of length $n - k$;

wherein each generator matrix of the set of ℓ generator matrices includes columns of
a $k \times k$ identity matrix and columns of a $k \times (n - k)$ check matrix;

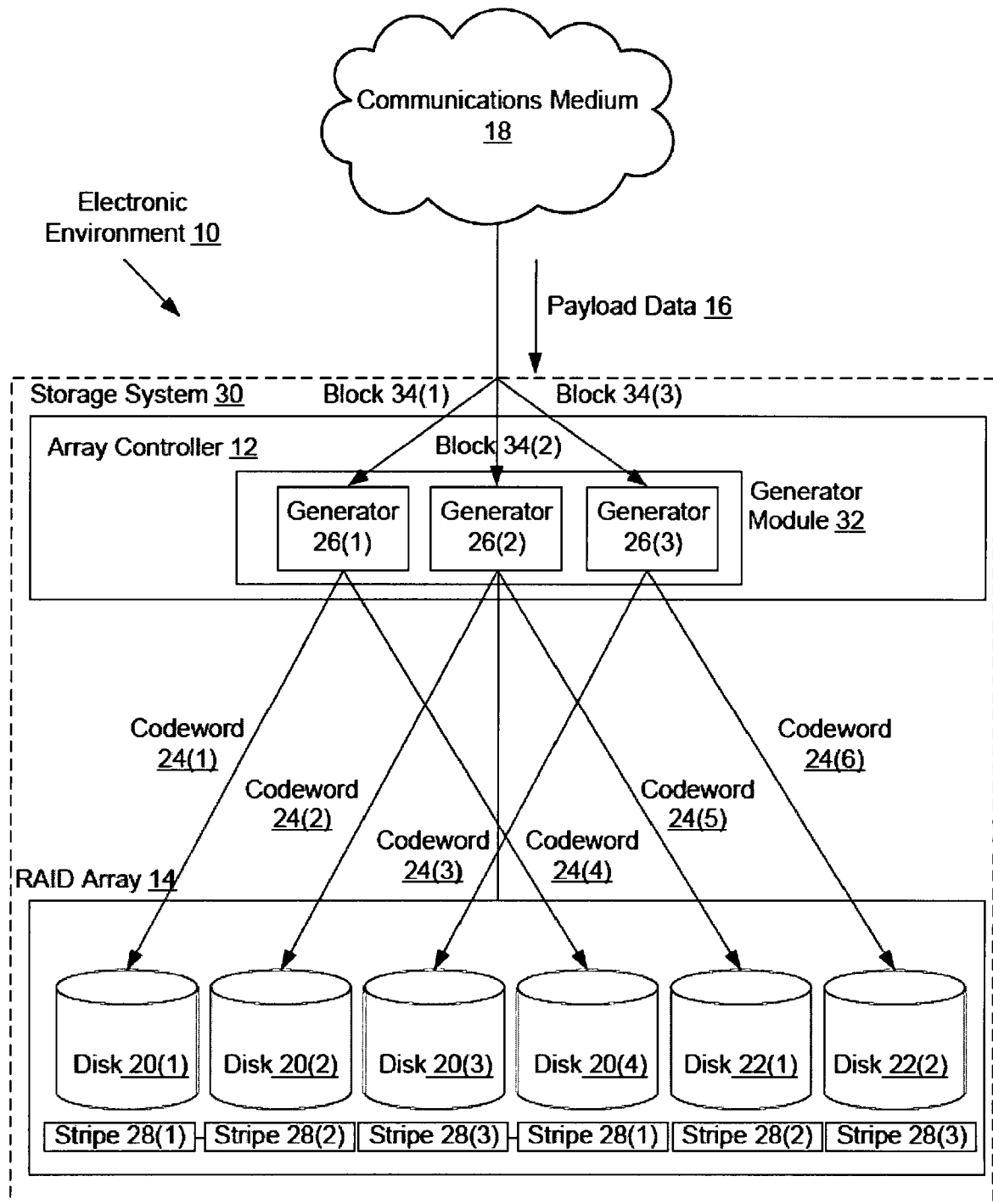
10

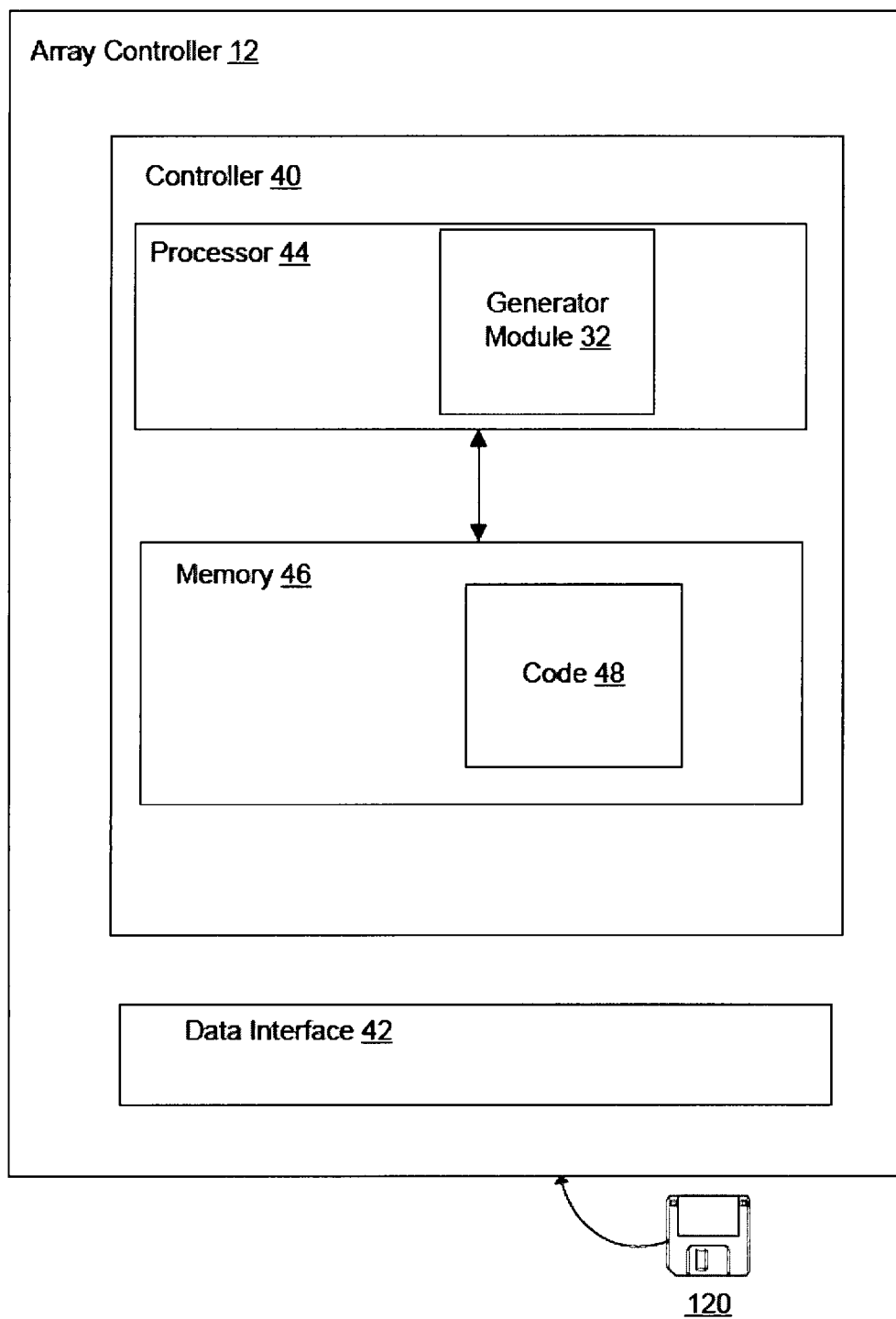
wherein a nominal generator matrix is defined to be a generator matrix with its first
 k columns being the $k \times k$ identity matrix; and

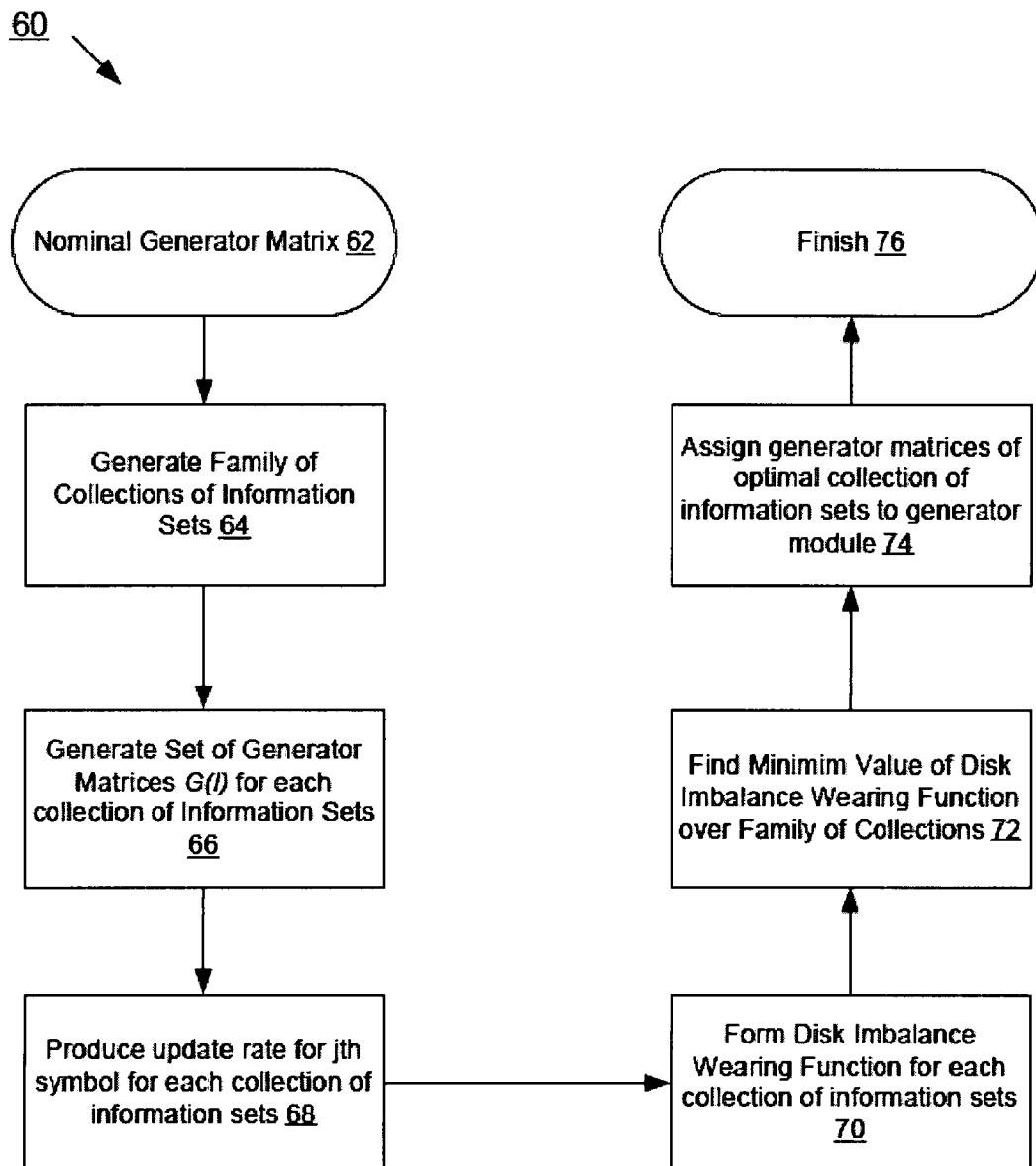
wherein constructing the set of ℓ generator matrices includes:

15

for each generator matrix of the set of ℓ generator matrices, distributing the columns
of the $k \times k$ identity matrix and the $k \times (n - k)$ check matrix within that generator matrix,
columns of the nominal generator matrix which have positions of the columns of that
generator matrix that are columns of the $k \times k$ identity matrix being linearly independent.

**Figure 1**

**Figure 2**

**Figure 3**

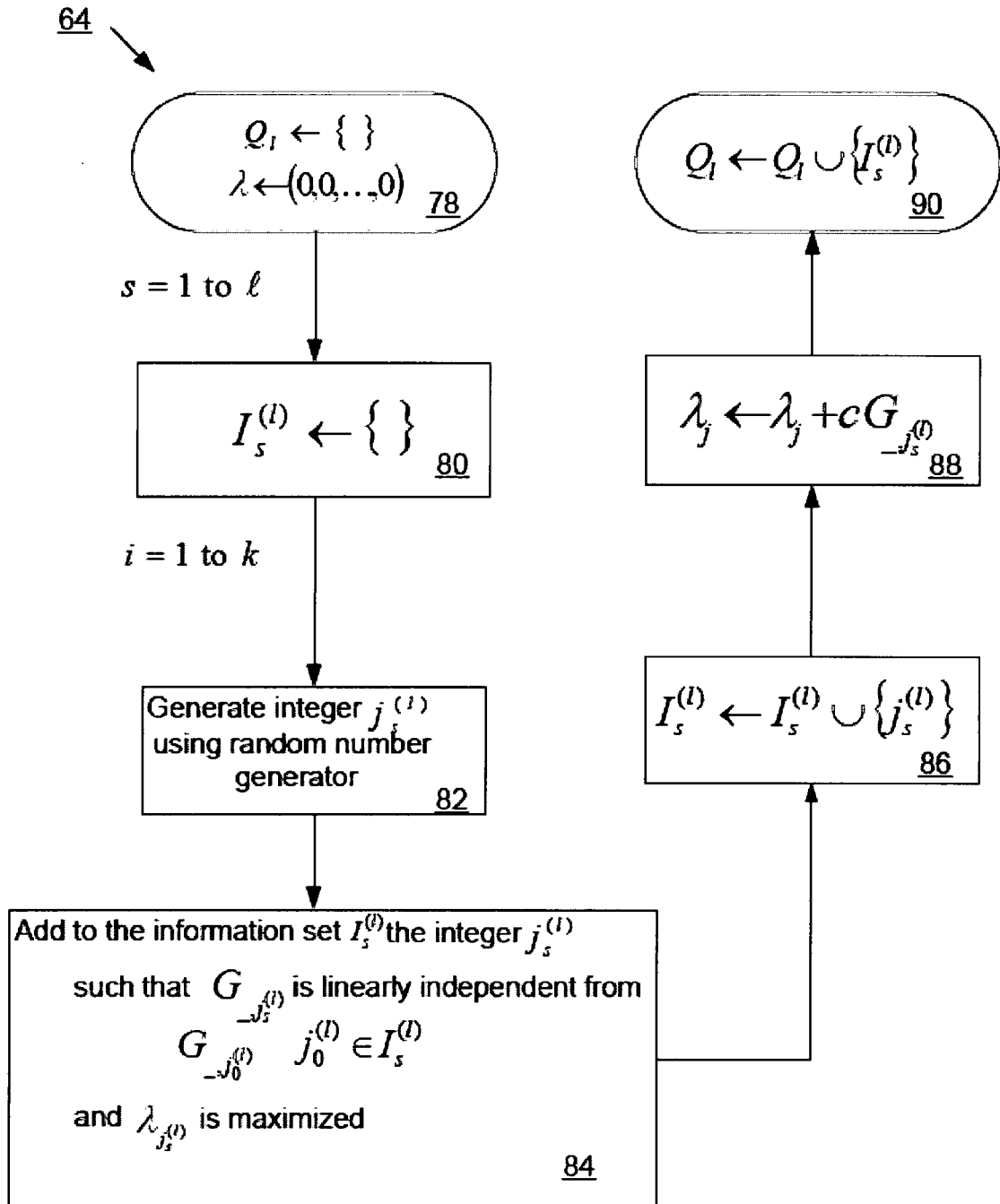
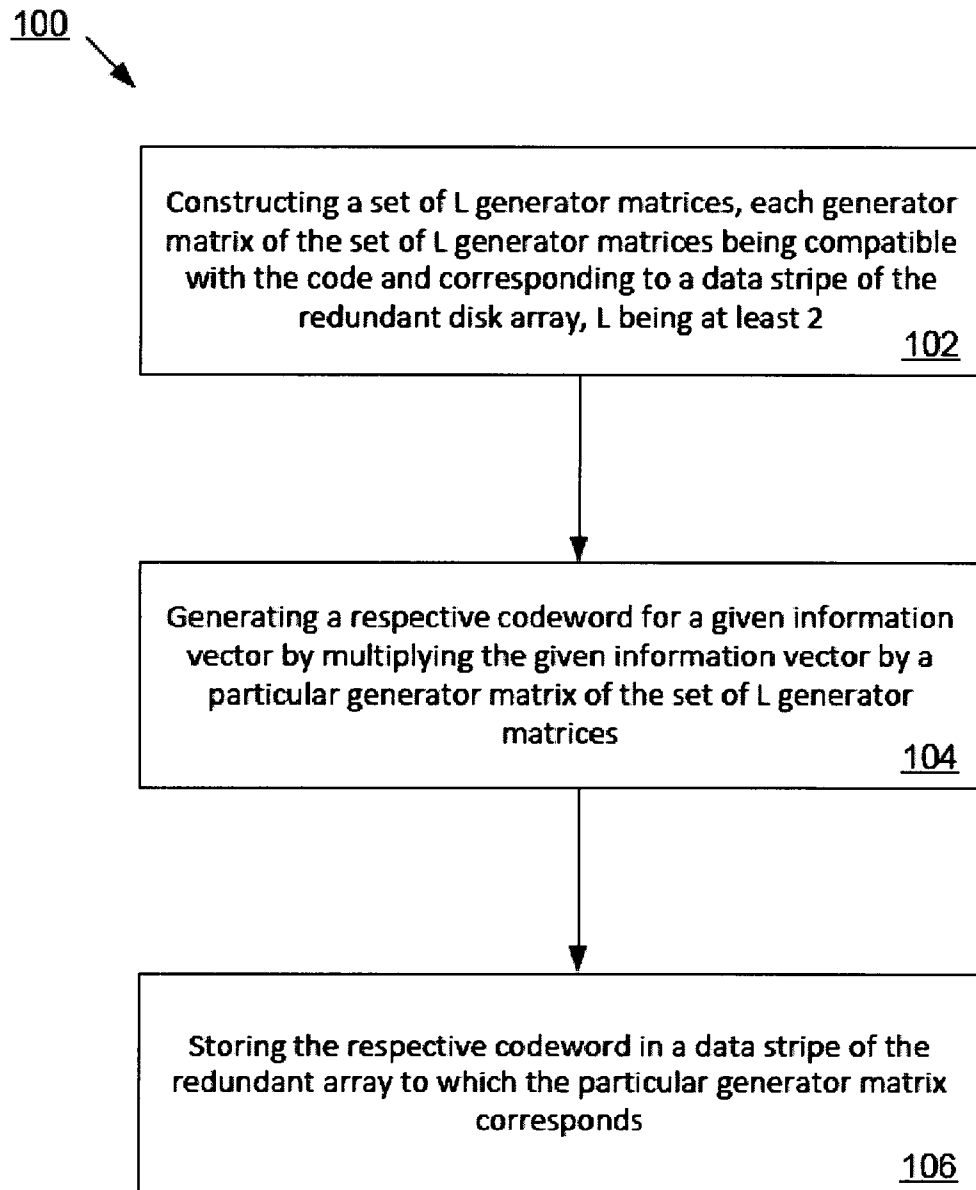


Figure 4

**Figure 5**

INTERNATIONAL SEARCH REPORT

International application No.

PCT/IB 2013/001188

A. CLASSIFICATION OF SUBJECT MATTER

*H03M 13/15 (2006.01)**G06F 12/00 (2006.01)*

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F 11/00-11/10, 12/00, 12/02, 12/06, 12/16, H03M 13/00, 13/03, 13/05, 13/13, 13/15, G11C 11/00, 15/00, 29/00

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatSearch (RUPTO internal), Esp@cenet, PAJ, USPTO, Information Retrieval System of FIPS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 2008/0115017 A1 (MICHAEL B. JACOBSON) 15.05.2008, paragraphs [0011], [0012], [0018], [0030]-[0033], [0035], [0039], [0046], [0047], [0050], claims 1, 3, 9, 14	1-4, 10-13, 19-20
A		5-9, 14-18
Y	US 2008/0183963 A1 (INTERNATIONAL BUSINESS MACHINES CORPORATION) 31.07.2008, paragraphs [0078], [0079], [0080]-[0082]	1-4, 10-13, 19-20
Y	US 2007/0290901 A1 (KONINKLIJKE PHILIPS ELECTRONICS, N.V.) 20.12.2007, paragraphs [0072], [0073], [0093]	2-4, 11-13

☐ Further documents are listed in the continuation of Box C.☐ See patent family annex.

* Special categories of cited documents:

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search

11 November 2013 (11.11.2013)

Date of mailing of the international search report

28 November 2013 (28.11.2013)

Name and mailing address of the ISA/ FIPS
Russia, 123995, Moscow, G-59, GSP-5,
Berezhkovskaya nab., 30-1

Facsimile No. +7 (499) 243-33-37

Authorized officer

I. Kryazhev

Telephone No. 8(499)240-25-91