



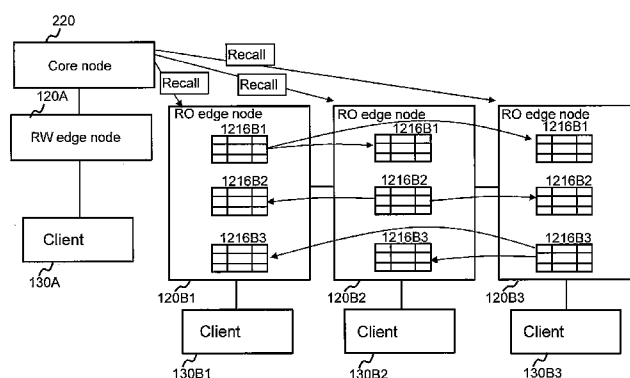
- (51) **International Patent Classification:**
H04L 29/08 (2006.01) *G06F 17/30* (2006.01)
- (21) **International Application Number:**
PCT/JP2011/001675
- (22) **International Filing Date:**
22 March 2011 (22.03.2011)
- (25) **Filing Language:** English
- (26) **Publication Language:** English
- (71) **Applicant (for all designated States except US):** **HITACHI, LTD.** [JP/JP]; 6-6, Marunouchi 1-chome, Chiyoda-ku, Tokyo, 1008280 (JP).
- (72) **Inventors; and**
- (75) **Inventors/Applicants (for US only):** **YAMADA, Hajime** [JP/JP]; c/o HITACHI, LTD., Disk Array Systems Division, 322-2, Nakazato, Odawara-shi, Kanagawa, 2500872 (JP). **SAIKA, Nobuyuki** [JP/JP]; c/o HITACHI, LTD., Disk Array Systems Division, 322-2, Nakazato, Odawara-shi, Kanagawa, 2500872 (JP).
- (74) **Agent:** **WILLFORT INTERNATIONAL;** Kanda-Ogawamachi Tosei Bldg. II 7F, 3, Kanda-Ogawamachi 3-chome, Chiyoda-ku, Tokyo 1010052 (JP).

(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, RO, RS, RU, 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, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, 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:** FILE SERVER SYSTEM AND STORAGE CONTROL METHOD**Fig. 10**

(57) **Abstract:** In accordance with a first aspect of the present invention, an edge node is provided with the storage destination management information that is corresponded to the edge node. The storage destination management information includes the information that indicates a storage destination for a core node of a file that has been written to a file storage region (an edge storage region) for an edge node that is corresponded to the information and includes the information that indicates a recall date that is a date that has been read by the edge node. The edge node judges whether or not there is an old file (a file before an update) in an edge storage region for the edge node based on the storage destination management information that is corresponded to the edge node and the storage destination information that is corresponded to another edge node, and create a stub for only an old file without creating a stub for a file after an update (a latest file).

Description

Title of Invention: FILE SERVER SYSTEM AND STORAGE CONTROL METHOD

Technical Field

[0001] The present invention relates to a storage control for a file server system that is provided with a plurality of file servers.

Background Art

[0002] A hierarchical file server system that is configured by a plurality of file servers that are coupled to a communication network is publicly known for instance. As a hierarchical file server system of such a type, there can be mentioned for instance a system that is provided with an edge node that is a file server that receives a request from a client (a local file server) and a core node that is a file server that receives a request from an edge node (a remote file server).

[0003] As a hierarchical file server system, there can also be mentioned for instance a system in which an edge node and a core node are disposed on a one-to-one basis (see Patent Literature 1) and a system in which an edge node and a core node (for instance, a cloud storage) are disposed on an N-to-one basis (N is an integer number equal to or larger than 2) (in other words, a system in which a plurality of edge nodes shares one core node).

[0004] The following problem may occur unfortunately for the system in which an edge node and a core node are disposed on an N-to-one basis.

[0005] For instance, a plurality of edge nodes has already read a file that is included in a core node, a file of an edge node is updated, and a file after the update is stored from the edge node into the core node. However, other edge nodes do not recognize that a file that has already been read has been updated. Consequently, in the case in which other edge nodes receive a read request that specifies a file that has already been read (a file before an update), the edge nodes transmit the file before an update to a client unfortunately.

[0006] As a method for solving the problem, it is thought that each of the edge nodes is provided with a stub of a file that is included in a core node for instance. The stub is an object (for instance, metadata) that is associated with the storage destination information of the file (the information that indicates a link destination). According to this method, in the case in which an edge node receives a read request that specifies a stub, the edge node reads a file of a link destination of the stub (that is, a file in a core node). Consequently, the edge node does not transmit the file before an update to a client.

- [0007] In this case however, in each case in which each of the edge nodes receive a read request that specifies a stub, the edge nodes access a core node. Consequently, accesses are concentrated to one core node, whereby a load to the core node is increased.
- [0008] As a method for solving the problem, it is thought that a plurality of core nodes and an intermediate apparatus that is coupled to the plurality of core nodes and a plurality of edge nodes are prepared for instance. As a literature that is associated with the method of this type, Patent Literature 1 and Patent Literature 2 can be mentioned for instance.
- [0009] In accordance with Patent Literature 2 for instance, a virtual server is configured by a plurality of servers, and a load distribution apparatus is adopted as an intermediate apparatus. The load distribution apparatus distributes a plurality of accesses executed to the virtual server from a plurality of clients to a plurality of servers that configure the virtual server.
- [0010] Moreover in accordance with Patent Literature 3 for instance, a file that is included in a first server is made to be a stub, and the file is migrated to a second server. In the case in which an intermediate apparatus receives an access request that specifies a stub from a client, the intermediate apparatus responds to the client by the file that is included in the second server.

Citation List

Patent Literature

- [0011] PTL 1: Japanese Patent Application Laid-Open Publication No. 2009-289252
PTL 2: Japanese Patent Application Laid-Open Publication No. 2009-265778
PTL 3: Japanese Patent Application Laid-Open Publication No. 2006-164211

Summary of Invention

Technical Problem

- [0012] However, in accordance with the above method, a plurality of core nodes is required for a plurality of edge nodes and an intermediate apparatus (such as a load distribution apparatus) is also required, thereby increasing a cost unfortunately.
- [0013] Moreover, even in the case in which a migration of a file from a core server to another core server is necessary, if a vendor of a core server is different from a vendor of another core server, a migration of a file from a core server to another core server cannot be executed.
- [0014] An object of the present invention is to suppress the number of times of accesses for a core node without transmitting a file before an update to a client even in the case in which a plurality of core nodes and an intermediate apparatus are not disposed.

Solution of Problem

- [0015] In accordance with a first aspect of the present invention, an edge node is provided

with the storage destination management information that is corresponded to the edge node. The storage destination management information includes the information that indicates a storage destination for a core node of a file that has been written to a file storage region (an edge storage region) for an edge node that is corresponded to the information and includes the information that indicates a recall date that is a date that has been read by the edge node. The edge node judges whether or not there is an old file (a file before an update) in an edge storage region for the edge node based on the storage destination management information that is corresponded to the edge node and the storage destination information that is corresponded to another edge node, and create a stub for only an old file without creating a stub for a file after an update (a latest file).

[0016] In accordance with a second aspect of the present invention, a plurality of edge nodes is divided into at least two groups. In the case in which an edge node receives a read request of a read target file that is a file that is corresponded to a stub from a client, the edge node executes the following processing:

(D) judging whether or not the read target file has been stored into a file storage region for any other second edge node that is included in a target group that is a group that includes the edge node;

(E) executing a recall of the read target file in the case in which the result of a judgment of the above (D) is negative; and

(F) transmitting a read request of the read target file to other edge node that is corresponded to a file storage region that has stored the read target file and transmitting the read target file that has been read by a response to the read request to a client in the case in which the result of a judgment of the above (D) is positive.

The at least two groups can be configured in such a manner that a duplicated file is not preferably included in a plurality of files that has been stored into at least two edge storage regions for at least two second edge nodes that are included in one group. Moreover, the storage destination management information described above can also be used in the judgment of the above (D). In the second aspect of the present invention, it is also possible that the storage destination management information does not include the information that indicates a recall date. Moreover, a storage destination for a core node of a read target file that has been stored into a file storage region for other edge node can also be equivalent to a storage destination that is indicated by a stub.

[0017] The file storage region described above can be disposed inside or outside an edge node. Moreover, although the file storage region is a file system in a representative pattern, the file storage region can also be a storage region of other type.

Brief Description of Drawings

[0018] [fig.1]Fig. 1 is a view showing a general outline in accordance with an embodiment 1

of the present invention.

[fig.2]Fig. 2 is a view showing a hardware configuration of an entire system in accordance with an embodiment 1 of the present invention.

[fig.3]Fig. 3 is a view showing a hardware configuration of an entire system in accordance with an embodiment 1 of the present invention.

[fig.4]Fig. 4 is a view showing a computer program and information that are included in an edge node 120.

[fig.5]Fig. 5 is a view showing a configuration of a file system.

[fig.6]Fig. 6 is a view showing a part related to a directory in an inode management table.

[fig.7]Fig. 7 is a view showing a part related to a file in an inode management table.

[fig.8]Fig. 8 is a view showing a relationship between a file "a.txt" and a location of data that configures the file.

[fig.9]Fig. 9A is a view showing a configuration of a storage destination management table 1216. Fig. 9B is a view showing a group management table 1217.

[fig.10]Fig. 10 is a view showing a flow of a processing in which a plurality of RO edge nodes 120B share the storage destination management table 1216.

[fig.11]Fig. 11 is a view showing a storage destination management table 1216B1 that is corresponded to an RO edge node 120B1 and a storage destination management table 1216B2 that is corresponded to an RO edge node 120B2.

[fig.12]Fig. 12A is a view showing an example of an edge table creation processing based on a file that has already been cached for each RO edge node. Fig. 12B is a view showing a result of the edge table creation processing shown in Fig. 12A. Fig. 12C is a view showing a file that has already been cached for each of the groups shown in Fig. 12B.

[fig.13]Fig. 13A is a view showing an example of a configuration of another edge group. Fig. 13B is a view showing an example of a problem for another edge group.

[fig.14]Fig. 14 is a view showing an example of a processing that is executed in the case in which an RO edge node 120B receives a read request that specifies a stub.

[fig.15]Fig. 15 is a part of a flowchart showing a flow of a processing of an RW mover 1210.

[fig.16]Fig. 16 is a remaining part of the flowchart of Fig. 15.

[fig.17]Fig. 17 is a part of a flowchart showing a flow of a processing of an RO mover 1210.

[fig.18]Fig. 18 is a flowchart showing a flow of a processing that is executed by a reception program 22110 of an RW edge node 120A.

[fig.19]Fig. 19 is a part of a flowchart showing a flow of a processing that is executed by a reception program 12110 of an RO edge node 120B.

[fig.20]Fig. 20 is a remaining part of the flowchart of Fig. 19.

[fig.21]Fig. 21 is a flowchart showing a flow of a processing that is executed by a storage destination information sharing program 1215.

[fig.22]Fig. 22 is a flowchart showing a flow of a processing that is executed by a load relaxation control program 1214.

[fig.23]Fig. 23 is a view showing a hardware configuration of an entire system in accordance with an embodiment 2 of the present invention.

[fig.24]Fig. 24 is a view showing that a storage destination management table 1216 that is included in a management server 300 is shared by a plurality of RO edge nodes 120B.

[fig.25]Fig. 25 is a flowchart showing a flow of a processing that is executed by a storage destination information sharing program 1215 in accordance with an embodiment 2 of the present invention.

[fig.26]Fig. 26 is a flowchart showing a flow of a processing that is executed by a storage destination information sharing program 310 in a management server 300.

Description of Embodiments

[0019] Some embodiments (examples) of the present invention will be described below in detail with reference to the drawings.

[0020] In the following descriptions, while a wide variety of information will be described in the expression of "xxx table" in some cases, a wide variety of information can be represented by a data structure other than a table. In order to indicate that a data structure is not depended on, "xxx table" can also be referred to as "xxx information".

[0021] In the following descriptions, while an ID (identifier) is used for specifying an element, a name or a number can also be used as an ID.

[0022] In the following descriptions, the processing will be described while a "program" is handled as a subject in some cases. In the case in which the program is executed by a processor (for instance, a CPU (Central Processing Unit)) that is included in a controller, the processor executes the predetermined processing by using a storage resource (such as a memory) and/or a communication interface apparatus (such as a communication port) as it decides proper. Consequently, a subject of a processing can also be a processor. The processing that is described while a program is handled as a subject can also be a processing that is executed by a file server (such as an RO (Read Only) edge node, an RW (Read Write) edge node, and a core node that are described later) or a management server. Moreover, the controller can include a hardware circuit that executes a part or a whole of a processing. A computer program can be installed from a program source to each of the computers (such as a file server and a management server). The program source can be a program distribution server or a

storage medium that can be read by a computer for instance.

[0023] A management server can also be at least one computer. For instance, in the case in which a management server indicates the information or a management server transmits the information for an indication to a remote computer, one computer is a management server. Moreover, in the case in which a function equivalent to a management server is implemented by using a plurality of computers for instance, the plurality of computers is a management server (the plurality of computers can include a computer for an indication in the case in which a computer for an indication executes an indication).

[0024] In the following descriptions moreover, in the case in which elements of the same type (for instance, a client, an edge node, and an RO edge node) are not distinguished for a description, only common signs among referential signs are used. On the other hand, in the case in which elements of the same type are distinguished for a description, referential signs (a combination of a common sign and an individual sign) are used, or the identification information that is added to an element is used as substitute for a referential sign.

[0025] Moreover, while a term of a "date" will be used in the following descriptions, the "date" includes at least any one of a year, a month, a day, an hour, a minute, and a second (that can include a second after the decimal point). For instance, the "date" can include a year, a month, and a day, and the "date" can also include an hour, a minute, and a second in addition to them.

Embodiment 1

[0026] Fig. 1 is a view showing a general outline in accordance with an embodiment 1 of the present invention.

[0027] A server system in accordance with the present embodiment is provided with a plurality of edge nodes 120 that are coupled to a plurality of clients (for instance, a computer such as a personal computer) 130 and a core node 220 that is coupled to a plurality of edge nodes 120. In the present embodiment, the server system is provided with one core node 220, and the core node 220 is shared by the plurality of edge nodes 120.

[0028] The edge node 120 is a local file server (that is, a file server that receives a request from a client 130), and the core node 220 is a remote file server (that is, a file server that receives a request from the edge node 120). The client 130 and the edge node 120 can be coupled to a first communication network (such as a LAN (Local Area Network)), and the edge node 120 and the core node 220 can be coupled to a second communication network (for instance, a communication network that is independent from the first communication network (such as the Internet)). In the example shown in Fig. 1, the clients 130A, 130B1, 130B2, and 130B3 are coupled to the edge nodes

120A, 120B1, 120B2, and 120B3, respectively.

[0029] Each of the edge nodes 120 is provided with a file storage region (hereafter referred to as an edge storage region). The core node 220 is also provided with a file storage region (hereafter referred to as a core storage region). The edge storage region and the core storage region are a file system representatively. The edge storage region (the core storage region) that is included in a node 120 (130) can be disposed inside or outside the node 120 (130). Moreover, each of the edge nodes 120 can write a file to an edge storage region that is included in the node 120 and can read a file from the edge storage region. However, each of the edge nodes 120 cannot write a file to an edge storage region that is not included in the node 120 and cannot read a file from the edge storage region.

[0030] A plurality of edge nodes 120 is configured by at least one RW edge node 120A and RO edge nodes 120B that are more than the RW edge node 120A. Although one RW edge node 120A and three RO edge nodes 120B are disposed in the example shown in Fig. 1, the number of the RO edge nodes 120B can be less or larger than 3 while one RW edge node 120A is disposed.

[0031] The RW edge node 120A is an edge node that receives both of a write request and a read request of a file, in other words, an edge node that can not only read a file from a core storage region but also write a file to a core storage region. More specifically, in the case in which the RW edge node 120A receives a write request of a file from a client 130A for instance, the RW edge node 120A writes the file to an edge storage region of the node 120A and writes the file to the core storage region. Moreover in the case in which the RW edge node 120A receives a read request of a file from a client 130A for instance, the RW edge node 120A returns the file in an edge storage region to the client 130A if the file is already in the edge storage region of the node 120A (that is, if the file has already been cached). On the other hand, in the case in which the file has not been cached, the RW edge node 120A reads the file from the core storage region (caches the file), and returns the file to the client 130A.

[0032] The RO edge node 120B is an edge node that receives a read request of a file and that does not receive a write request of a file, in other words, an edge node that reads a file from a core storage region and that does not write a file to a core storage region. More specifically, in the case in which the RO edge node 120B1 receives a read request of a file from a client 130B1 for instance, the RO edge node 120B1 returns the file in an edge storage region to the client 130B1 if the file is already in the edge storage region of the node 120B1 (that is, if the file has already been cached). On the other hand, in the case in which the file has not been cached, the RO edge node 120B1 reads the file from the core storage region and returns the file to the client 130B1 (the file that has been read can be cached). Moreover in the case in which a file that is specified by a

read request is a stub of the file, the RO edge node 120B1 reads the file from a storage destination that is specified by the stub (that is, a region in the core storage region) and returns the file to the client 130B1. In the processing in which the RO edge node 120B reads a file from the core storage region, the RO edge node 120B transmits a read request of a file to a core node 220, and the core node 220 responds to the read request and returns the file that is specified by the read request to the RO edge node 120B.

[0033] Each of the RO edge nodes 120B is provided with a table (hereafter referred to as a storage destination management table) 1216 configured to manage a storage destination (a region in the core storage region) of a file that has been cached by the RO edge node 120B. The storage destination management table 1216 can be managed inside or outside the RO edge node 120B. The storage destination management table 1216 is managed inside the RO edge node 120B in the present embodiment, and the storage destination management table 1216 is managed outside the RO edge node 120B in an embodiment 2 describer later.

[0034] The following processing is executed by the above configuration in the present embodiment for instance.

[0035] The RO edge node 120B executes a recall of a file (that is, reads a file from a core storage region and caches the file) in a repetitive manner (on a regular basis or on an irregular basis). At this time, the RO edge node 120B updates the storage destination management table 1216 that is included in the node 120B.

[0036] For instance, a file that has been cached by the RW edge node 120A is updated, and a file after the update is stored into a core storage region from the RW edge node 120A. After that, the RO edge node 120B1 executes a recall of the file after the update from the core storage region, and updates the storage destination management table 1216 that is included in the node 120B1.

[0037] Consequently, although the file after the update is cached to the RO edge node 120B1, a file before the update is cached to the RO edge nodes 120B2 and 120B3 and a file after the update is not cached to the RO edge nodes 120B2 and 120B3. Consequently, the RO edge node 120B2 (or 120B3) receives a read request that specifies a file before the update from the client 130B2 (or 130B3), and the RO edge node 120B2 (or 120B3) responds to the read request and returns the cached file before the update to the client 130B2 (or 130B3).

[0038] To solve the above problem, it is thought that a stub is created for all files that have been cached by the RO edge nodes 120B2 and 120B3. However in this case, in each case in which the RO edge nodes 120B2 and 120B3 receive a read request, the RO edge nodes 120B2 and 120B3 access a core node 220 (transmit a read request to the core node 220). Consequently, a load of the core node 220 is increased unfortunately.

[0039] In the present embodiment, a stub is created for only a file before an update and a

stub is not created for a file that have not been updated among all files that have been cached by the RO edge nodes 120B1 to 120B3 by the following processing. In other words, the number of files for which a stub is created can be reduced.

- [0040] More specifically, the RO edge node 120B2 compares the storage destination management table 1216 that is managed by the node 120B2 with the storage destination management table 1216 that is included in each of the other RO edge nodes 120B1 and 120B3 for instance. The RO edge node 120B2 specifies a file before an update (a file for which a stub is to be created) among at least one file that has been cached for the node 120B2 based on the result of the comparison. The RO edge node 120B2 then creates a stub for only the specified file before an update (a stub in which a link destination is a file after an update in a core storage region).
- [0041] By the above configuration, in the case in which the RO edge node 120B2 receives a read request that specifies a stub from a client 130B2, the RO edge node 120B2 can read a file after an update that is specified by the stub from a core storage region and can return the file after the update to the client 130B2. In other words, in the present embodiment, a file before an update can be prevented from being transmitted to a client 130B. In the present embodiment moreover, since the number of files for which a stub is created can be reduced as described above, the number of accesses to a core node 220 can be reduced.
- [0042] Fig. 2 is a view showing a hardware configuration of an entire system in accordance with the present embodiment.
- [0043] The system is provided with an edge system 100 and a core system 200. The edge system 100 is a stronghold that includes a local computer system, for instance a stronghold such as a branch office or a sales office in which a user executes a business as a practical matter. The core system 200 is a stronghold that includes a remote computer system, for instance a stronghold for managing a server or a storage apparatus in a lump or a stronghold for providing a cloud service.
- [0044] Although a plurality of edge systems 100 are shown in the example of Fig. 2, only one edge system 100 can also be disposed. Moreover, although a plurality of clients 130 that are coupled to an edge node 120 is shown in the example of Fig. 2, only one client can also be disposed.
- [0045] The edge systems 100 is provided with a RAID (Redundant Array of Independent (or Inexpensive) Disks) system 110, an edge node 120, and a client 130.
- [0046] The edge node 120 is coupled to the client 130 via a communication network (such as a LAN) 10 for instance. Moreover, the edge node 120 is coupled to the RAID system 110 via a communication network (such as a SAN (Storage Area Network)) that is equivalent to or separate from the communication network 10 for instance.
- [0047] The RAID system 110 is classified broadly into a controller part and a storage part.

The controller part is provided with a CHA (Channel Adaptor) 111 and a DKC (DisK Controller) 112. The storage part is provided with a DISK 113. The CHA 111 and the DISK 113 are coupled to the DKC 112. The CHA 111 is a communication interface apparatus that is coupled to the edge node 120. The DKC 112 is a controller. The DISK 113 is a physical storage device of a disk type (such as an HDD (Hard Disk Drive)). As a physical storage device, a physical storage device of other type (such as a flash memory device) can also be adopted. Although one DISK 113 is disposed in the example shown in Fig. 2, a plurality of DISKs 113 can also be disposed. A plurality of DISKs 113 can configure at least one RAID group. In the present embodiment, the RAID system 110 is shared by a plurality of edge nodes 120. A plurality of edge storage regions (such as an LU (Logical Unit) that is used as a file system) that are corresponded to a plurality of edge nodes 120 are disposed based on the DISK 113.

- [0048] The RAID system 110 receives an I/O (Input/Output) request of a block level that has been transmitted from the edge node 120 by using the CHA 111, and executes an I/O to the appropriate DISK 113 based on a control of the DKC 112.
- [0049] The edge node 120 is provided with a memory 121, a CPU (Central Processing Unit) 122, and a NIC (Network Interface Card) 123, and an HBA (Host Bus Adaptor) 124. The CPU 122 is coupled to the memory 121, the NIC 123, and the HBA124.
- [0050] The NIC 123 is a communication interface apparatus that is configured to communicate with the core node 220 and the client 130.
- [0051] The HBA124 is a communication interface apparatus that is configured to communicate with the RAID system 110.
- [0052] The memory 121 is a storage region (such as a RAM (Random Access Memory) and a ROM (Read Only Memory)) that can be read and written by the CPU 122. The CPU 122 executes a computer program that has been stored into the memory 121, whereby a wide variety of functions can be implemented for the edge node 120. The edge node 120 can also be provided with storage resources of other kinds in addition to or as substitute for the memory 121.
- [0053] The edge node 120 receives an I/O request of a file level from the client 130 via the NIC 123. The edge node 120 creates an I/O request (an I/O request of a block level) for an I/O of a data block that configures a file that is specified by the I/O request. The edge node 120 transmits an I/O request of a block level to the RAID system 110 via the HBA 124.
- [0054] The client 130 is provided with a memory 131, a CPU 132, a NIC 133, and a DISK 134. The client 130 can also be provided with storage resources of other kinds in addition to or as substitute for the memory 131 and/or the DISK 134.
- [0055] The CPU 132 executes a computer program that has been stored into the memory 131, whereby a wide variety of functions can be implemented for the client 130. The

client 130 transmits an I/O request of a file level to the file storage apparatus 120 via the NIC 133. The DISK 134 is a physical storage device of a disk type (such as an HDD (Hard Disk Drive)).

[0056] The core system 200 is provided with a RAID system 210 and a core node 220. The RAID system 210 is coupled to the core node 220.

[0057] The RAID system 210 is provided with a CHA 211, a DKC 212, and a DISK 213. In the example shown in Fig. 2, the configuration of the RAID system 210 in the core system 200 is equivalent to that of the RAID system 110 in the edge system 100. Consequently, for the RAID system 210, the CHA 211 receives an I/O request of a block level that has been transmitted from the core node 220, and an appropriate I/O to the DISK 213 is executed based on a control of the DKC 212. It is also possible that the configuration of the RAID system 210 is different from that of the RAID system 110.

[0058] The core node 220 is provided with a memory 221, a CPU 222, a NIC 223, and an HBA 224. The core node 220 can also be provided with storage resources of other kinds in addition to or as substitute for the memory 221. The CPU 222 executes a computer program that has been stored into the memory 221, whereby a wide variety of functions can be implemented for the core node 220. Moreover, the core node 220 communicates with the edge node 120 via the NIC 223 and a communication network 1 (such as the Internet). The core node 220 transmits an I/O request of a block level to the RAID system 210 via the HBA 224.

[0059] Fig. 3 is a view showing a hardware configuration of an entire system in accordance with an embodiment 1 of the present invention.

[0060] The RAID system 110 (210) is provided with an LU 1101 (2101) in which an OS of the edge node 120 is not stored and an LU (OS LU) 1100 (2100) in which an OS of the edge node 120 is stored as an LU. The LU is a logical storage device. The LU can be a substantive LU based on at least one DISK 113 (213) and can be a virtual LU that conforms to Thin Provisioning. The LU is configured by a plurality of blocks (storage regions), and a file is stored into the LU.

[0061] The memory 221 of the edge node 120 stores at least one computer program. The computer program will be described later in detail.

[0062] The memory 231 of the client 130 stores an application 1310, a file system program 1311, and a kernel/driver 1312.

[0063] The application 1310 is software (an application program) that is used according to a purpose of a work by the client 130. The descriptions of the file system program 1311 and the kernel/driver 1312 are omitted. This is because the file system program 1311 is almost equivalent to a file system program 1211 that is included in the edge node 120 and that is described later (see Fig. 4), and the kernel/driver 1312 is almost equivalent to a kernel/driver 1212 that is included in the edge node 120 and that is described later

(see Fig. 4).

- [0064] The memory 221 of the core node 220 stores a data mover program 2210, a file system program 2211, and a kernel/driver 2212. The descriptions of the data mover program 2210, the file system program 2211, and the kernel/driver 2212 are omitted. This is because the data mover program 2210 is almost equivalent to a data mover program 1210 that is included in the edge node 120 and that is described later (see Fig. 4), the file system program 2211 is almost equivalent to the file system program 1211 described above, and the kernel/driver 2212 is almost equivalent to the kernel/driver 1212 described above.
- [0065] In the following descriptions, the data mover program 1210 in the edge node 120 is referred to as an edge mover 1210, and the data mover program 2210 in the core node 220 described later is referred to as a core mover 2210. In the case in which the edge mover 1210 and the core mover 2210 are not distinguished in particular, the edge mover 1210 and the core mover 2210 are referred to as a data mover program. A file is transmitted and received between the edge node 120 and the core node 220 via the edge mover 1210 and the core mover 2210.
- [0066] The computer program and information that are included in the edge node 120 will be described in the following with reference to Fig. 4.
- [0067] The memory 121 of the edge node 120 stores an edge mover 1210, a file system program 1211, a kernel/driver 1212, a file sharing program 1213, a load relaxation control program 1214, a storage destination information sharing program 1215, a storage destination management table 1216, and a group management table 1217.
- [0068] The edge mover 1210 reads a file of a replication target from the edge storage region, transfers the file to the core node 220, and creates a stub in which the file is a link destination. The core mover 2210 receives the file of a replication target from the edge node 120, and writes the file to the core storage region.
- [0069] In the case in which the edge mover 1210 receives a read request from the client 130 to the stub (metadata), the edge mover 1210 transmits a read request that specifies a file of a link destination of the stub to the core mover 2210. By this configuration, the file in the core storage region is read from the core mover 2210 to the edge mover 1210.
- [0070] The file system program 1211 manages a file system. The file system program 1211 is provided with a reception program 12110. The reception program 12110 reads a file or writes a file in accordance with an I/O request (a write request or a read request) from the client 130.
- [0071] The kernel/driver 1212 executes the general control and a control unique to hardware such as a schedule control of a plurality of programs (processes) that are operated on the edge node 120 and a handling of an interrupt from hardware.

- [0072] The file sharing program 1213 provides a file sharing service to the client 130 by using a predetermined communication protocol (such as a CIFS (Common Internet File System) and an NFS (Network File System)).
- [0073] The load relaxation control program 1214 executes a processing for relaxing a load of the core node 220 (more specifically, an edge group creation processing described later).
- [0074] The storage destination information sharing program 1215 shares a storage destination management table between the RO edge nodes 120B.
- [0075] The storage destination management table 1216 is provided with the information that indicates a storage destination of a file that has been cached in the edge storage region (a location in the core storage region).
- [0076] The group management table 1217 is provided with the information related to an edge group that has been created by the load relaxation control program 1214. The "edge group" is an aggregate of at least one RO edge node 120B.
- [0077] An example of a file system will be described in the following with reference to Figs. 5 to 8.
- [0078] The file system 500 is provided with a super block 501, an inode management table 502, and a plurality of data blocks 503 as shown in Fig. 5.
- [0079] The super block 501 holds the information related to a file system. The information includes information related to each of files in the file system (such as the information that indicates a size and a location of a file) for instance.
- [0080] The inode management table 502 is provided with information related to an inode that is corresponded to each individual directory or file. Fig. 6 is a view showing an example of a part that is provided with information related to an inode that is corresponded to a directory in an inode management table. Fig. 7 is a view showing an example of a part that is provided with information related to an inode that is corresponded to a file in an inode management table.
- [0081] The data blocks 503 stores data that configures a file.
- [0082] In accordance with an example shown in Figs. 6 to 8, a file "a.txt" is read as described in the following. In other words, a path of /home/user-01/a.txt is gone through based on the information shown in Fig. 6. More specifically, the data block 100 is accessed according to the flow of "/" (inode number 2) -> "home" (inode number "10") -> "user-01" (inode number "15") -> "a.txt" (inode number 100). The following data (8-1) to (8-3) are obtained as shown in Fig. 8 based on the information shown in Fig. 7.

(8-1) data in three consecutive blocks in which a block of an address 100 is beginning;

(8-2) data in two consecutive blocks in which a block of an address 200 is beginning;

and

(8-3) data in five consecutive blocks in which a block of an address 100 is beginning. A file that includes data (8-1) to (8-3) is a file "a.txt".

[0083] Fig. 9A is a view showing a configuration of a storage destination management table 1216.

[0084] The storage destination management table 1216 associates and stores a file storage destination hash value 12161, a file entity hash value 12162, and the time and date of recall 12163 for every file that has been cached in the edge storage region that is included in the RO edge node 120B that is provided with the table 1216.

[0085] The file storage destination hash value 12161 is a hash value of the information that indicates a storage destination in the core storage region (a location of the core storage region for a file that has been cached). The storage destination of a file is indicated by the URL for instance.

[0086] The file entity hash value 12162 is a hash value of a file that has been cached.

[0087] The time and date of recall 12163 is the information that indicates the latest date when a recall request of a file was transmitted to the core node 220 (the latest date when a recall of a file has been executed). The recall request is a read request that specifies a file of a link destination of a stub.

[0088] Fig. 9B is a view showing a group management table 1217.

[0089] The group management table 1217 is provided with a group ID 12171, a group configuration 12172, and a representative flag 12173 for every edge group.

[0090] The group ID is an ID of an edge group.

[0091] The group configuration 12172 includes an ID of each of the RO edge nodes 120B that configure an edge group.

[0092] The representative flag 12173 is a flag that indicates whether each of the RO edge nodes 120B that configure an edge group is a representative node or not. The RO edge node 120B that is corresponded to the representative flag 12173 of "1" is a representative node of an edge group that includes the node 120B. The RO edge node 120B that is corresponded to the representative flag 12173 of "0" is a non-representative node of an edge group that includes the node 120B.

[0093] Fig. 9B indicates the following matters for instance:

(*) a group in which the group ID 12171 is "01" is configured by the RO edge nodes 120B1, 120B2, and 120B3; and

(*) the representative node of the group is the RO edge node 120B1, and other RO edge nodes 120B2 and 120B3 are non-representative nodes.

[0094] The flow of a processing that is executed in the present embodiment will be described in the following.

[0095] Fig. 10 is a view showing a flow of a processing in which a plurality of RO edge

nodes 120B share the storage destination management table 1216.

[0096] For instance, one RW edge node 120A and three RO edge nodes 120B1 to 120B3 are coupled to one core node 220. The clients 130A and 130B1 to 130B3 are coupled to the RW edge node 120A and the RO edge nodes 120B1 to 120B3, respectively.

[0097] Each of the storage destination management tables 1216 is shared by a plurality of RO edge nodes 120B according to the storage destination information sharing program 1215 that is included in each of the RO edge nodes 120B (see Fig. 4). For instance, the information that is included in the storage destination management table 1216 in the RO edge node 120B1 is copied to all of other the RO edge nodes 120B2 and 120B3. By this configuration, the RO edge nodes 120B2 and 120B3 are provided with the storage destination management table 1216B1 (substantially, a duplicate of the storage destination management table 1216B1 in the RO edge node 120B1).

[0098] The similar processing is also executed for the storage destination management table 1216B2 (1216B3) in the RO edge node 120B2 (120B3). By this configuration, each of the RO edge nodes 120B can be provided with not only the storage destination management table 1216 for a file that has been stored into an edge storage region (a file system) that is included in the node 120B but also the storage destination management table 1216 for a file that has been stored into an edge storage region that is included in other node 120B.

[0099] It is assumed that all storage destination management tables 1216 are shared by all RO edge nodes 120B. Moreover, it is assumed that a file in a core storage region (hereafter referred to as a file X1) is cached in all edge storage regions.

[0100] In this state, it is assumed that a write request is transmitted from the client 130A to the RW edge node 120A, the file X1 that has been stored into an edge storage region of the RW edge node 120A is updated, and the file after the update (hereafter referred to as a file X2) is stored into a core storage region for instance. Moreover, the RO edge node 120B1 executes a recall of the file X2 after the update from the core storage region, caches the file X2 after the update to an edge storage region that is included in the node 120B1, and updates the storage destination management table 1216B1 that is included in the node 120B1.

[0101] In this stage, although the file X2 after the update has been cached in the edge storage region that is included in the RO edge node 120B1, the file X1 before an update has been cached in the edge storage region that is included in the RO edge nodes 120B2 and 120B3 and the file X2 after the update has not been cached. In other words, the storage destination management tables 1216B2 and 1216B3 have not been updated.

[0102] The RO edge node 120B transmits the information that is included in the storage destination management table 1216 that is corresponded to the node 120B to other RO

edge node 120B. The information to be transmitted can be all information that is included in the storage destination management table 1216, or can be only information on the line that has been updated by an execution of a recall of a file.

[0103] The RO edge node 120B receives the information that is included in the storage destination management table 1216 that is corresponded to other edge node 120B from the other RO edge node 120B. The RO edge node 120B updates the storage destination management table 1216 that is corresponded to the other edge node 120B.

[0104] The RO edge node 120B compares the storage destination management table 1216 that is corresponded to the node 120B with the storage destination management table 1216 that is corresponded to other node 120B. Based on the result of the comparison, the RO edge node 120B can specify whether a file that has been cached is a file after an update or a file before an update.

[0105] As described with reference to Fig. 11 for instance, the RO edge node 120B2 can specify a file in which a file entity hash value 12162 is xxxyyy123 is a file before an update among three files that have been cached by the RO edge node 120B2 by comparing tables 1216B1 and 1216B2 that are included in the RO edge node 120B2. More specifically, the RO edge node 120B2 executes the following processing for instance:

(a) selecting at least two files that are provided with the same file storage destination hash value 12161;

(b) comparing the file entity hash values 12162 with each other for the at least two files that have been selected in (a);

(c) comparing the time and date of recall 12163 for the at least two files that have been selected in (a) in the case in which the file entity hash values 12162 are different from each other; and

(d) specifying a file in which the time and date of recall 12163 are older as a file before an update.

Consequently, in accordance with the example shown in Fig. 11, it is specified that a file in which a file entity hash value 12162 is xxxyyy123 is a file before an update.

[0106] The RO edge node 120B2 creates a stub for only a file before an update and does not create a stub for a file after an update. A file in which a stub is created is limited to a file before an update among a plurality of files that have been cached. Consequently, the number of accesses to the core node 220 can be suppressed. In other words, in the case in which the RO edge node 120B2 receives a read request that specifies a stub, the RO edge node 120B2 transmits a recall request (a read request) of a file of a link destination of the stub to the core node 220. However, since a file in which a stub is created is limited to a file before an update, the number of transmissions of a recall request to the core node 220 can be suppressed.

- [0107] Moreover, in the case in which the RO edge node 120B receives a read request that specifies a file that has been cached from the client 130B2, the RO edge node 120B transmits the file that has been cached to the client 130B2. However, since it is expected that the file that has been cached is only a file after an update, it can be expected that a file before an update is not returned to the client 130B2.
- [0108] In accordance with one ingenuity for suppressing the number of accesses to the core node 220 (hereafter referred to as an ingenuity 1), a file before an update is specified among at least one file that has been cached, and a stub is created for only a file after the update.
- [0109] The present embodiment is also provided with another ingenuity 2 for suppressing the number of accesses to the core node 220. In accordance with the ingenuity 2, at least one edge group that is configured by at least one RO edge node 120B is created. Even in the case in which the RO edge node 120B receives a read request that specifies a stub, it is expected that the RO edge node 120B does not transmit a recall request (a read request) to the core node 220, obtains a file after an update (a latest file) that is corresponded to the stub, and returns the file after an update. More specifically, the RO edge node 120B can judge whether or not a file after an update that is corresponded to the stub that has been specified by a read request has already been cached to an edge storage region that is included in other node 120 in an edge group that includes the node 120B based on the storage destination management table 1216 that is corresponded to other node 120B. In the case in which a core storage destination (a storage destination in a core storage region) that is corresponded to the stub is equivalent to a core storage destination of a file that has been cached in an edge storage region that is included in other node 120B, it is decided that a file after an update that is corresponded to the stub has been cached in an edge storage region that is included in other node 120B in an edge group. In the case in which the result of the judgment is positive, the RO edge node 120B can obtain the file from other node 120B and return the file to the client 130B.
- [0110] In other words, in the case in which a file that has been received from the client 130B by the RO edge node 120B and that has been specified by a read request has been cached in an edge group that includes the node 120B, a recall request is not issued from the node 120B, and the file has been cached is transmitted to the client 130B.
- [0111] Consequently, in the case in which many files have been cached in one edge group as much as possible and duplicated files are included in the files as less as possible, it is thought that a probability of an issuance of a recall request is low.
- [0112] An edge group is created according to such an aspect. The creation processing of an edge group will be described in the following.
- [0113] In the creation processing of an edge group, the following processing is executed:

- (a) a selection of a representative node; and
- (b) an allocation for every non-representative node.

By the above configuration, at least one edge group is created. The edge group is configured by one representative node and at least zero non-representative node.

[0114] The selection of a representative node is executed as described in the following.

[0115] The RO edge node 120B (the load relaxation control program 1214) refers to all the storage destination management tables 1216 in the node 120B, and specifies an RO edge node 120B that satisfies both (or one) of the following two conditions:

(Condition 1) an amount of cache is equal to or larger than a certain value; and

(Condition 2) an amount of duplicated files is equal to or larger than a certain value.

The specified node 120B is selected as a representative node.

[0116] The amount of cache described above can be the total amount of at least one file that has already been cached (for instance, a total size, a total number, or a rate (a total size of at least one file that has already been cached to a capacity of a whole or a predetermined part of an edge storage region (a file system))).

[0117] The amount of duplicated files described above is a file that is corresponded to the following (A) and (B) among at least one file that has been cached in an edge storage region that is included in the node 120B (it can be specified whether a storage destination is equivalent to each other or not based on the storage destination management table 1216 of each of the RO edge nodes 120B).

(A) a core storage destination is equivalent to a core storage destination of a file that has been cached in an edge storage region that is included in other RO edge node 120B; and

(B) duplicated by a file that has been cached in an edge storage region that is included in other RO edge node 120B (the content is the same).

Consequently, in the case in which a file that has been cached for a node 120B1 and a file that has been cached for a node 120B2 are duplicated by each other for instance, it is said that a duplicated file exists for the node 120B1 and the node 120B2.

[0118] The amount of duplicated files described above can be the total amount of at least one duplicated file (for instance, a total size, a total number, or a rate to an amount of cache (a total number of at least one duplicated file to a total number of at least one file that has already been cached)).

[0119] The Condition 1 and Condition 2 described above conform to the above described aspect.

[0120] More specifically, as described above, it is preferable that many files have already been cached in an edge group as much as possible. However, there is a possibility that an RO edge node 120B other than a representative node does not exist in the edge group. Consequently, it is preferable that an RO edge node 120B that is provided with

a larger amount of cache is selected as a representative node. Therefore, the Condition 1 is adopted as one condition of a representative node.

[0121] Moreover, as described above, it is preferable that there are many files that are provided with different contents (duplicated files are included as less as possible). An edge group exists for every representative node. Consequently, it is preferable that at least two RO edge nodes 120B that are provided with many duplicated files exist in separate edge groups. Therefore, the Condition 2 is adopted as one condition of a representative node.

[0122] An amount of cache and an amount of duplicated files can be calculated based on the storage destination management table 1216.

[0123] For instance, an amount of cache can be the number of files that have already been cached, which is specified by the storage destination management table 1216. In the example shown in Fig. 11, for the RO edge node 120B1, it is found that the number of files that have already been cached is 3 from the storage destination management table 1216B1. Consequently, an amount of cache can be 3 for the RO edge node 120B1.

[0124] The duplicated file is one of at least two files in which a combination of a file storage destination hash value 12161 and a file entity hash value 12162 conforms to each other. In other words, in the example shown in Fig. 11, there is a file that is corresponded to a combination of a file storage destination hash value 12161 aaabbb123 and a file entity hash value 12162 xxxyyy456 (hereafter referred to as a file K1) as a file that has already been cached for the node 120B1. It is also found that there is a file that is corresponded to a combination of a file storage destination hash value 12161 aaabbb123 and a file entity hash value 12162 xxxyyy456 (hereafter referred to as a file K2) as a file that has already been cached for the node 120B2. In this case, the file K1 is a duplicated file for the node 120B1, and the file K2 is a duplicated file for the node 120B2.

[0125] After a representative node is selected as described above, a non-representative node (that is, an RO edge node that has not been selected as a representative node) is allocated to any of edge groups. An edge group that is an allocated destination is an edge group in which there are least cached files that are duplicated with a file that has already been cached for the non-representative node among at least two edge groups.

[0126] In the case in which an allocation for all non-representative nodes is completed, the creation processing of an edge group is terminated.

[0127] A concrete example of the creation processing of an edge group will be described in the following.

[0128] The example shown in Fig. 12A indicates the following statuses:

- (*) the RO edge node 120B1 is provided with files A to H;
- (*) the RO edge node 120B4 is provided with files A to D and J to M;

- (*) the RO edge node 120B2 is provided with files L and M;
- (*) the RO edge node 120B3 is provided with files D and J; and
- (*) the RO edge node 120B5 is provided with files G and H.

- [0129] In accordance with the Condition 1 and Condition 2 described above, the nodes 120B1 and 120B4 of the nodes 120B1 to 120B5 are representative nodes. The nodes 120B1 and 120B4 of the nodes 120B1 to 120B5 are provided with the largest files that have already been cached and the largest duplicated files A to D.
- [0130] An edge group 01 that includes a representative node 120B1 and an edge group 04 that includes a representative node 120B4 are prepared. The RO edge nodes 120B2, 120B3, and 120B5 other than a representative node are non-representative nodes. Each of the non-representative nodes is allocated to an edge group in which there are least cached files that are duplicated with a file that has already been cached for the non-representative node. As a result, as shown in Fig. 12A, the non-representative nodes 120B2 and 120B3 are allocated to the edge group 01, and the non-representative node 120B5 is allocated to the edge group 02.
- [0131] As a result, as shown in Fig. 12B, the edge group 01 is configured by the representative node 120B1 and the non-representative nodes 120B2 and 120B3. Moreover, the edge group 02 is configured by the representative node 120B4 and the non-representative node 120B5.
- [0132] Consequently, as shown in Fig. 12C, files that have already been cached for the edge group 01 are files A to H, L, M, and J, and files that have already been cached for the edge group 02 are files A to D, J to K, L, G, and H.
- [0133] In other words, since an edge group is created from an aspect described above, there is a low possibility that a plurality of duplicated files is concentrated to one edge group, and there is a high possibility that a plurality of duplicated files is distributed to a plurality of edge groups. As shown in Fig. 12C, the duplicated files A to D, G, H, J, L, and M exist in both of edge groups 01 and 02. Consequently, in the case in which a node 120B that belongs to any of the edge groups 01 and 02 receives a read request that specifies a stub of any one of the duplicated files, a file that has already been cached that is corresponded to the stub can be obtained from other node 120B in the edge group 01 or 02 that includes the node 120B. Therefore, a recall request is not necessary.
- [0134] However, in the case in which an edge group is created without conforming to the above aspect, there is a high probability that a recall request is necessary unfortunately.
- [0135] For instance, as shown in Fig. 13A, it is assumed that an edge group 03 that is configured by nodes 120B2, 120B3, and 120B5 and an edge group 04 that is configured by nodes 120B1 and 120B4 are prepared.
- [0136] In this case, as shown in Fig. 13B, duplicated files A to D that have already been

cached for the node 120B1 and duplicated files A to D that have already been cached for the node 120B4 exist as a file that has already been cached for the same edge group 04, and files A to D do not exist as a file that has already been cached for other edge group 03. In other words, a plurality of duplicated files is not distributed to a plurality of edge groups in a uniform manner and is placed in a disproportionate manner.

[0137] Consequently, even in the case in which any node 120B in the edge group 03 receives a read request that specifies a stub of any duplicated file (such as a file A), since there is not a file A as a file that has already been cached in the edge group 03, it is necessary that a recall request is issued.

[0138] As described above, in the case in which an edge group is created without conforming to the above aspect, there is a high probability that a recall request is issued unfortunately.

[0139] As a result, creating an edge group according to the above described aspect is the ingenuity 2 in accordance with the present embodiment. By this configuration, it is possible to reduce the number of times of an issuance of a recall request. Although both of the ingenuity 1 and the ingenuity 2 is adopted in the present embodiment, it is also possible to reduce the number of times of the accesses to a core node even in the case in which any one of the ingenuity 1 and the ingenuity 2 is adopted.

[0140] Fig. 14 is a view showing an example of a processing that is executed in the case in which an RO edge node 120B receives a read request that specifies a stub.

[0141] The edge group 01 is configured by the representative node 120B1 and the non-representative nodes 120B2 and 120B3.

[0142] It is assumed that an RO edge node 120B3 receives a read request that specifies a stub of a file A from a client 130B3 (S1401).

[0143] In this case, the RO edge node 120B3 judges whether or not the file A that is corresponded to the stub has already been cached to an edge group 01 that includes the node 120B3 based on all of the storage destination management tables 1216 that are included in the node 120B3 (S1402).

[0144] In the case in which the result of a judgment of the S1402 is positive, the RO edge node 120B3 obtains the file A that has already been cached from an edge storage region of other node 120B (S1403), and transmits the file A to the client 130B3 (S1405).

[0145] In the case in which the RO edge node 120B3 specifies that a plurality of files A that have already been cached that are corresponded to the stub that is specified by the read request exist in the same edge group, the RO edge node 120B3 selects one file A that has already been cached from the plurality of files A that have already been cached. The file A that is selected can be a file that conforms to a predetermined policy. The policy conforms to at least one aspect of (a) a distance from the node 120B3, (b) an

access load of other node 120B, and (c) the latest date when other node 120B executes a recall for other node 120B in the edge group 01 for instance. More specifically, the file that is selected is a file that is corresponded to at least one of the following (A) to (C):

(A) a file that has been stored into an edge storage region of other node 120B in which a distance from the node 120B3 is shortest;

(B) a file that has been stored into an edge storage region of a node 120B in which an access load is lowest among at least one other node 120B in the edge group 01; and

(C) a file in which the time and date of recall 12163 is latest (see Fig. 9).

[0146] The distance from the node 120B3 can be the number of hops, that is, relay apparatuses that are disposed between the node 120B3 and the other node 120B (such as a router and a switch device that is provided with a function similar to the router). Each of the nodes 120B can be provided with the information that indicates a distance from each of other nodes 120B in an edge group that includes the node 120B. The distance from the node 120B3 can be specified by the information.

[0147] The access load of a node 120B can be the number of recalls that have been executed by the node 120B (or a CPU usage rate of the node 120B) for instance.

[0148] In the case in which there are at least two files that are corresponded to at least one of the above (A) to (C), a file that satisfies the above (A) to (C) as many as possible can be selected. Moreover, in the case in which there are at least two files that satisfies separate conditions of the above (A) to (C), a file can be selected based on each of the priority levels of the above (A) to (C).

[0149] In the case in which the result of a judgment of the S1402 is negative, the RO edge node 120B3 obtains the file A from the RW edge node 120A, and caches the file A in an edge storage region of the node 120B3 (S1404). The node 120B3 then transmits the file A to the client 130B3 (S1405).

[0150] The reason why the node 120B3 accesses the RW edge node 120A in the case in which a file that has already been cached and that is corresponded to a stub does not exist in the edge group 01 is in the following. This is because there is a possibility that a file after an update for the RW edge node 120A (a file after an update (the latest file) that is corresponded to a stub and that exists in an edge storage region of the RW edge node 120A) is not reflected to a core node 220 (is not migrated to a core node 220) in the case in which a file that has already been cached and that is corresponded to a stub does not exist in the edge group 01. In other words, it is thought that there is a high possibility that a file after an update (the latest file) that is corresponded to a stub can be obtained in the case in which the node 120B3 transmits a read request of a file that is corresponded to a stub to not the core node 220 but the RW edge node 120A in this case. Moreover in this case, since a read request is transmitted to not the core node 220

but the RW edge node 120A, a load of the core node 220 can be reduced.

[0151] The processing that is executed by each of the programs will be described in the following. In the following descriptions, an edge mover that is executed by the RW edge node 120A is referred to as an RW mover, and an edge mover that is executed by the RO edge node 120B is referred to as an RO mover.

[0152] Fig. 15 is a part of a flowchart showing a flow of a processing of an RW mover 1210. Fig. 16 is a remaining part of the flowchart of Fig. 15. In the descriptions of Fig. 15 and Fig. 16, the RW edge node 120A that executes the RW mover 1210 is referred to as a target node 120A, and an edge storage region that is included in the target node 120A is referred to as an RW edge storage region.

[0153] The RW mover 1210 checks whether or not an event has occurred (S1501). The event means any one of an elapse of a certain period of time, a replication request, a synchronization request, and a recall request.

[0154] The elapse of a certain period of time means an elapse of a certain period of time after the S1501 is terminated.

[0155] The replication request means a request of a replication of a file from the RW edge storage region to the core storage region.

[0156] The synchronization request means a request of a synchronization of a file to which a replication from the RW edge storage region to the core storage region has been executed (a file in the core storage region) and a file that has been updated by the RW edge node 120A (that is, a request of writing a file after an update in the RW edge storage region to a file in the core storage region).

[0157] The recall request means a request of reading a file from the core storage region as described before.

[0158] The RW mover 1210 can receive at least one of a replication request, a synchronization request, and a recall request from an apparatus other than the RW edge node 120A (for instance, a client), and can also receive that from other program in the RW edge node 120A.

[0159] In the next place, in the case in which the RW mover 1210 decides that an event has not occurred (S1502: No), S1501 is executed.

[0160] In the case in which the RW mover 1210 decides that an event has occurred (S1502: Yes), the RW mover 1210 judges a type of the event that has occurred (S1503).

[0161] In the case in which a type of the event is an elapse of a certain period of time, the RW mover 1210 checks the remaining capacity of an LU 1101 (an RW edge storage region (file system)) (S1504). The remaining capacity of an LU 1101 is a free capacity (a capacity that can be used) of an LU 1101 that has been provided to the target node 120A.

[0162] In the case in which the remaining capacity of an LU 1101 is larger than a threshold

value (S1508: No), the RW mover 1210 terminates the processing.

- [0163] In the case in which the remaining capacity of an LU 1101 is equal to or less than a threshold value (S1508: Yes), the RW mover 1210 selects files of older final access date in order until the remaining capacity of an LU 1101 (a file system) exceeds a threshold value (S1509).
- [0164] The RW mover 1210 deletes a file and creates the stub of the deleted file for all files that have been selected in the S1509 (S1510).
- [0165] In the case in which a type of the event is a replication request, the RW mover 1210 obtains the information that indicates a storage destination of a file of a replication target from the core node 220 (S1505).
- [0166] The RW mover 1210 sets the information of the storage destination that has been obtained in the S1505 to the metadata of a file of a replication target (S1511).
- [0167] The RW mover 1210 obtains an object (a file or a directory) that is specified by a replication request and the metadata of the object from the LU 1101 (an RW edge storage region) (S1512).
- [0168] The RW mover 1210 transfers the object and the data that have been obtained in the S1512 to the core node 220 (S1513). By this step, the object (a file or a directory) that has been obtained in the S1512 is stored into the core storage region.
- [0169] In the case in which a type of the event is a synchronization request or a recall request, the RW mover 1210 checks whether the type of the event is a synchronization request or a recall request (S1506).
- [0170] In the case in which a type of the event is a synchronization request, the RW mover 1210 obtains an object that has been updated after the replication and its metadata from the LU 1101 (an RW edge storage region) (S1514).
- [0171] The RW mover 1210 transfers the object that has been obtained and its metadata to the core node 220 (S1515). By this step, a file that has been updated by the RW edge node 120A (a file in the RW edge storage region) is reflected to the core storage region.
- [0172] In the case in which a type of the event is a recall request, the RW mover 1210 checks reads a file that is linked to a stub from the core node 220 (the core storage region) and caches the file to the RW edge storage region (S1507).
- [0173] Fig. 17 is a flowchart showing a flow of a processing of an RO mover 1210. In the descriptions of Fig. 17, the RO edge node 120B that executes the RO mover 1210 is referred to as a target node 120B, and an edge storage region that is included in the target node 120B is referred to as an RO storage region.
- [0174] The RO mover 1210 checks whether or not an event has occurred (S1701). The event means any one of a first elapse of a certain period of time, a second elapse of a certain period of time, receiving the effect that an update has occurred from the core node, and

a recall request.

- [0175] The first elapse of a certain period of time means an elapse of a certain period of time after the S1701 is terminated.
- [0176] The second elapse of a certain period of time means an elapse of a certain period of time after the S1708 is terminated.
- [0177] Receiving the effect that an update has occurred from the core node means a reception of the information that indicates that a file that has been stored into the core storage region and that is corresponded to a stub that is included in the target node 120B has been updated.
- [0178] In the case in which an event has not occurred (S1702: No), the RO mover 1210 executes the S1701.
- [0179] In the case in which an event has occurred (S1702: Yes), the RO mover 1210 judges a type of the event that has occurred (S1703).
- [0180] In the case in which a type of the event is a first elapse of a certain period of time, the RO mover 1210 checks the remaining capacity of an LU 1101 (an RO edge storage region (file system)) (S1704).
- [0181] In the case in which the remaining capacity of an LU 1101 is larger than a threshold value (S1705: No), the RO mover 1210 terminates the processing.
- [0182] In the case in which the remaining capacity of an LU 1101 is equal to or less than a threshold value (S1705: Yes), the RO mover 1210 selects files of older final access date in order until the remaining capacity of an LU 1101 exceeds a threshold value (S1706).
- [0183] The RO mover 1210 deletes a file and creates the stub of a file for all files that have been selected in the S1706 (S1707).
- [0184] In the case in which a type of the event is a first elapse of a certain period of time or receiving the effect that an update has occurred from the core node, the RO mover 1210 executes the following processing (S1708):
- (S1708-1) judging whether or not a file before an update exists in at least one file that has already been cached in the RO edge storage region based on all of the storage destination management tables 1216 that are included in the target node 120B; and
- (S1708-2) deleting the file before an update and creating a stub that is corresponded to a file after an update that has been stored into the core storage region in the case in which the result of a judgment of the S1708-1 is positive.
- [0185] In the case in which a type of the event is a recall request, the RO mover 1210 reads a file that is linked to a stub from the core node 220 (the core storage region) and caches the file to the RW edge storage region (S1709). After the file is cached, the RO mover 1210 updates the file entity hash value 12162 and the time and date of recall 12163 that are corresponded to the file that has been cached for the storage destination

management table 1216 that is corresponded to the target node 120B. The information that includes the file entity hash value 12162 and the time and date of recall 12163 after the update is transmitted to all other RO edge nodes 120B.

[0186] Fig. 18 is a flowchart showing a flow of a processing that is executed by a reception program 22110 of an RW edge node 120A. In the descriptions of Fig. 18, the reception program is referred to as an RW reception program, the RW edge node 120A is referred to as a target node 120A, and an edge storage region that is included in the target node 120A is referred to as an RW storage region.

[0187] The RW reception program 22110 receives an I/O request from the client 130A and judges a type of the I/O request (S1801).

[0188] In the case in which an I/O request is a write request, the RW reception program 22110 executes the following processing (S1802):

(S1802-1) making the edge mover 1210 of the target node 120A to execute a recall of a file that is corresponded to the file accompanying the write request from the core storage region;

(S1802-2) storing the file that has been obtained in the S1802-1 into the RW edge storage region; and

(S1802-3) writing the file accompanying the write request to the file that has been stored into the RW edge storage region in the S1802-2 (updating the file).

At this point of time, the file after the update in the S1802 (the file that has been written in accordance with the write request) has not been written to the core storage region. Consequently at this point of time, the latest file exists in not the core storage region but the RW edge storage region only. In this state, there is not a file that has already been cached of a read target in an edge group as described above, the latest file is read from the target node 120A by transmitting a read request of the file to the target node 120A.

[0189] In the case in which an I/O request is a read request, the RW reception program 22110 judges whether or not a file that is specified by the read request has already been recalled (cached) (S1803).

[0190] In the case in which the result of a judgment of the S1803 is positive (S1803: Yes), the RW reception program 22110 returns a file that has been recalled (a file that has already been cached) to the client 130A (S1805).

[0191] In the case in which the result of a judgment of the S1803 is negative (S1803: No), the RW reception program 22110 executes the following processing (S1804):

(S1804-1) making the edge mover 1210 of the target node 120A to execute a recall of a file that is specified by the read request (a file that is corresponded to the stub that is specified by the read request) from the core storage region; and

(S1804-2) storing the file that has been obtained in the S1804-1 into the RW edge

storage region.

The RW reception program 22110 returns a file that has been obtained in the S1804 to the client 130A (S1805).

- [0192] Fig. 19 is a part of a flowchart showing a flow of a processing that is executed by a reception program 12110 of an RO edge node 120B. Fig. 20 is a remaining part of the flowchart of Fig. 19. In the descriptions of Fig. 19 and Fig. 20, the reception program is referred to as an RO reception program, the RO edge node 120B that executes the reception program is referred to as a target node 120B, an edge storage region that is included in the target node 120B is referred to as an RO edge storage region, and an edge group that includes the target node 120B is referred to as a target group.
- [0193] In the case in which the RO reception program 22110 receives a read request from the client 130, the RO reception program 22110 judges whether or not a file that is specified by the read request has already been recalled (cached) (S1901).
- [0194] In the case in which the result of a judgment of the S1901 is positive (S1902: Yes), the RO reception program 22110 returns a file that has already been recalled (a file that has already been cached) to the client 130B (S1905).
- [0195] In the case in which the result of a judgment of the S1901 is negative (S1902: No), the RO reception program 22110 judges whether or not there is a target group based on a group configuration table 1217 in the target node 120B (S1903).
- [0196] In the case in which the result of a judgment of the S1903 is negative (S1904: No), the RO reception program 22110 executes the following processing (S1905):
- (S1905-1) making the edge mover 1210 of the target node 120B to execute a recall of a file that is specified by the read request (a file that is corresponded to the stub that is specified by the read request) from the core storage region;
 - (S1905-2) caching the file that has been obtained in the S1905-1 (storing the file that has been obtained in the S1905-1 into the RO edge storage region); and
 - (S1905-3) registering the information related to the file that has been cached in the S1905-2 to the storage destination management table 1216 that is corresponded to the target node 120B.
- [0197] In the case in which the result of a judgment of the S1903 is positive (S1904: Yes), the RO reception program 22110 judges whether or not a file that is corresponded to the stub that is specified by the read request has already been cached in the target group (S1906). More specifically, the RO reception program 22110 executes the following processing:
- (S1906-1) specifying all other RO edge nodes 120B that are included in an edge group that includes the target node 120B based on the group management table 1217; and
 - (S1906-2) judging whether or not a file that is corresponded to the stub that is

specified by the read request exists based on the storage destination management table 1216 that is corresponded to the all other RO edge nodes 120B that have been specified in the S1906-1 among all the storage destination management tables 1216 that are included in the target node 120B.

[0198] In the case in which the result of a judgment of the S1906 is positive (S1907: Yes), the RO reception program 22110 executes the following processing (S1908):

(S1908-1) obtaining a file that is corresponded to the stub from other RO edge node 120B in the target group;

(S1908-2) caching the file that has been obtained in the S1908-1 (storing the file that has been obtained in the S1908-1 into the RO edge storage region); and

(S1908-3) updating the information that is corresponded to the file that has been cached in the S1908-2 in the storage destination management table 1216 that is corresponded to the target node 120B.

In the case in which at least two files that are corresponded to the stub that is specified by the read request have already been cached in the target group, the RO reception program 22110 obtains a predetermined policy, for instance, a file that is corresponded to at least one of the following three files:

(*) a file in which the time and date of recall 12163 is latest;

(*) a file in which other node 120B in which a distance from the target node 120B is shortest (for instance, an RO edge node in which the number of hops described above is least) has already been cached; and

(*) a file in which an RO edge node 120B in which the number of times of recall is least has already been cached.

[0199] In the case in which the result of a judgment of the S1906 is negative (S1907: No), the RO reception program 22110 executes the following processing (S1909):

(S1909-1) transmitting a read request of a file of a read target to the RW edge node 120A (that is, obtaining a file of a read target from the RW edge node 120A);

(S1909-2) caching the file that has been obtained in the S1909-1 (storing the file that has been obtained in the S1909-1 into the RO edge storage region); and

(S1909-3) updating the information that is corresponded to the file that has been cached in the S1909-2 in the storage destination management table 1216 that is corresponded to the target node 120B.

[0200] The RO reception program 22110 returns a file of a read target that has been cached in the RO edge storage region to the client 130B (S1910).

[0201] Fig. 21 is a flowchart showing a flow of a processing that is executed by a storage destination information sharing program 1215. In the descriptions of Fig. 21, the RO edge node 120B that executes the program 1215 is referred to as a target node 120B.

[0202] The storage destination information sharing program 1215 waits till an event occurs

(S2101). The event means any one of a "recall processing", an "advent of a transmission time", and "receiving the storage destination information from other RO edge node".

[0203] The program 1215 checks whether or not an event has occurred (S2102). In the case in which an event has not occurred (S2102: No), the program 1215 executes the S2101.

[0204] In the case in which a recall processing by other program has not occurred (S2103: Yes), the program 1215 updates a line of the storage destination management table 1216 that is corresponded to the file that has been recalled and that has been cached in the edge storage region (S2106). In other words, although the program that has executed a recall updates the storage destination management table 1216, the program that has executed a recall can call the storage destination information sharing program 1215, and the program 1215 can update the storage destination management table 1216.

[0205] In the case in which the current time is a transmission time (S2103: No, S2104: Yes), the program 1215 specifies the storage destination information to be transferred (the information of the storage destination management table 1216) for every RO edge node 120B (S2107). More specifically, the program 1215 compares the final transmission date when the storage destination management information is transmitted to other RO edge node 120B with the time and date of recall 12163 of the storage destination management table 1216 for every file that has already been cached for instance. The information that is included in a line (a line of the storage destination management table 1216) that is corresponded to a file that has already been cached in which the time and date of recall 12163 is later than the final transmission date is the information that has not been transferred to other RO edge node 120B. The information that indicates the final transmission date for every file that has already been cached can be managed by the storage destination management table 1216, or can also be managed by a table other than the storage destination management table 1216.

[0206] The program 1215 transfers the information (the storage destination information) that has been specified in the S2107 to all other RO edge nodes 120B (S2106). By this configuration, the storage destination management table 1216 that is corresponded to the target node 120B is updated for all other RO edge nodes 120B.

[0207] In the case in which the storage destination information is received from other RO edge node 120B (S2103: No, S2104: No, S2105: Yes), the program 1215 writes the storage destination information that has been received to the storage destination management table 1216 that is corresponded to other RO edge node that has transmitted the information (S2109).

[0208] Fig. 22 is a flowchart showing a flow of a processing that is executed by a load relaxation control program 1214. In the descriptions of Fig. 22, the RO edge node 120B

that executes the program 1214 is referred to as a target node 120B, and an edge storage region that is included in the target node 120B is referred to as an RO storage region.

- [0209] The load relaxation control program 1214 judges whether or not there is an RO edge node 120B in which a recall frequency is high based on all of the storage destination management tables 1216 that are included in the target node 120B (S2201). This can be judged based on whether or not the number of lines of the storage destination management table 1216 is equal to or larger than a predetermined value. In the case in which the number of lines of the storage destination management table 1216 is equal to or larger than a predetermined value, it is judged that a recall frequency is high. On the other hand, in the case in which the number of lines of the storage destination management table 1216 is less than a predetermined value, it is judged that a recall frequency is low. A level of a recall frequency can also be judged as described in the following as substitute for or in addition to the number of lines of the storage destination management table 1216. That is, a count value that indicates the number of times of recall (a count value that is reset to an initial value (such as zero) for every certain period of time) is set, and a level of a recall frequency can be judged based on whether or not the count value is equal to or larger than a predetermined value.
- [0210] In the case in which the result of a judgment of the S2201 is negative (S2202: No), the program 1214 judges whether or not an edge group has already been configured based on the group management table 1217 (S2203). In the case in which an edge group has not been configured (S2203: No), the program 1214 terminates the processing. In the case in which an edge group has been configured (S2203: Yes), the program 1214 deletes all edge groups (or only an edge group that is configured by only at least one RO edge node in which a recall frequency is less than a predetermined value) (S2204). The deletion of an edge group means a deletion of the information that is included in the group management table 1217 for instance.
- [0211] In the case in which the result of a judgment of the S2201 is positive (S2202: Yes), the program 1214 executes an edge group creation processing that is configured by the S2205 to S2208.
- [0212] In other words, the program 1214 specifies at least two RO edge nodes 120B in which an amount of cache is equal to or larger than a certain value and an amount of duplicated files is equal to or larger than a certain value as a representative node based on all of the storage destination management tables 1216 that are included in the target node 120B (S2205).
- [0213] The program 1214 creates a group for every representative node that has been specified (S2206). For instance, the program 1214 registers the information related to a group for every representative node to the group management table 1217.

- [0214] The program 1214 allocates each of the non-representative nodes to the edge group in which files that have already been cached are duplicated with a file that has already been cached for the non-representative node as less as possible based on the storage destination management table 1216 of each of the representative nodes and the storage destination management table 1216 of at least one non-representative node (S2207).
- [0215] The program 1214 registers the result of the S2207 to the group management table 1217 (S2208).
- [0216] In accordance with the example shown in Fig. 22, in the case in which it is specified that there is an RO edge node 120B in which a recall frequency is high, an edge group is created. However, a reorganization of an edge group (deleting all of the edge groups and executing an edge group creation processing again) can also be executed for every certain period of time.

Embodiment 2

- [0217] An embodiment 2 in accordance with the present invention will be described in the following. In the descriptions, points that are different from the embodiment 1 will be described in large part, and the detailed descriptions of elements equivalent to those of the embodiment 1 are omitted or simplified.
- [0218] As shown in Fig. 23, in the embodiment 2, a management server 300 that are coupled to all of edge nodes 120 is disposed. The management server 300 is provided with a storage destination information sharing program 310.
- [0219] As shown in Fig. 24, the program 310 can obtain the information that is included in the storage destination management table 1216 that is included in each of the RO edge nodes 120B via the communication network 1. By this configuration, the management server 300 can be provided with all of the storage destination management tables 1216 (a duplicate of the table 1216). The table 1216 that has been obtained is stored into a storage resource (such as a memory) that is included in the management server 300.
- [0220] Each of the RO edge nodes 120B can access the management server 300, whereby the RO edge node 120B can refer to all of the storage destination management tables 1216. In the case in which a storage destination management table that is corresponded to the RO edge node 120B is updated for the node 120B, the RO edge node 120B transmits the information that is included in the table 1216 after the update to the management server 300 (in other words, the transmission destination of the storage destination information after the update is not other RO edge node 120B but the management server 300). By this configuration, the update of the table 1216 for the RO edge node 120B is reflected to the table 1216 that is included in the management server 300.
- [0221] Fig. 25 is a flowchart showing a flow of a processing that is executed by a storage destination information sharing program 1215 in accordance with an embodiment 2 of

the present invention.

[0222] The storage destination information sharing program 1215 judges whether or not an event has occurred (S2301).

[0223] In the case in which an event has occurred (S2302: Yes), the program 1215 judges whether or not the event that has occurred is a recall processing (S2303).

[0224] In the case in which the event that has occurred is a recall processing (S2303: Yes), the program 1215 transmits the storage destination information related to a file that has been recalled (a file that has been cached) (for instance, the information that includes a file storage destination hash value 12161, a file entity hash value 12162, and the time and date of recall 12163) to the management server 300 (S2304).

[0225] Fig. 26 is a flowchart showing a flow of a processing that is executed by a storage destination information sharing program 310 in a management server 300.

[0226] The storage destination information sharing program 310 judges whether or not an event has occurred (S2701).

[0227] In the case in which an event has occurred (S2702: Yes), the program 310 judges a type of the event that has occurred (S2703). The event is a "reference request" or an "update request". The requests are transmitted from the RO edge node 120B for instance.

[0228] In the case in which the event is a "reference request", the program 310 returns the information that is included in at least one storage destination management table 1216 that conforms to the reference request to the RO edge node 120B of the reference request source (S2704). The program 310 manages the final transmission date when the information that is included in the storage destination management table 1216 is transmitted to each of the RO edge nodes 120B. The program 310 can compare the final transmission date with the time and date of recall 12163, thereby deciding the information (the information in the storage destination management table 1216) to be transmitted to the RO edge node 120B that has issued the reference request.

[0229] In the case in which the event is an "update request", the program 310 registers the storage destination information accompanying the update request to the storage destination management table 1216 that is corresponded to the RO edge node 120B of the update request source (S2705).

[0230] While the preferred embodiments in accordance with the present invention have been described above, the present invention is not restricted to the embodiments.

[0231] For instance, an identifier of private or public can also be added to a file for at least one of the embodiments 1 and 2. In the case in which an identifier of public is added to a file that has been cached, the RO edge node 120B registers the information related to the file to the storage destination management table 1216. In the case in which an identifier of private is added to a file that has been cached, it is necessary that the RO

edge node 120B register the information related to the file to the storage destination management table 1216.

- [0232] Moreover for instance, one of a plurality of programs that are executed by an edge node (for instance, the load relaxation control program 1214) can also be executed by the management server 300. More specifically, the management server 300 can execute the edge group creation processing for instance.
- [0233] Moreover for instance, one virtual server system can also be configured by at least two edge nodes 120. That is, the edge nodes 120 can be handled as one server system by a client.
- [0234] Moreover for instance, it is possible that each of the RO edge nodes 120B does not receive a write request from the client 130B. In the case in which an RO edge node 120B receives a write request, the write request can be transferred to the RW edge node 120A.
- [0235] Moreover for instance, an edge node that transmits a write request to the core node 220 and that does not transmit a read request to the core node 220 can also be adopted as substitute for the RW edge node 120A.

Claims

[Claim 1]

A file server system comprising a plurality of edge nodes that are coupled to a core node that is one file server and that are a plurality of file servers that receive a write request and/or a read request of a file from at least one client, wherein:

- there is an edge storage region that is a storage region that stores a file for each of the edge nodes;
- the plurality of edge nodes includes at least one first edge node and second edge nodes that are more than the first edge node;
- the first edge node is an edge node that transmits a write request of a file to the core node;
- each of the second edge nodes is an edge node that does not transmit a write request of a file to the core node and that transmits a read request of a file to the core node;
- each of the edge nodes writes the file that has been obtained by a recall of the file to the edge storage region of the edge node;
- each of the second edge nodes is provided with the storage destination management information that is corresponded to the second edge node;
- the storage destination management information includes the information that indicates a storage destination for the core node of the file that has been written to the edge storage region for the second edge node that is corresponded to the information and includes the information that indicates a recall date that is a date that has been read by the second edge node; and
- each of the second edge nodes executes the following processing based on the storage destination management information that is corresponded to each of the plurality of second edge nodes:
 - (A) judging whether or not a first file that has been written to a first edge storage region that is an edge storage region for the second edge node is different from a second file that has been written to an edge storage region different from the first edge storage region in which a storage destination for the core node is equal to that of the first file;
 - (B) judging whether or not the first file is older than the second file based on a recall date of the first file and a recall date of the second file in the case in which the result of a judgment of the (A) is positive; and
 - (C) not creating a stub for the first file in the case in which the result of a judgment of the (A) or (B) is negative, and creating a stub for the first

- [Claim 2] file in the case in which the result of a judgment of the (B) is positive.
A file server system according to claim 1, wherein:
the plurality of second edge nodes is divided into at least two groups;
in the case in which each of the second edge nodes receives a read request of a read target file that is a file that is corresponded to a stub from a client, the second edge node executes the following processing:
(D) judging whether or not the read target file has been stored into a file storage region for any other second edge node that is included in a target group based on the storage destination management information of each of the other second edge nodes that is included in the target group that is a group that includes the second edge node;
(E) executing a recall of the read target file in the case in which the result of a judgment of the above (D) is negative; and
(F) transmitting a read request of the read target file to other second edge node that is corresponded to an edge storage region that has stored the read target file and transmitting the read target file that has been read by a response to the read request to a client in the case in which the result of a judgment of the above (D) is positive.
- [Claim 3] A file server system according to claim 2, wherein:
the at least two groups are configured in such a manner that a duplicated file is not preferably included in a plurality of files that has been stored into at least two edge storage regions for at least two second edge nodes that are included in one group.
- [Claim 4] A file server system according to claim 3, wherein:
each of the groups is configured by one representative node and at least zero non-representative node;
at least two edge nodes in which an amount of caches is equal to or larger than a certain value and an amount of duplicated files is equal to or larger than a certain value for the plurality of second edge nodes are at least two representative nodes that are corresponded to the at least two groups;
the amount of caches for the second edge node is a total amount of at least one file that has been stored into an edge storage region for the second edge node;
the amount of duplicated files for the second edge node is a total amount of at least one file that is duplicated with a file that has been stored into an edge storage region for other second edge node among at least one file that has been stored into an edge storage region for the

second edge node; and

the non-representative node is a second edge node that is not a representative node.

[Claim 5]

A file server system according to claim 2, wherein:

in the case in which the first edge node receives a write request of a file, the first edge node stores the file into an edge storage region for the first edge node and then transmits the write request of the file to the core node; and

in the (E), a read request for a recall of the read target file is transmitted to not the core node but the first edge node, and the first edge node obtains a file that conforms to the read request from the edge storage region for the first edge node by a response to the read request, and transmits the file to a second edge node that is a transmission source of the read request.

[Claim 6]

A file server system according to claim 2, wherein:

in the (F), in the case in which the at least two read target files have been stored into at least one edge storage region for at least one other second edge node that is included in the target group, one is selected from the at least two read target files, and a read request is transmitted to the second edge node that is corresponded to an edge storage region that has stored the selected read target file; and

the selected read target file is a file that is corresponded to at least one of the following (f1) to (f3):

(f1) a file that has been stored into an edge storage region of other second edge node in which a distance to a second edge node that executes the (F) is shortest among at least one other second edge node that is included in the target group;

(f2) a file that has been stored into an edge storage region of the second edge node in which an access load is lowest among at least one other second edge node in the target group; and

(f3) a file in which a recall date is latest.

[Claim 7]

A file server system according to claim 2, wherein:

the plurality of groups is configured in the case in which a frequency of a recall that is executed by the second edge node is equal to or larger than a certain value.

[Claim 8]

A file server system according to claim 7, wherein:

in the case in which a frequency of a recall that is executed by the second edge node is less than a certain value, a group that includes the

second edge node is released.

[Claim 9]

A file server system according to claim 7, wherein:

in the case in which the first edge node receives a write request of a file, the first edge node stores the file into an edge storage region for the first edge node and then transmits the write request of the file to the core node;

in the case in which there is a group that includes the second edge node that executes the (E), a read request for a recall of the read target file is transmitted to not the core node but the first edge node in the (E); and
in the case in which there is not a group that includes the second edge node that executes the (E), a read request for a recall of the read target file is transmitted to the core node in the (E).

[Claim 10]

A file server system according to claim 1, wherein:

the storage destination management information includes a hash value of a file; and

in the (A), it is judged whether or not a hash value of the first file is different from a hash value of the second file.

[Claim 11]

A file server system according to claim 1, wherein:

the storage destination management information includes the information related to a file to which the information that means a public has been added and does not include the information related to a file to which the information that means a private has been added.

[Claim 12]

A storage control method wherein the first edge node that is a file server that is coupled to a core node that is one file server and to a client executes the following processing based on the storage destination management information of the first edge node and the storage destination management information of the second edge node:

(A) judging whether or not a first file that has been written to a first file storage region that is a file storage region for the first edge node is different from a second file that has been written to a file storage region for the second edge node in which a storage destination for the core node is equal to that of the first file;

(B) judging whether or not the first file is older than the second file based on a recall date of the first file and a recall date of the second file in the case in which the result of a judgment of the (A) is positive; and

(C) not creating a stub for the first file in the case in which the result of a judgment of the (A) or (B) is negative, and creating a stub for the first file in the case in which the result of a judgment of the (B) is positive,

and

the storage destination management information includes the information that indicates a storage destination for the core node of the file that has been written to the file storage region for the edge node that is corresponded to the information and includes the information that indicates a recall date that is a date that has been read by the edge node.

[Claim 13]

An edge node that is a first edge node that is a file server that is coupled to a core node that is one file server and to a client and that stores a file into a first file storage region, comprising:

a communication interface apparatus;

a storage resource; and

a controller that is coupled to the communication interface apparatus and the storage resource,

wherein the storage resource stores the storage destination management information,

the storage destination management information includes the information that indicates a storage destination for the core node of the file that has been written to the file storage region for the edge node that is corresponded to the information and includes the information that indicates a recall date that is a date that has been read by the edge node, and

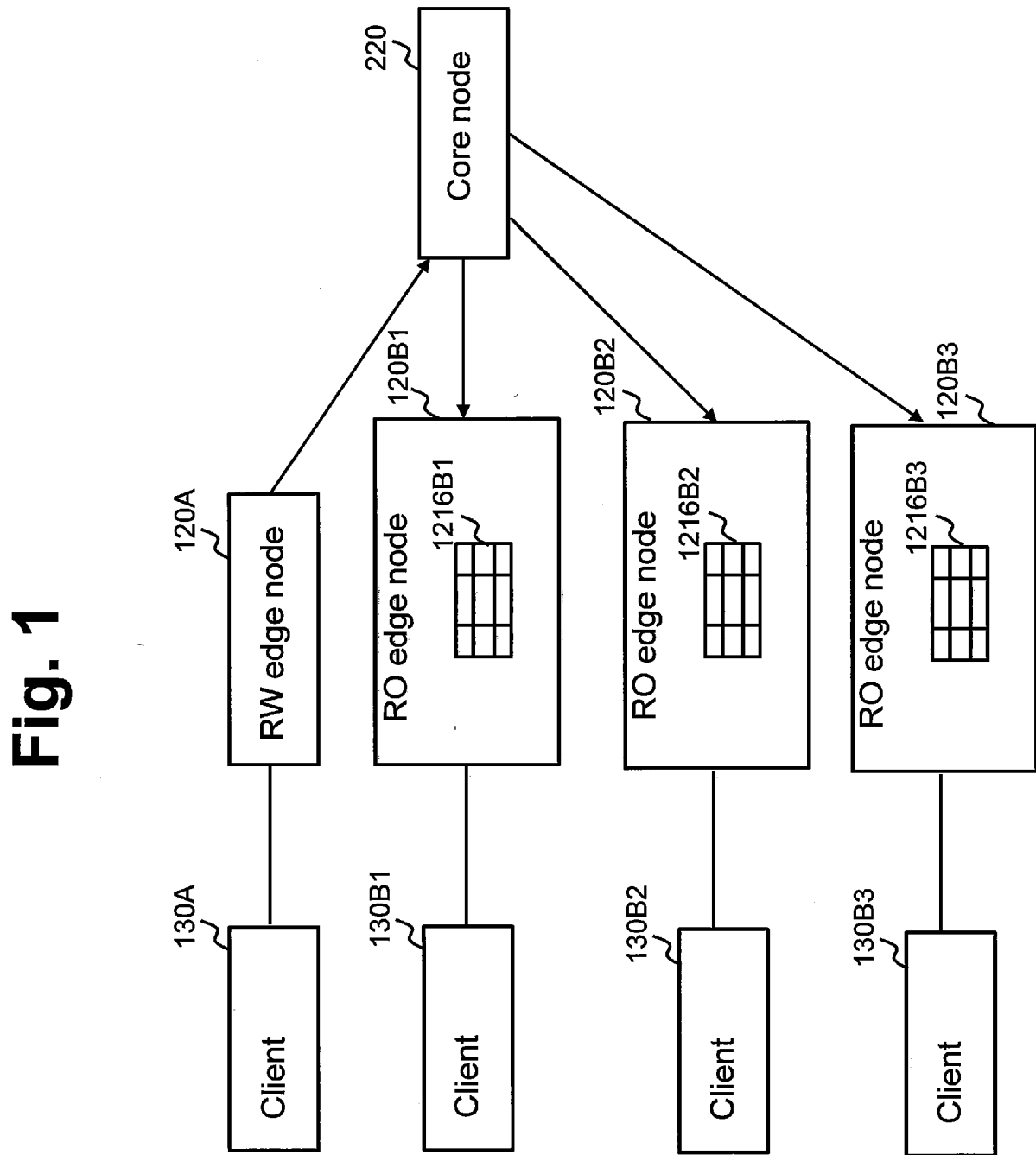
the controller executes the following processing based on storage destination management information of an edge node that is separate from the storage destination management information:

(A) judging whether or not a first file that has been written to the first file storage region is different from a second file that has been written to a file storage region for the separate edge node in which a storage destination for the core node is equal to that of the first file;

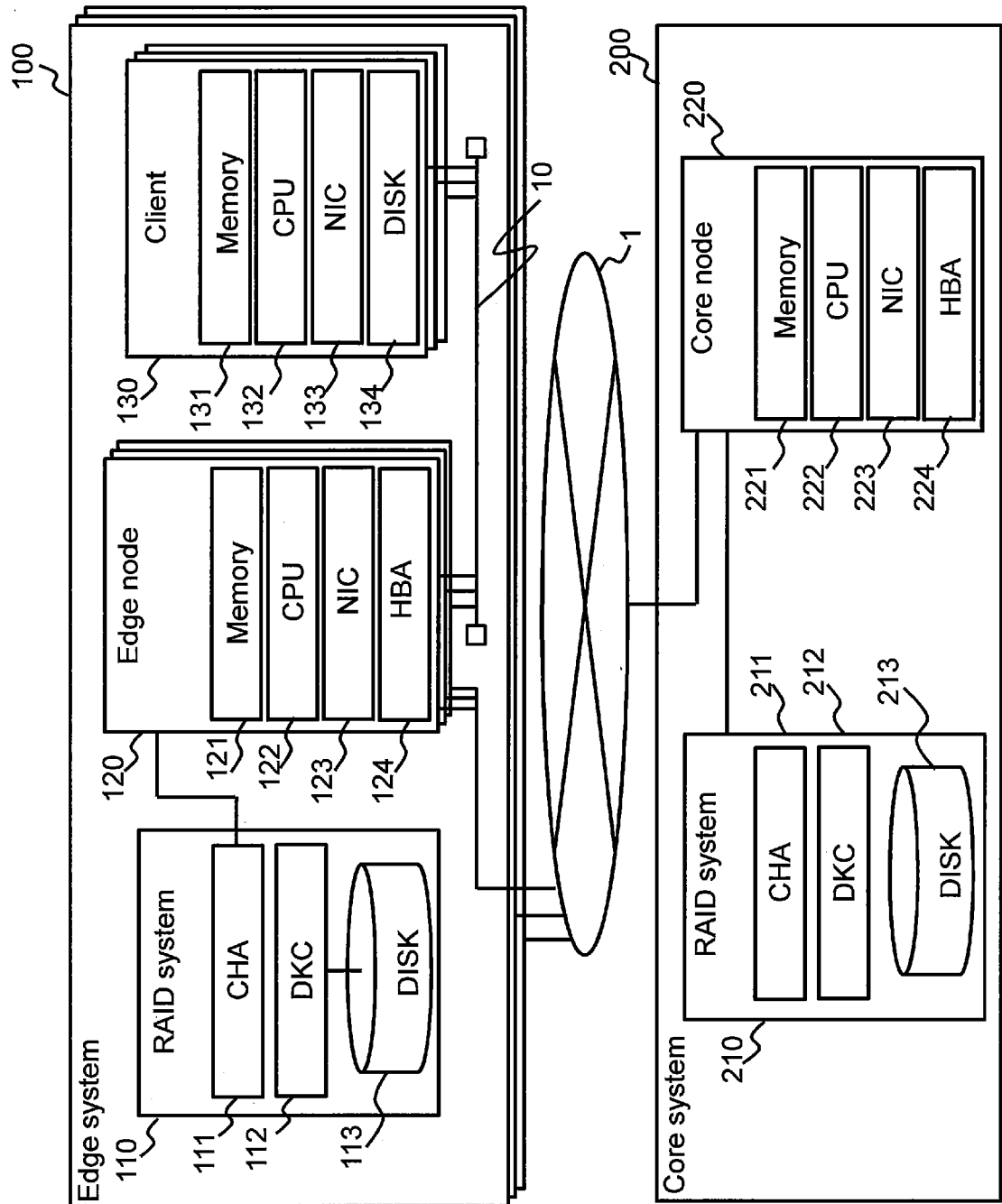
(B) judging whether or not the first file is older than the second file based on a recall date of the first file and a recall date of the second file in the case in which the result of a judgment of the (A) is positive; and

(C) not creating a stub for the first file in the case in which the result of a judgment of the (A) or (B) is negative, and creating a stub for the first file in the case in which the result of a judgment of the (B) is positive.

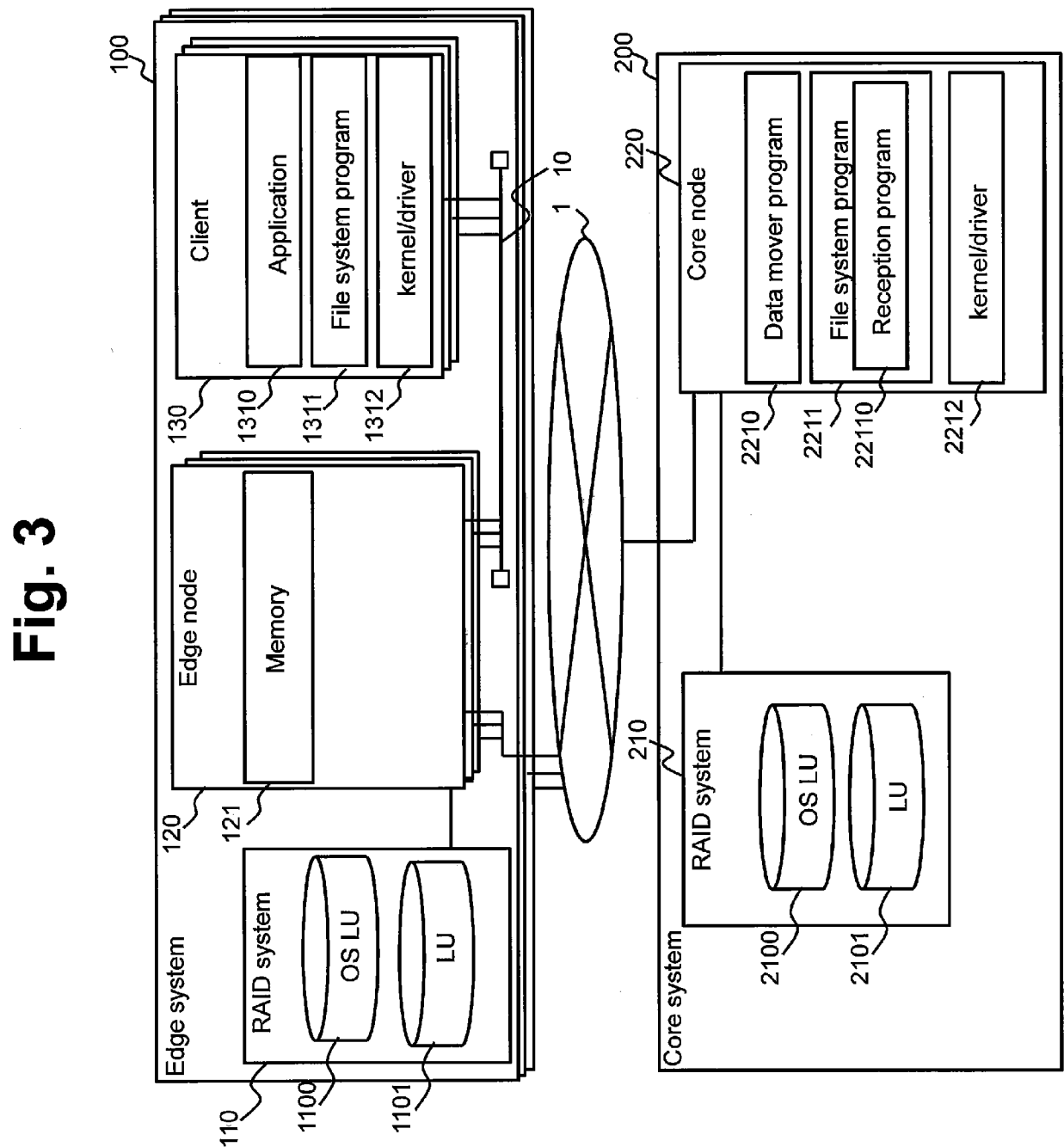
[Fig. 1]



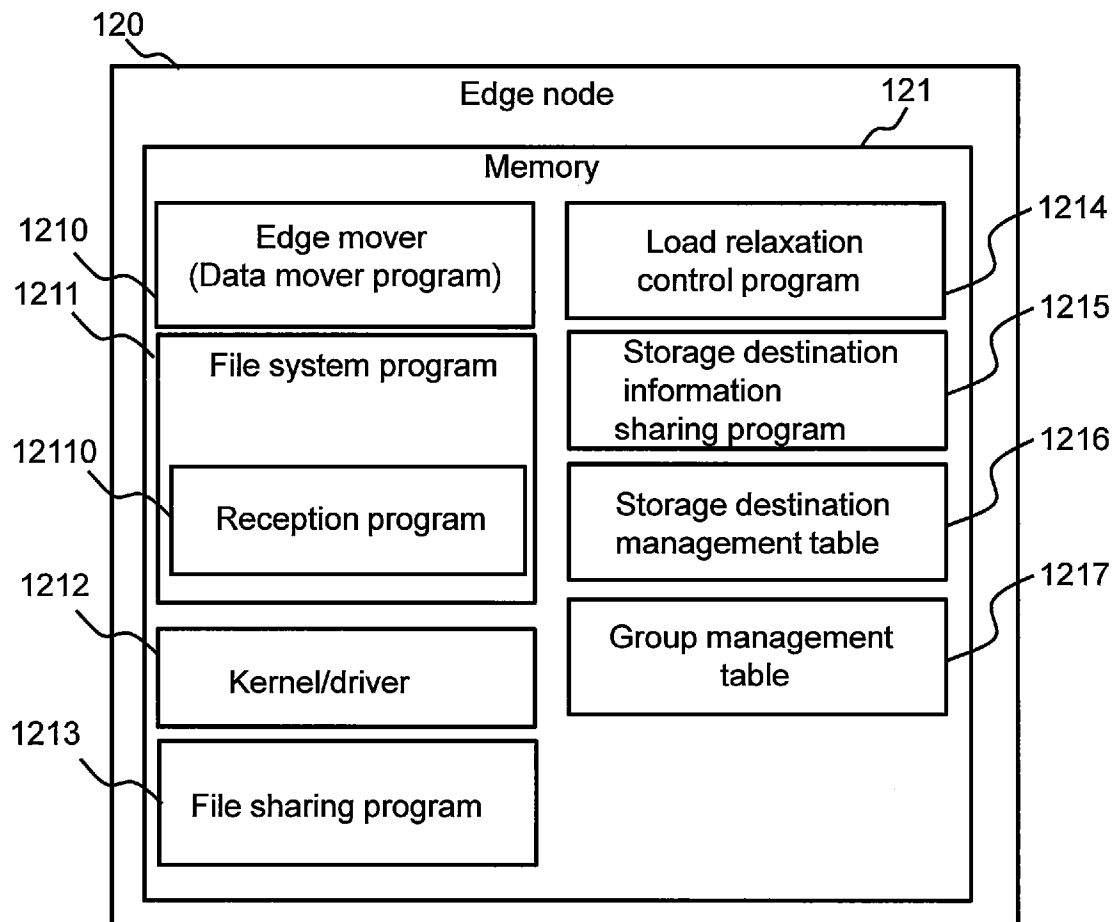
[Fig. 2]

Fig. 2

[Fig. 3]

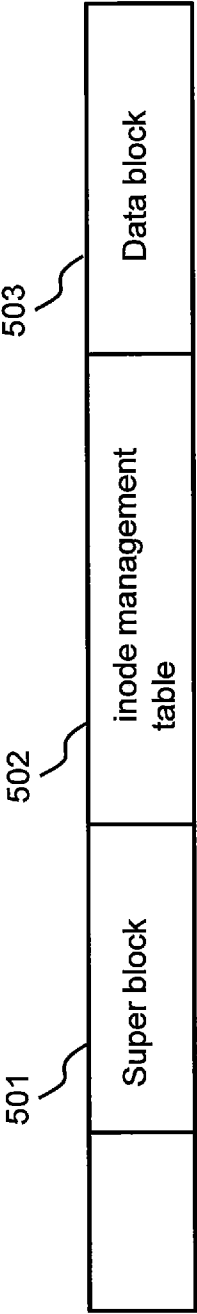


[Fig. 4]

Fig. 4

[Fig. 5]

Fig. 5



[Fig. 6]

Fig. 6

Parent directory	inode number	Child directory inode number
/	2	10
home	10	15
user-01	15	100
a.txt	100	Data block

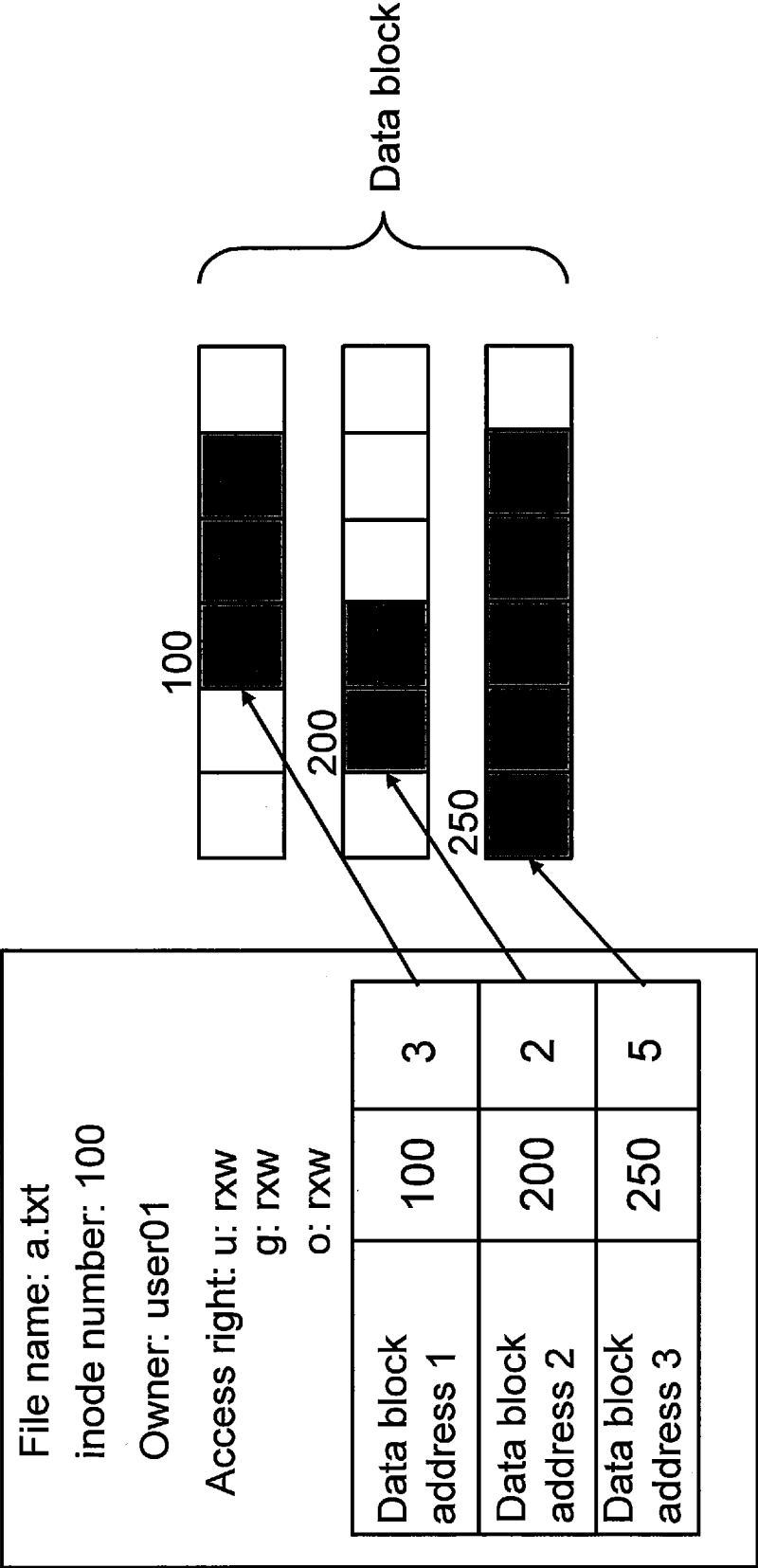
[Fig. 7]

Fig. 7

Inode number	Owner	Access right	Size	Time and date of update	Data block address 1	Data block address 2	Data block address 3
...	...	...	...		...	...	...
100	user001	u:g:o rxw rxw rxw	1000KB	200912/23 12:00:00	100-3	200-2	250-5
...	...	...	...		...	...	...

[Fig. 8]

Fig. 8



[Fig. 9]

Fig. 9A

<u>1216</u>	12161	12162	12163
File storage destination hash value	File entity hash value	Time and date of recall	
aaabbbccc	xxxxyyzzz	2010/10/22 12:00:10	
aaabbb123	xxxxyy456	2010/10/22 11:00:10	
aaabbb256	xxxxyy269	2010/10/22 11:00:10	

Fig. 9B

<u>1217</u>	12171	12172	12173
Group ID	Group configuration	Representative flag	
01	120B1	1	
	120B2	0	
	120B3	0	
02	120B4	1	
	120B5	0	

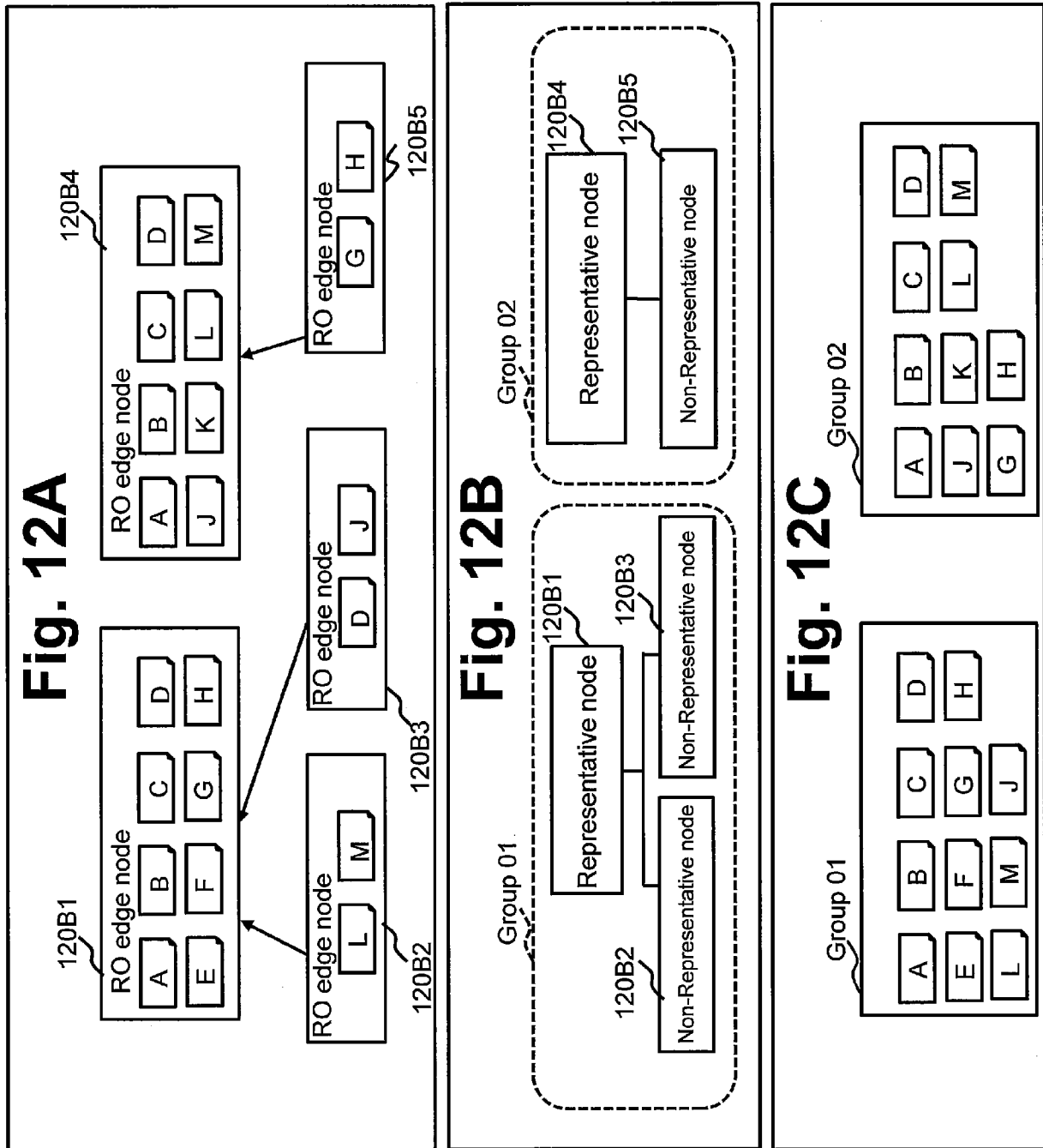
[Fig. 11]

Fig. 11

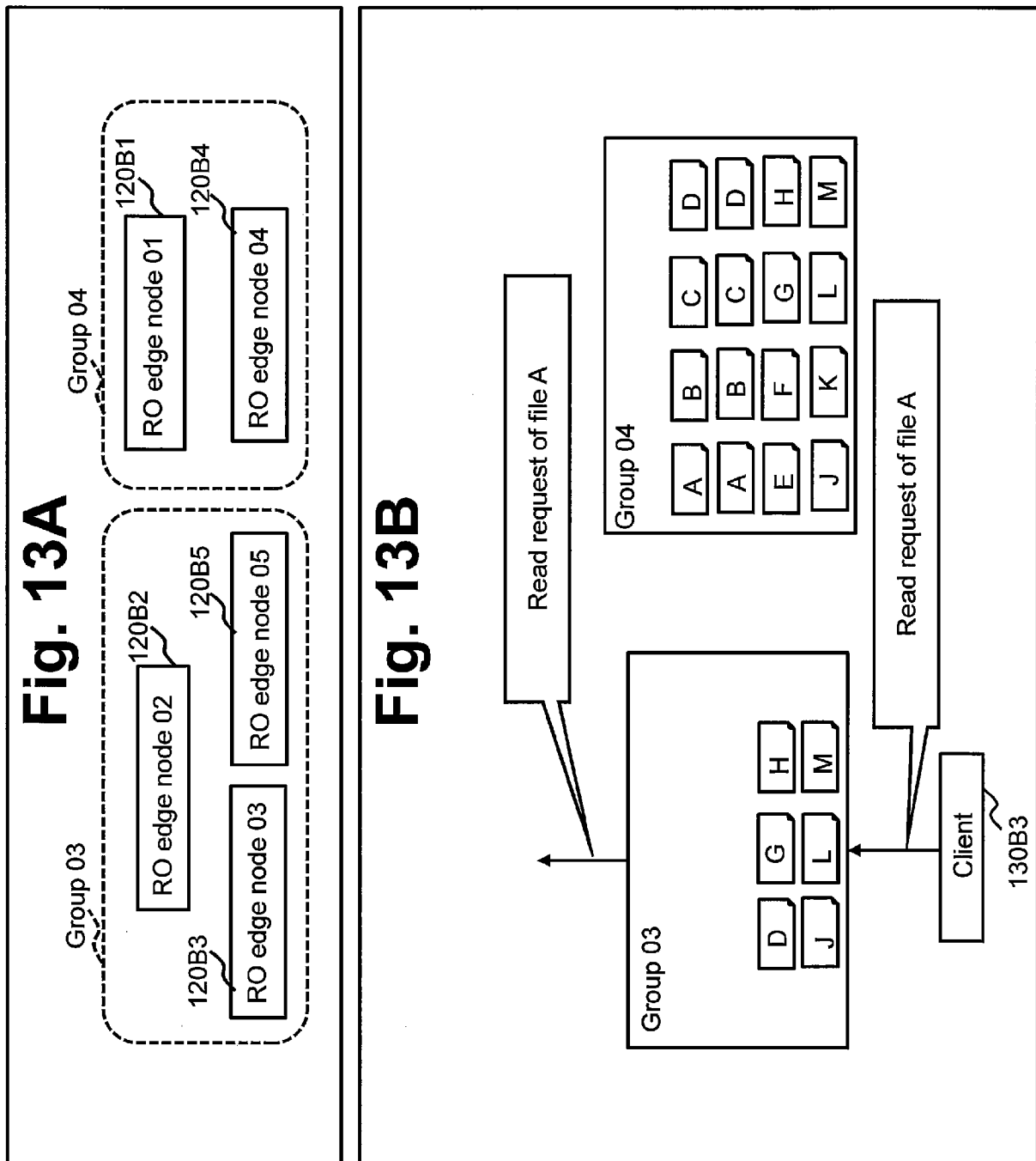
1216B1	12161	12162	12163
File storage destination hash value	File entity hash value	Time and date of recall	
aaabbbccc	xxxxyyzzz	2010/10/22 12:00:10	
aaabbb123	xxxxyy456	2010/10/22 11:00:10	
aaabbb256	xxxxyy269	2010/10/22 11:00:10	

1216B2	12161	12162	12163
File storage destination hash value	File entity hash value	Time and date of recall	
aaabbbccc	xxxxyy123	2010/10/22 11:00:10	
aaabbb123	xxxxyy456	2010/10/22 11:00:10	
aaabbb256	xxxxyy269	2010/10/22 11:00:10	

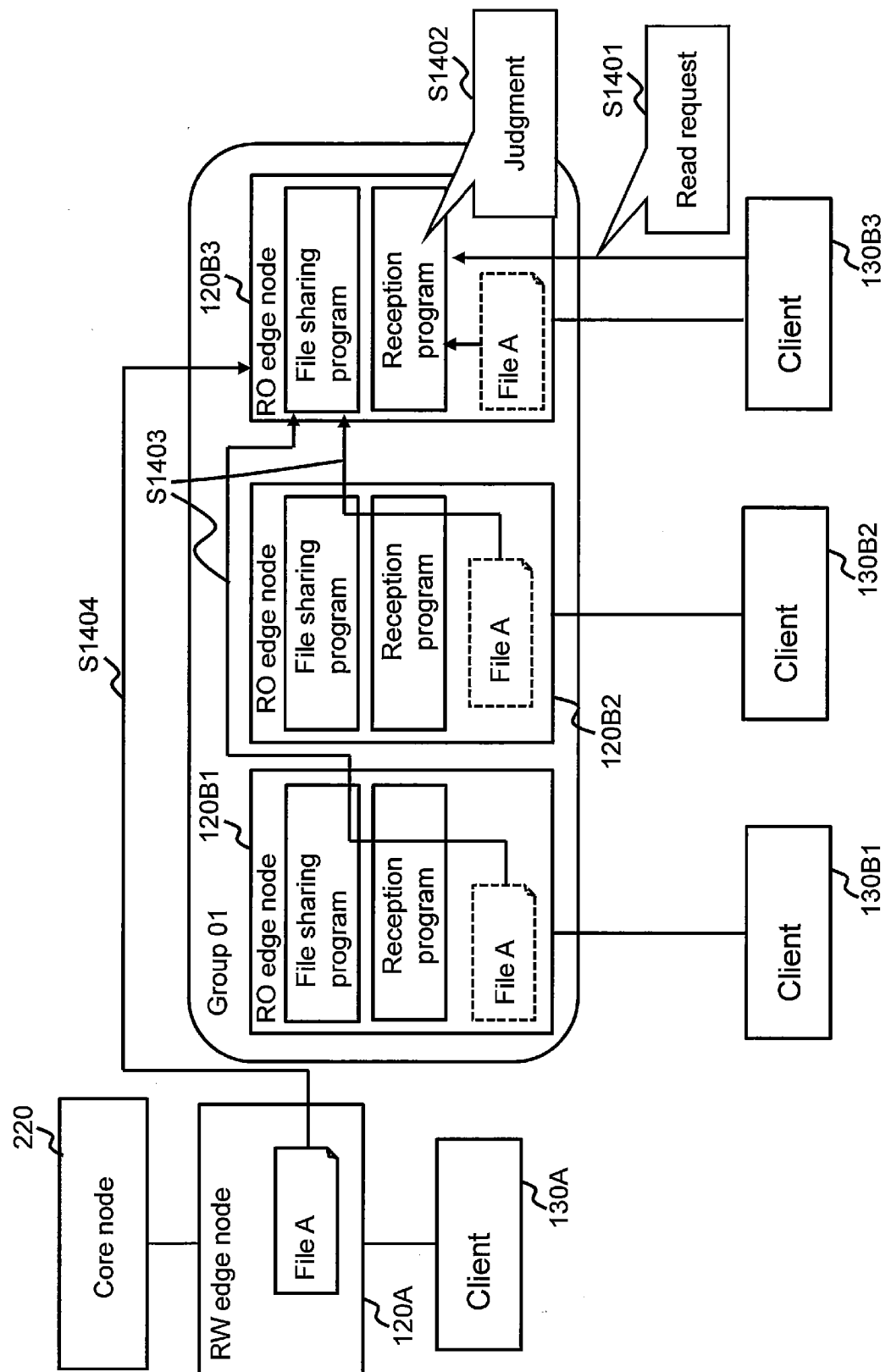
[Fig. 12]



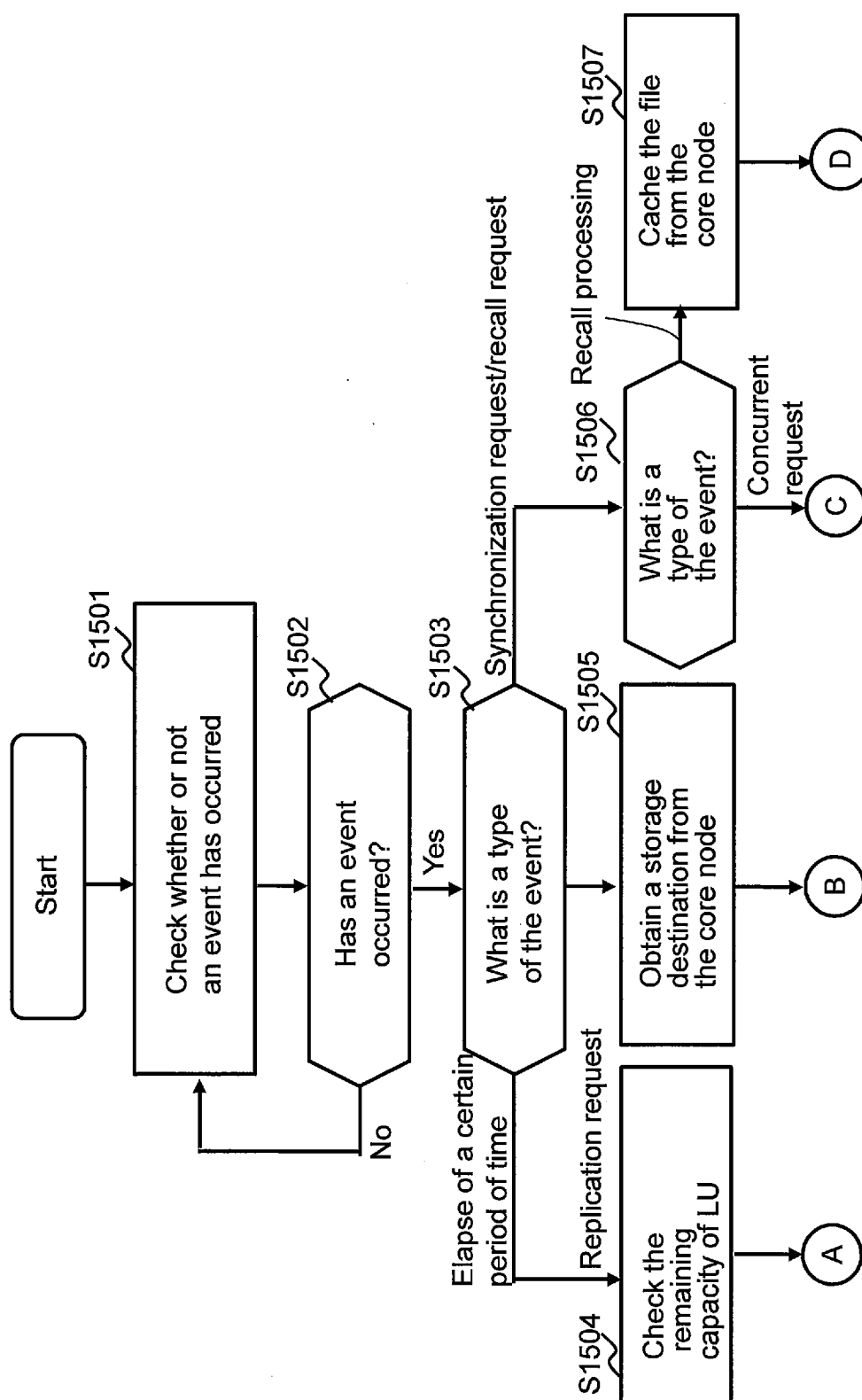
[Fig. 13]



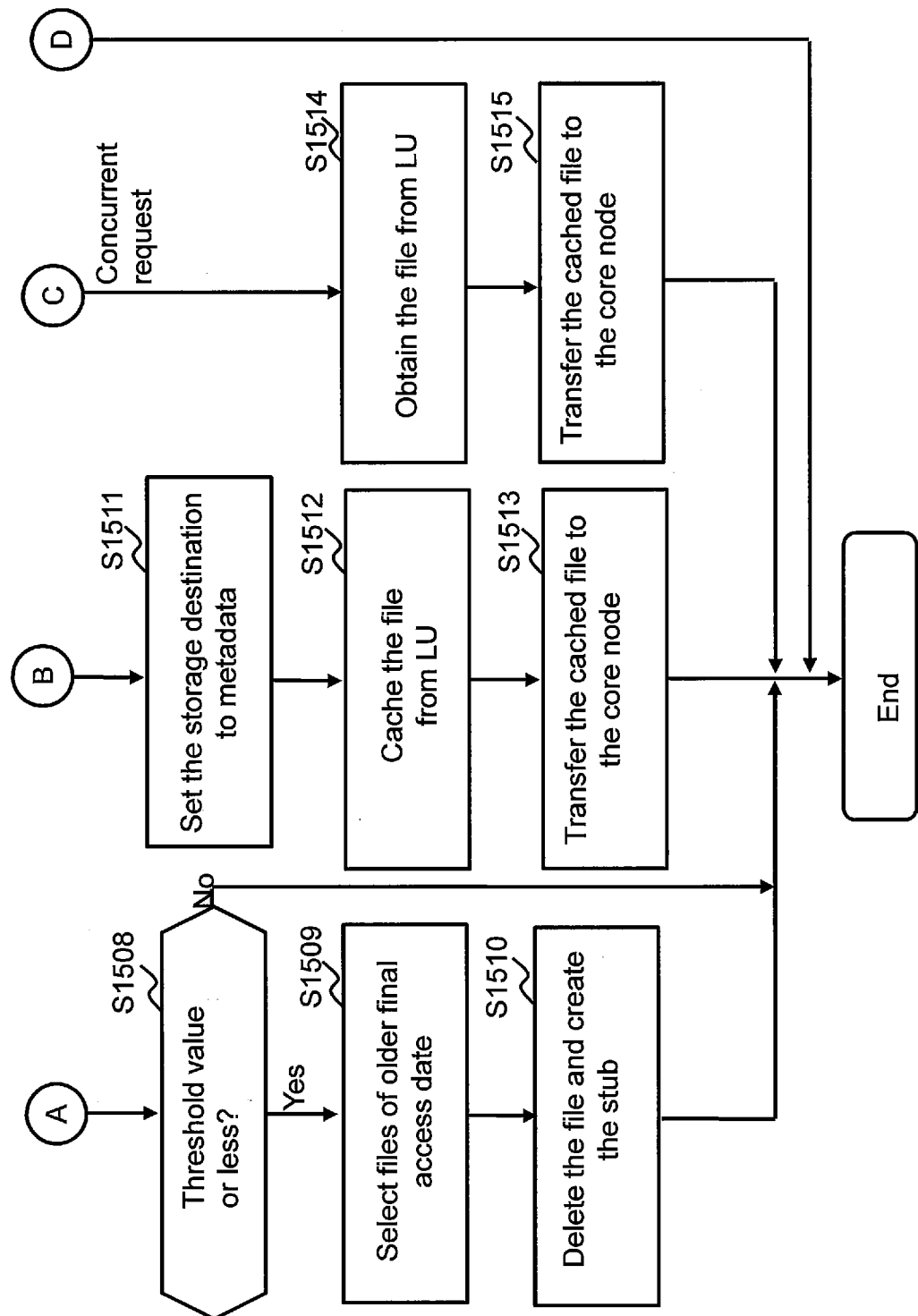
[Fig. 14]

Fig. 14

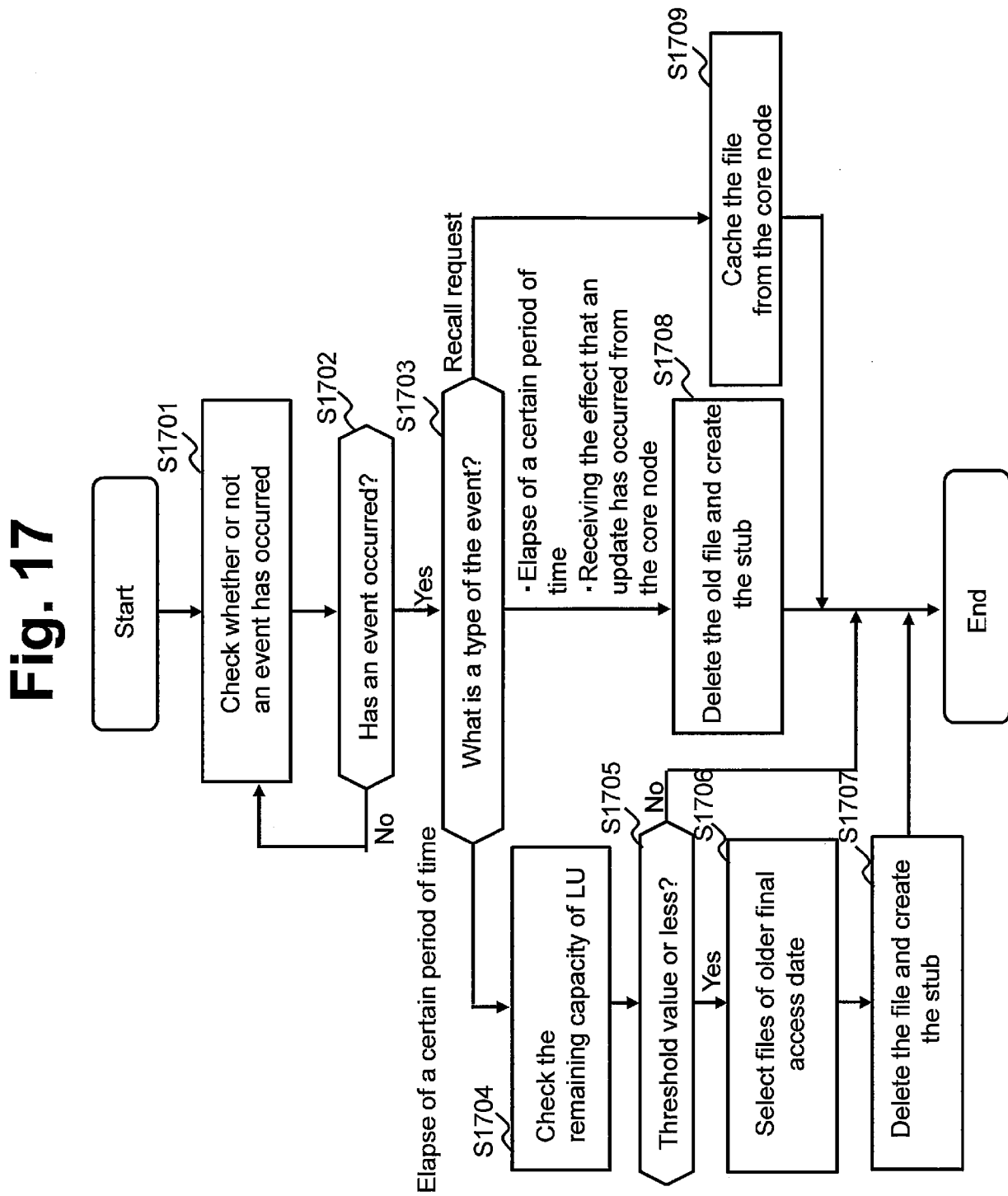
[Fig. 15]

Fig. 15

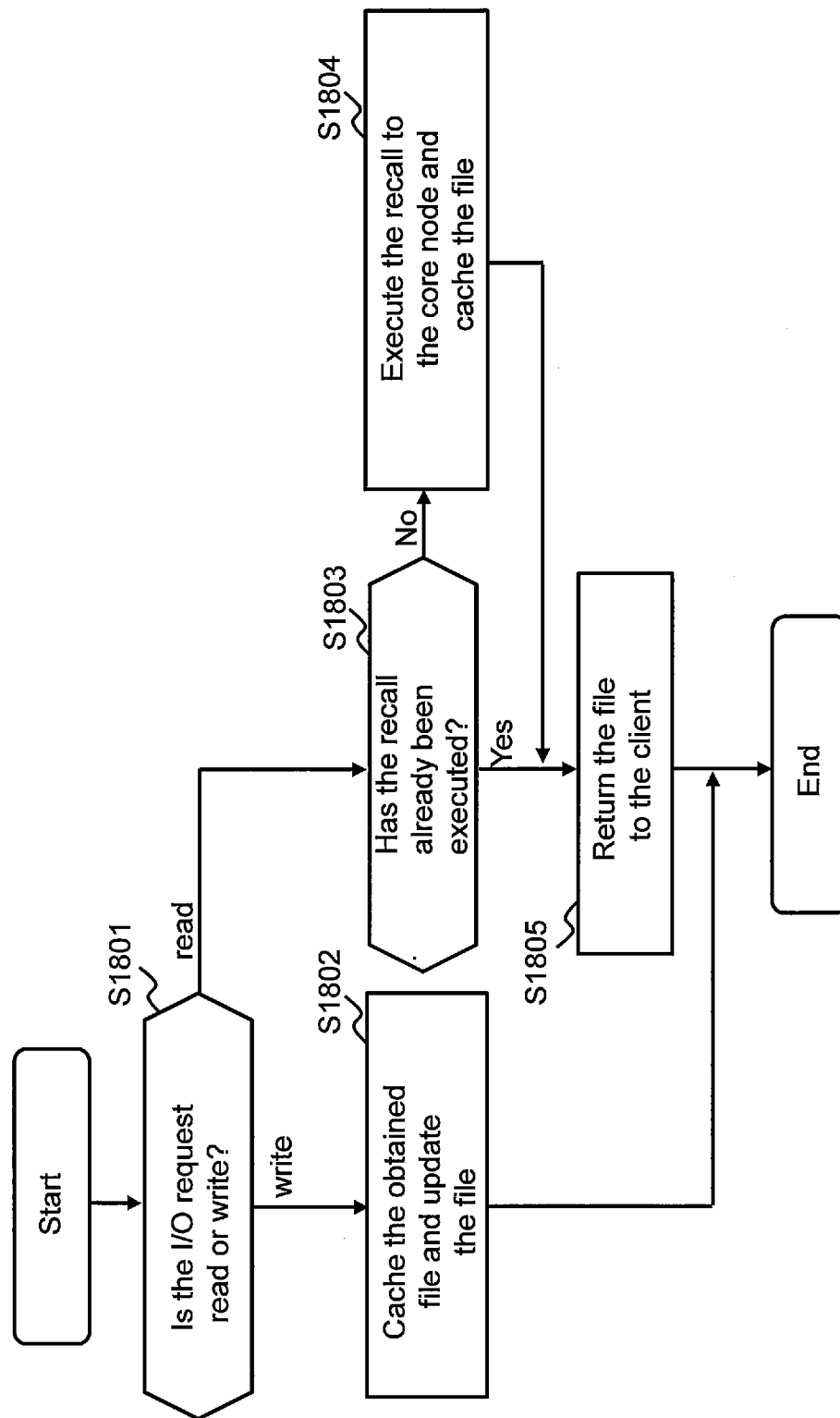
[Fig. 16]

Fig. 16

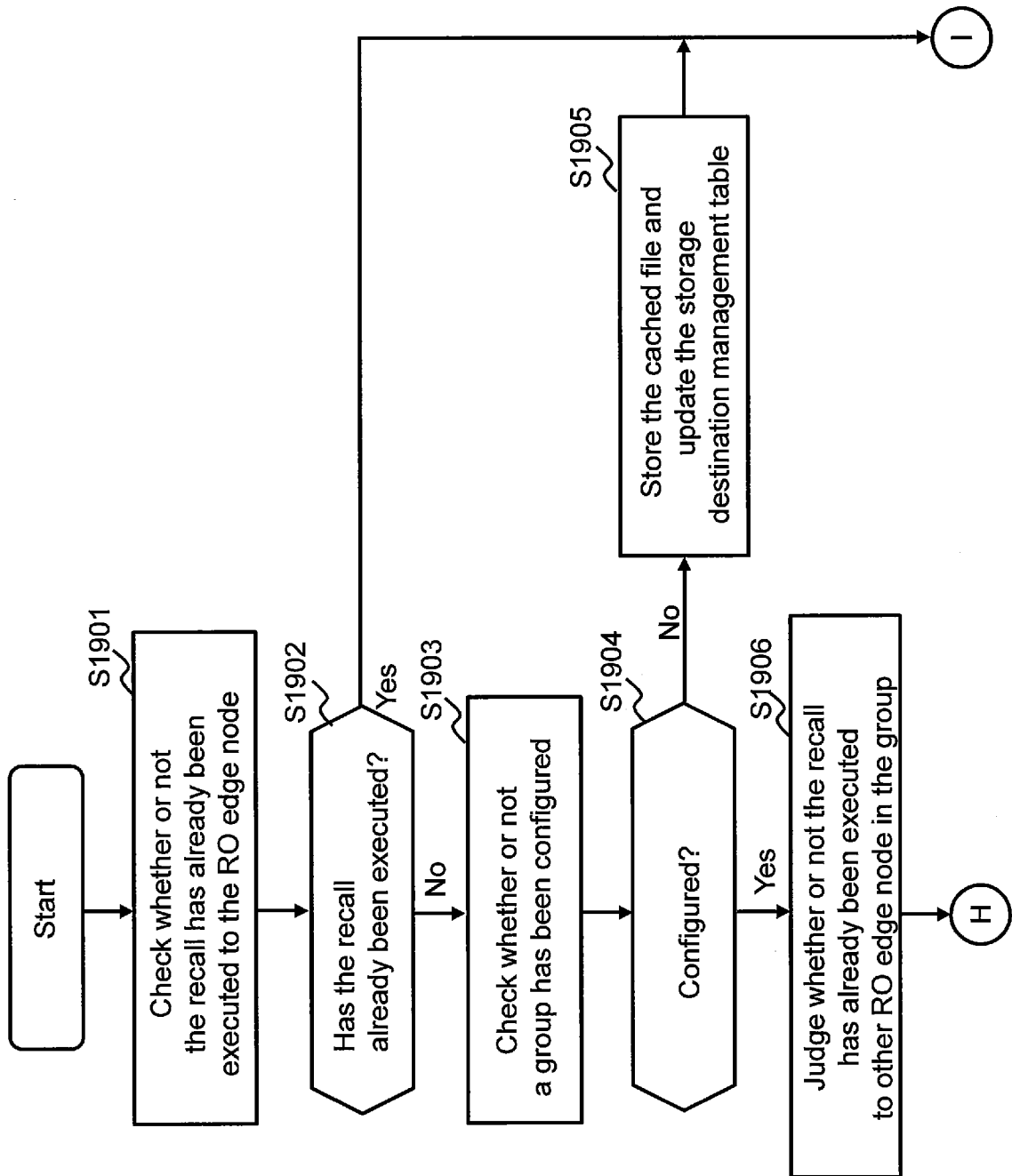
[Fig. 17]



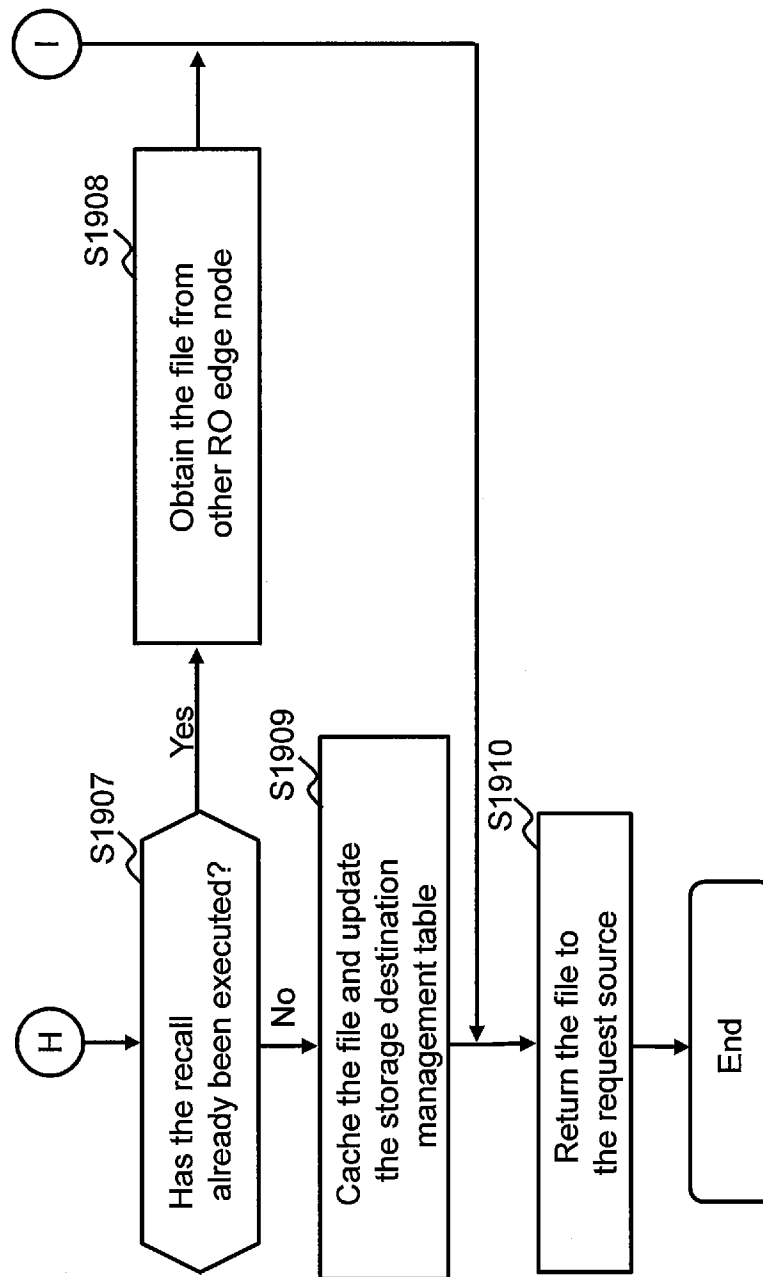
[Fig. 18]

Fig. 18

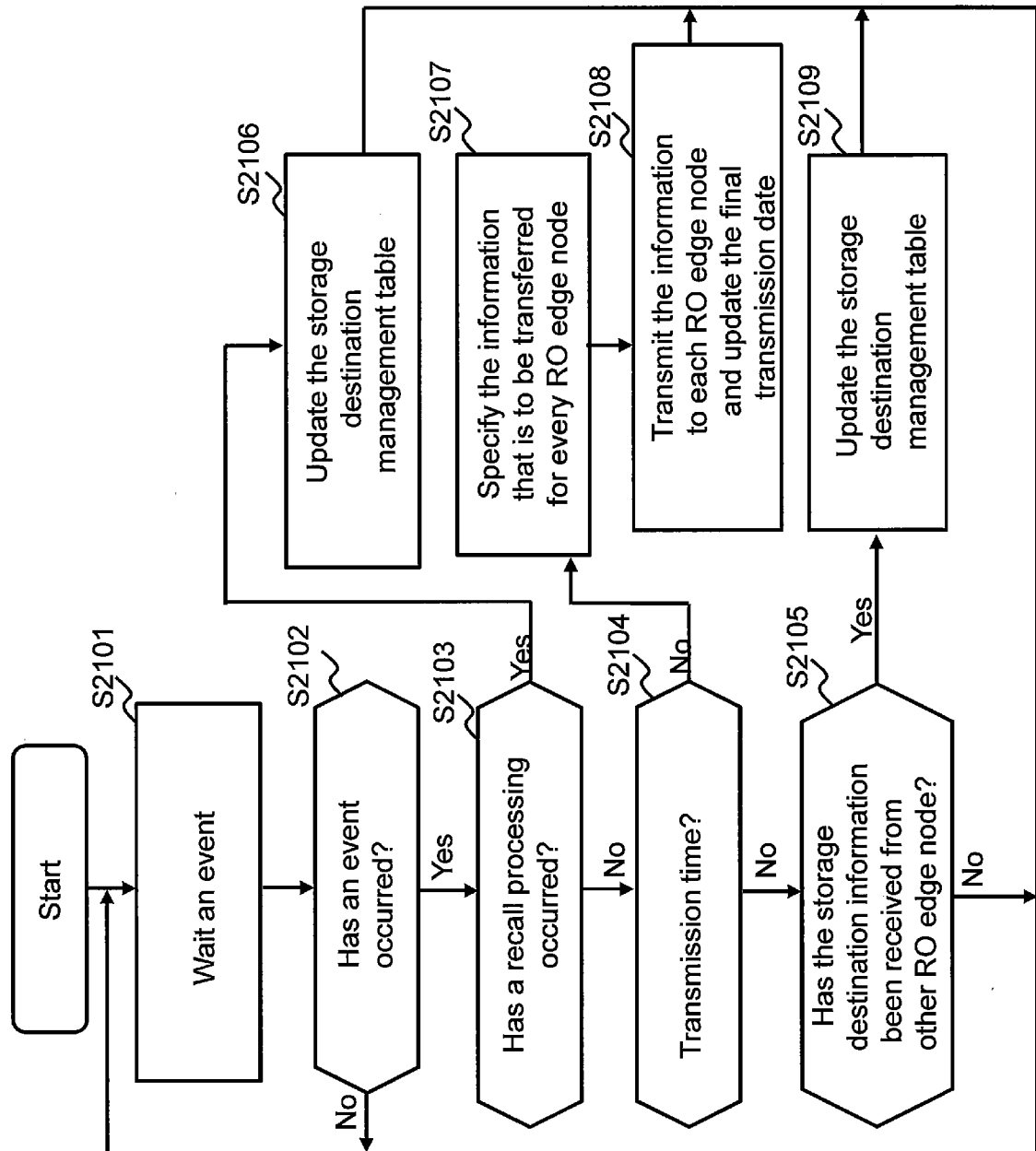
[Fig. 19]

Fig. 19

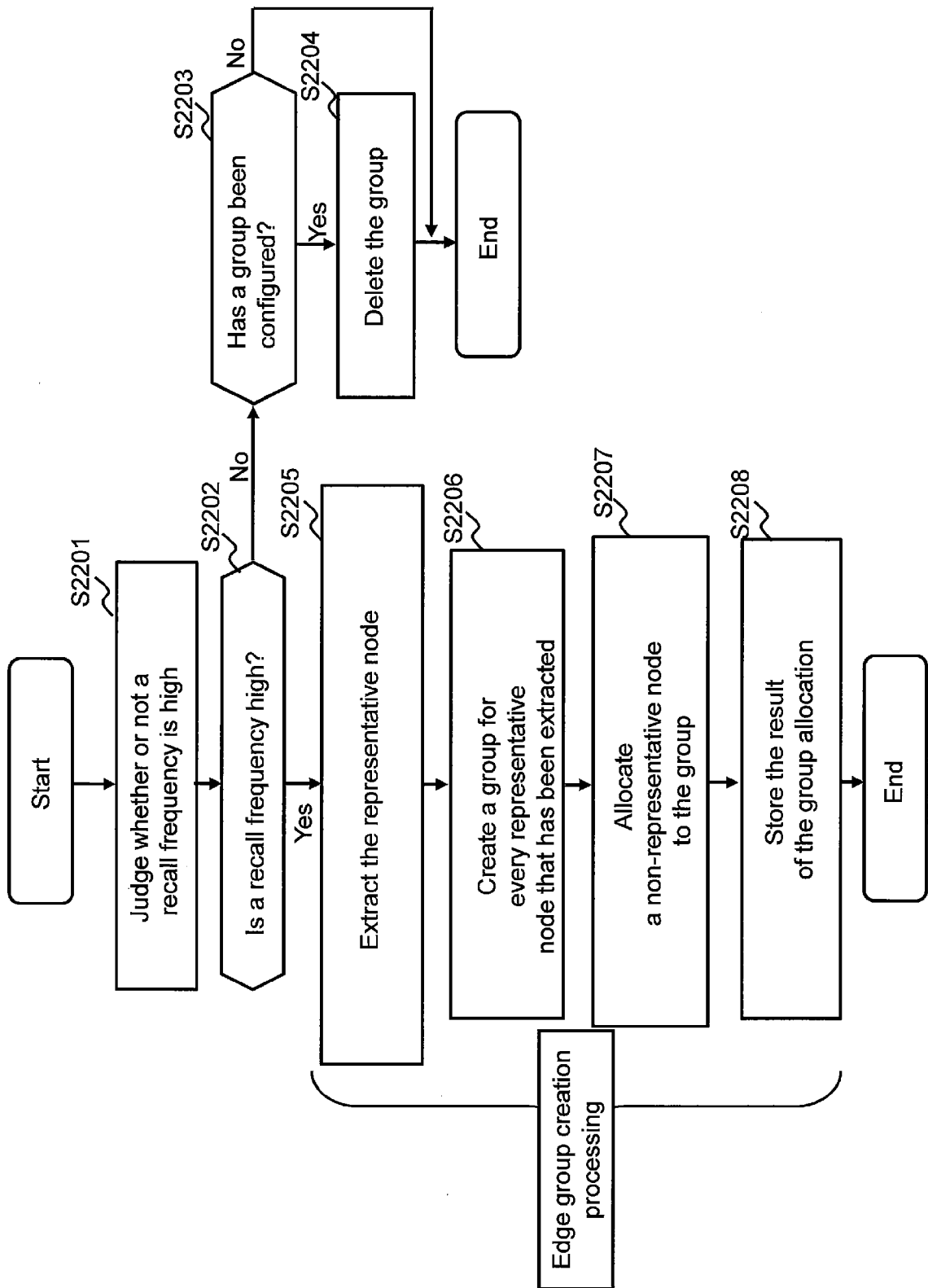
[Fig. 20]

Fig. 20

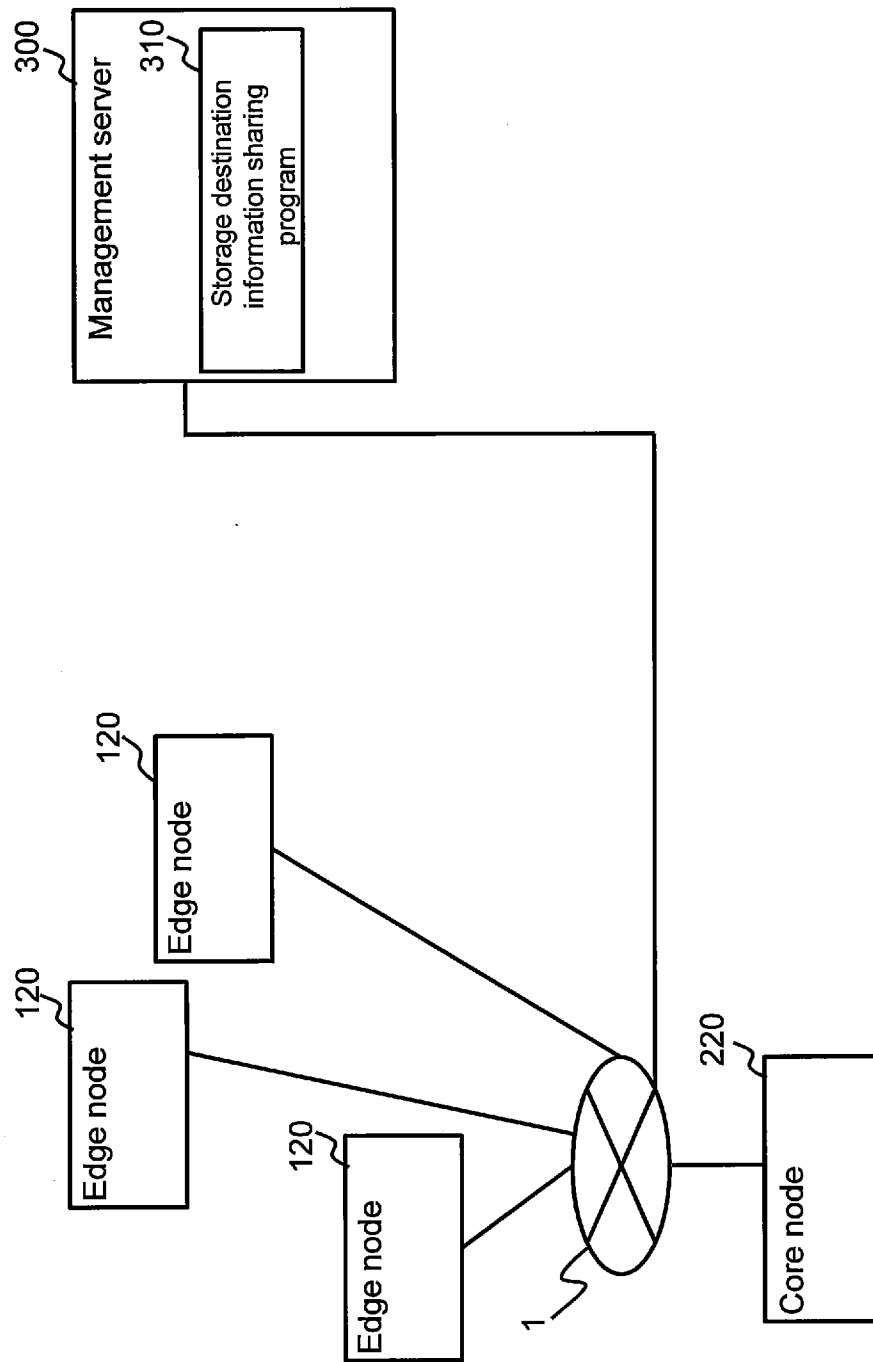
[Fig. 21]

Fig. 21

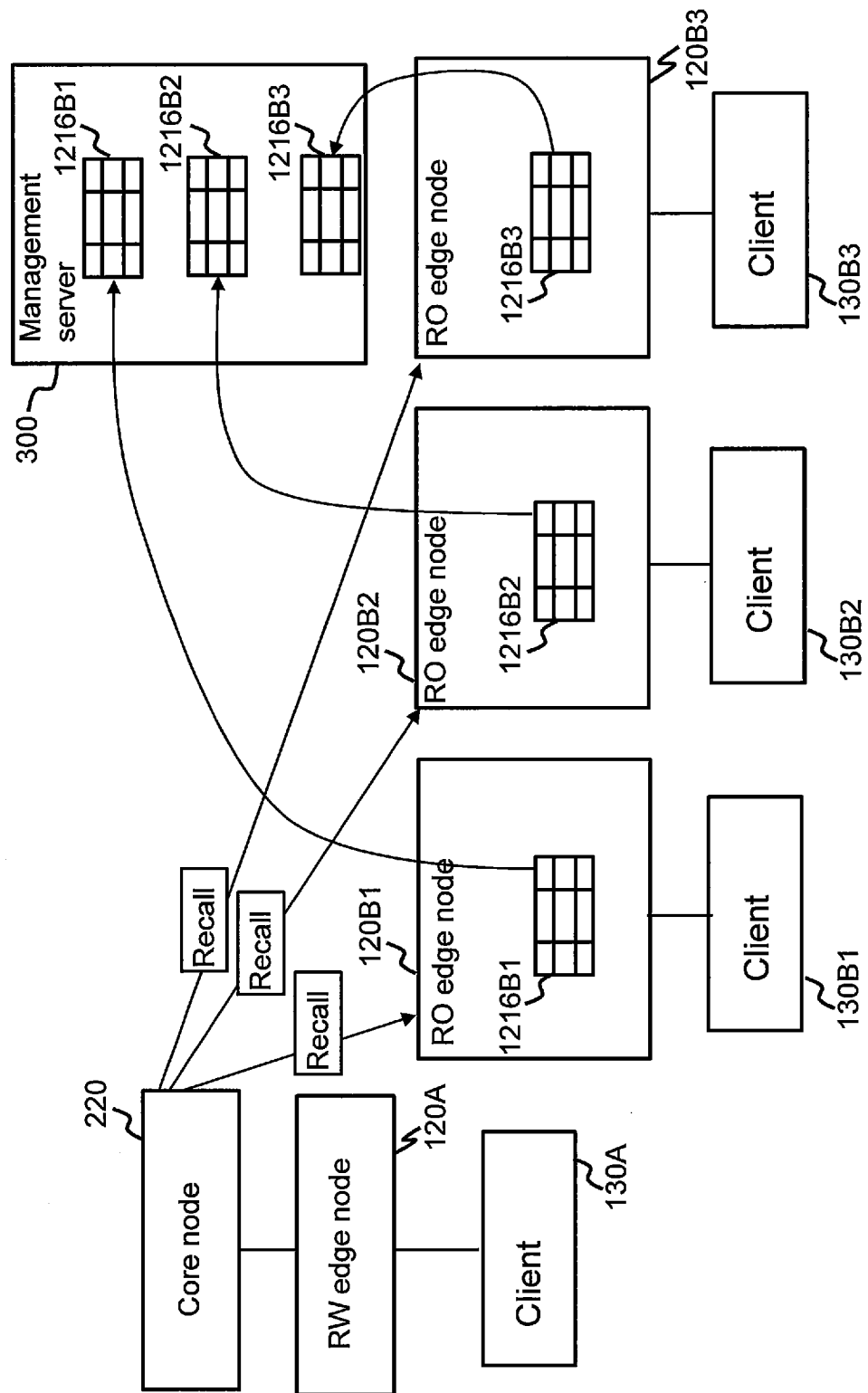
[Fig. 22]

Fig. 22

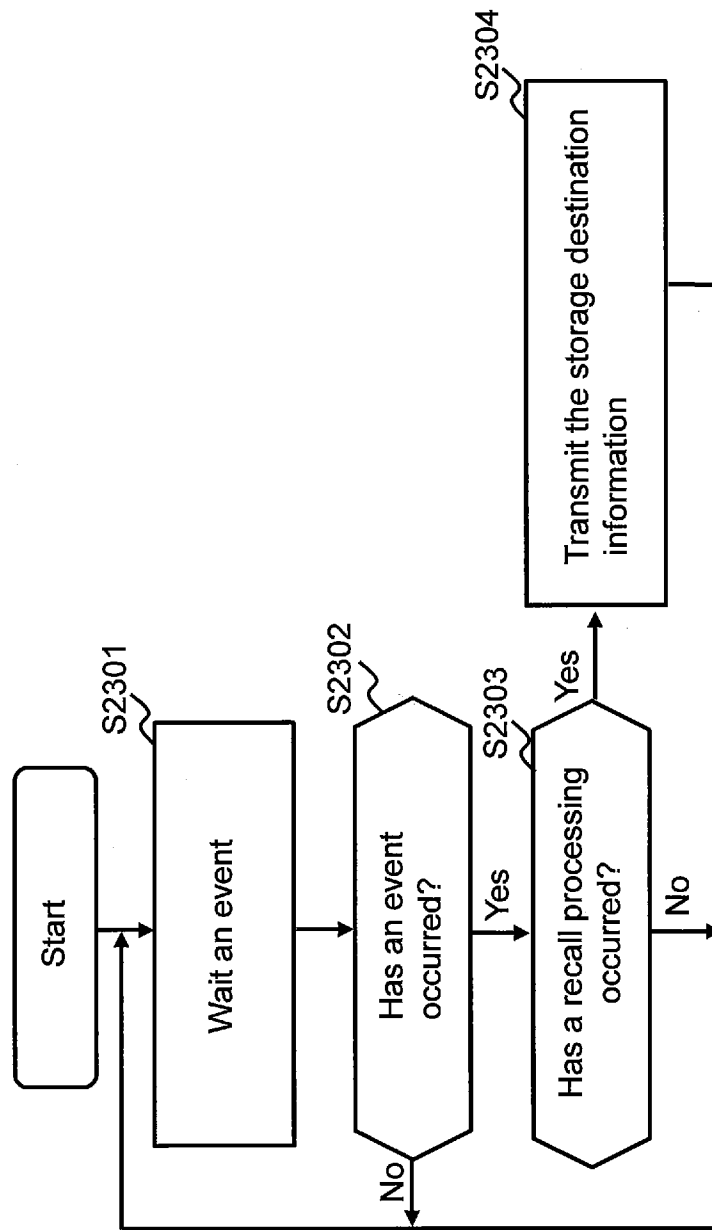
[Fig. 23]

Fig. 23

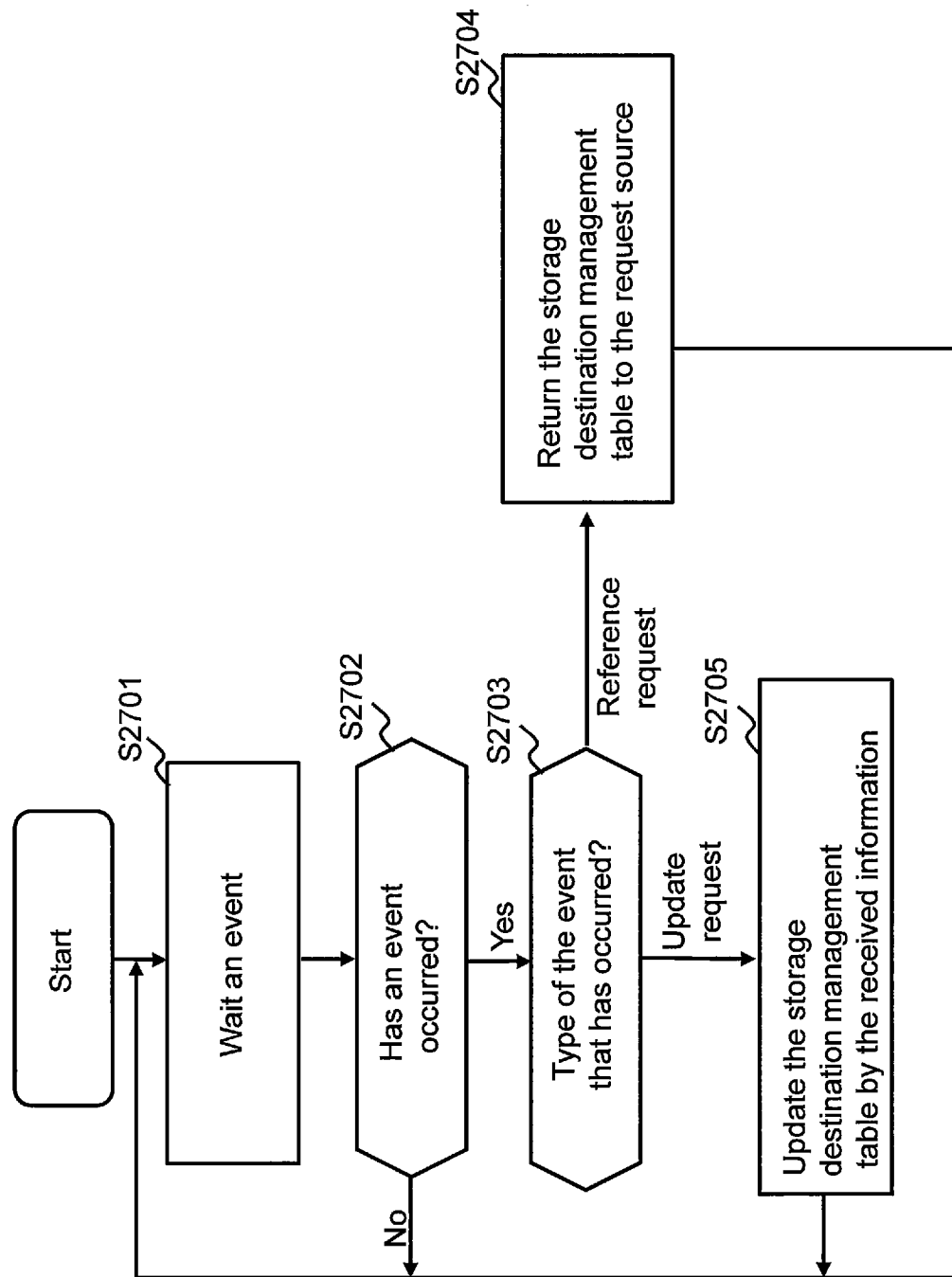
[Fig. 24]

Fig. 24

[Fig. 25]

Fig. 25

[Fig. 26]

Fig. 26

INTERNATIONAL SEARCH REPORT

International application No
PCT/JP2011/001675

A. CLASSIFICATION OF SUBJECT MATTER
INV. H04L29/08 G06F17/30
ADD.

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)
H04L G06F

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 practical, search terms used)

EPO-Internal

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Ibm: "Tivoli Storage Manager for HSM for Windows Version 6.1", 1 January 2010 (2010-01-01), pages 1-176, XP55010002, Retrieved from the Internet: URL: http://publib.boulder.ibm.com/infocenter/tsminfo/v6/topic/com.ibm.itsm.hsmwin.doc/b_hsm_guide_win.pdf [retrieved on 2011-10-19] page 35 - page 81 ----- -/-	1-13



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

31 October 2011

Date of mailing of the international search report

07/11/2011

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Warry, Lawrence

INTERNATIONAL SEARCH REPORT

International application No

PCT/JP2011/001675

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	Michael O'connor ET AL: "INTRODUCING FILE AREA NETWORKS", 1 January 2007 (2007-01-01), pages 1-208, XP55010007, Retrieved from the Internet: URL:http://www.snia.org/sites/default/files2/Intro_to_FAN.pdf [retrieved on 2011-10-19] page 117 - page 144 -----	1-13

INTERNATIONAL SEARCH REPORT

International application No.
PCT/JP2011/001675

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:
2. ☒ Claims Nos.: 1-13(partially)
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210
3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.
2. ☐ As all searchable claims could be searched without effort justifying an additional fees, this Authority did not invite payment of additional fees.
3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:
4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

- ☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
- ☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
- ☐ No protest accompanied the payment of additional search fees.

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 1-13(partially)

The subject-matter of claim 1 which can be at least partly understood is: (B) judging whether or not the first file is older than the second file based on a recall date of the first file and a recall date of the second file in the case in which the result of a judgment of the (A) is positive; and (C) not creating a stub for the first file in the case in which the result of a judgment of the (A) or (B) is negative, and creating a stub for the first file in the case in which the result of a judgment of the (B) is positive. This subject matter does not appear inventive with respect to PCT Article 33(3). Document D1 discloses a commonly known file server system as would be envisaged by the application. Document D2 discloses the commonly known techniques for reducing bandwidth by accessing server nodes only when necessary (for example when a file has changed). See, for example D2, page 2, last three paragraphs, "migration job defines a set of files and their migration behavior .. replace the original files with stub files ... depending on your configuration". It would be a mere implementation detail for the skilled person to provide a configuration such as that claimed in claim 1 where the files are compared by way of the recall date and the old files (i.e. files which do not have a new version) or stubbed. Such an implementation detail is regarded common knowledge for the skilled person and it would be an obvious step to implement such an option to the network administrator.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guideline C-VI, 8.2), should the problems which led to the Article 17(2) declaration be overcome.