(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
3 April 2014 (03.04.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/049635 A1

(51) International Patent Classification:
G06F 3/06 (2006.01)

(21) International Application Number:
PCT/JP2012/006097

(22) International Filing Date:
25 September 2012 (25.09.2012)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **HITACHI, LTD.** [JP/JP]; 6-6, Marunouchi 1-chome, Chiyoda-ku, Tokyo, 1008280 (JP).

(72) Inventors: **TSUKADA, Masaru**; c/o HITACHI, LTD., IT Platform Division Group, 322-2, Nakazato, Odawara-shi, Kanagawa, 2500872 (JP). **HIRAMATSU, Shinichi**; c/o HITACHI, LTD., IT Platform Division Group, 322-2, Nakazato, Odawara-shi, Kanagawa, 2500872 (JP). **GOTO, Yuya**; c/o HITACHI, LTD., IT Platform Division Group, 322-2, Nakazato, Odawara-shi, Kanagawa, 2500872 (JP). **MIYASHITA, Jun**; c/o HITACHI, LTD., IT Platform Division Group, 322-2, Nakazato, Odawara-shi, Kanagawa, 2500872 (JP).

(74) Agent: **ISSHIKI & CO.**; Rookin-Shinbashi Bldg., 12-7, Shinbashi 2-chome, Minato-ku, Tokyo, 1050004 (JP).

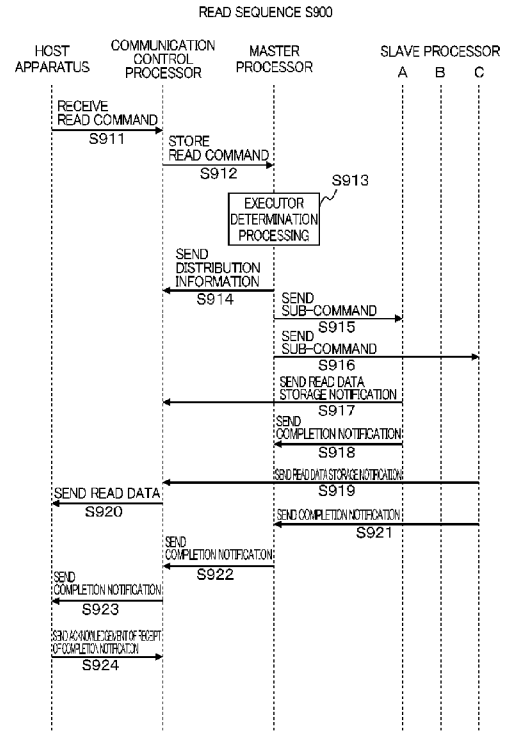
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE,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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

[Continued on next page]

(54) Title: STORAGE SYSTEM AND METHOD THEREOF FOR I/O LOAD DISTRIBUTION WITH DATA STORED IN CKD FORMAT

[Fig. 9]



(57) Abstract: A storage apparatus 10 includes a communication control processor 11 that is a processor configured to control communications with a host apparatus 3, and plurality of main processors 12 each configured to perform an I/O process on a storage drive 171 according to an I/O request received from the host apparatus 3. The storage apparatus 10 manages data to be stored or that has been stored in the storage drive 171 in accordance with the CKD format. The storage apparatus 10 distributes the I/O process to the plurality of main processors 12 in units of the data fields on basis of the operation rates of the respective main processors 12, information on a C field, a K field, and a D field that are data fields forming a record of data targeted by the I/O process and managed in CKD format, and an I/O load indicator being a load indicator of the I/O process currently running on each of the main processors.

WO 2014/049635 A1

Published:

— *with international search report (Art. 21(3))*

Description

Title of Invention: STORAGE SYSTEM AND METHOD THEREOF FOR I/O LOAD DISTRIBUTION WITH DATA STORED IN CKD FORMAT

Technical Field

[0001] The present invention relates to a storage apparatus and a method of controlling a storage apparatus.

Background Art

[0002] PTL 1 discloses a storage system including a plurality of microprocessors and configured to prevent load concentration on one microprocessor in the following manner. Specifically, when one of the plurality of microprocessors performing an I/O process for a received I/O request has a load heavier than a first load, the microprocessor allocates at least an I/O process part of the I/O process to any of the other microprocessors and the other microprocessor performs at least the I/O process part (the I/O process part is a process part including a cache control process involving reserving one area of a cache memory 20, i.e., a cache area for storing data in the cache memory 20).

Citation List

Patent Literature

[0003] PTL 1: Japanese Patent Application Laid-open Publication No. 2007-249729

Summary of Invention

Technical Problem

[0004] Nowadays, more and more storage apparatuses are equipped with multi-core processors in order to enhance processing performance and response performance. A storage apparatus equipped with a multi-core processor is capable of performing parallel processing among core processors and is expected to achieve performance improvement of the whole storage apparatus.

[0005] Even being equipped with a multi-core processor, however, the storage apparatus cannot always receive the benefit of the multi-core processor to a process not subjected to the parallel processing. For example, when the storage apparatus receives a read request from a mainframe, only one core processor performs a data transfer process for the read request. In this case, effects on performance improvement can be expected if the apparatus operating under a heavy load as a whole has multiple core processors operating in parallel. However, if the apparatus only operates under a low load with some core processors not running any processes, the load is concentrated on the core processor in charge of the process, and the effects produced by the multi-core

processor cannot be expected.

[0006] Meanwhile, mainframes widely employ, as a data management method, a so-called CKD (Count Key Data) format in which one track consists of a predetermined number of records each including three data fields of a C field, a K field, and a D field. Currently, however, no effective load balancing among core processors is made for data management using this CKD format.

[0007] The present invention has been made in view of such circumstances, and has an objective to provide a storage apparatus and a method of controlling a storage apparatus, which are capable of improving performance of the storage apparatus.

Solution to Problem

[0008] One aspect of the present invention to achieve the foregoing objective is a storage apparatus including a communication control processor that is a processor configured to control communications with an external apparatus, and a plurality of main processors each configured to perform an I/O process on a storage drive according to an I/O request received from the external apparatus, wherein the storage apparatus manages data to be stored or data that has been stored in the storage drive in accordance with CKD format, and distributes the I/O process to a plurality of the main processors in units of data fields on basis of operation rates of the respective main processors and information on a C field, a K field, and a D field that are data fields forming a record of data targeted by the I/O process and managed in CKD format.

[0009] Other problems and solutions thereto disclosed by the present application will be clarified in the description of embodiments, drawings, and the like.

Advantageous Effects of Invention

[0010] According to the present invention, the performance of a storage apparatus can be improved.

Brief Description of Drawings

[0011] [fig.1]Fig. 1 is a diagram illustrating a configuration of an information processing system 1.

[fig.2]Fig. 2 is a diagram illustrating hardware of a communication control processor 11.

[fig.3]Fig. 3 is a diagram illustrating hardware of a main processor 12.

[fig.4]Fig. 4 is a diagram illustrating a configuration of a CCW 410.

[fig.5]Fig. 5 is a diagram for explaining a basic operation of a storage apparatus 10 for a CCW chain 420.

[fig.6]Fig. 6 is a diagram for explaining the basic operation of the storage apparatus 10 for the CCW chain 420.

[fig.7]Fig. 7 is a diagram illustrating a configuration of a CKD format.

[fig.8]Fig. 8 is a diagram for explaining data management units (record and track) in the CKD format.

[fig.9]Fig. 9 is a diagram for explaining a read sequence S900.

[fig.10]Fig. 10 is a diagram for explaining a write sequence S1000.

[fig.11]Fig. 11 is a flowchart for explaining executor determination processing S913.

[fig.12]Fig. 12 is a flowchart for explaining executor selection processing S1131.

[fig.13]Fig. 13 is a diagram illustrating an example of operation information 1300.

[fig.14]Fig. 14 is a flowchart for explaining distribution target selection processing S1222.

[fig.15]Fig. 15 is a flowchart for explaining distribution processing S1223.

[fig.16]Fig. 16 is a flowchart for explaining allocation processing S1513.

[fig.17]Fig. 17 is a diagram illustrating an example of allocation information 1700.

[fig.18]Fig. 18 is a flowchart for explaining read data transfer processing S1800.

[fig.19]Fig. 19 is a diagram illustrating an example of an entry 1900.

Description of Embodiments

[0012] Fig. 1 illustrates an information processing system 1 described as an embodiment. As illustrated in Fig. 1, the information processing system 1 includes one or more host apparatuses 3 (external apparatuses) and one or more storage apparatuses 10.

[0013] The host apparatuses 3 are information processing apparatuses (computers) that provide services such as an automated teller service in a bank and a web page browsing service on the Internet. The host apparatuses 3 in the present embodiment are mainframes (general-purpose computers).

[0014] The host apparatuses 3 and the storage apparatuses 10 are communicatively coupled to each other through a communication network 5. The communication network 5 is, for example, a LAN (Local Area Network), a WAN (Wide Area Network), a SAN (Storage Area Network), the Internet, a public communication network, a dedicated line, or the like.

[0015] Communications through the communication network 5 are performed in accordance with protocols such for example as TCP/IP, iSCSI (internet Small Computer System Interface), Fibre Channel Protocol, FICON (Fibre Connection) (registered trademark), ESCON (Enterprise System Connection) (registered trademark), ACONARC (Advanced Connection Architecture) (registered trademark), and FIBARC (Fibre Connection Architecture) (registered trademark).

[0016] The host apparatuses 3 each include a CPU 31(Central Processing Unit), a memory 32 (RAM (Random Access Memory), ROM (Read Only Memory), or NVRAM (Non Volatile Random Access Memory)), a storage device 33 (for example, a hard disk drive, or a semiconductor storage device (SSD (Solid State Drive))), an input device 34

such as a keyboard and a mouse, an output device 35 such as a liquid crystal monitor or a printer, and a network interface (hereinafter, referred to as a network I/F 36) such as NIC (Network Interface Card) or HBA (Host Bus Adapter).

[0017] When accessing a storage area provided by any of the storage apparatuses 10, the host apparatus 3 sends the storage apparatus 10 a data frame (hereinafter, simply referred to as a frame) including an I/O request (such as a write request or a read request). This frame is, for example, a Fibre Channel (FC) frame (FC frame).

[0018] Each of the storage apparatuses 10 is an apparatus configured to provide data storage areas to the host apparatuses 3, and is a disk array apparatus, for example. As illustrated in Fig. 1, the storage apparatus 10 includes one or more communication control processors 11, one or more main processors 12, a data storage buffer 14, a shared memory 15, an internal switch 16, a storage device 17, and a maintenance device 18.

[0019] The communication control processors 11, the main processors 12, the data storage buffer 14, and the shared memory 15 are communicatively coupled to each other through the internal switch 16. In addition, the maintenance device 18 is communicatively coupled to the communication control processors 11, the main processors 12, the data storage buffer 14, the shared memory 15, and the internal switch 16 through appropriate communication means (for example, LAN, USB (Universal Serial Bus), RS-232C or the like).

[0020] Fig. 2 illustrates hardware of each of the communication control processors 11. As illustrated in Fig. 2, the communication control processor 11 includes a central processing unit 111, a memory 112, an external network interface (such as NIC or HBA)(hereinafter referred to as an external network I/F 113), and an internal network interface (hereinafter referred to as an internal network I/F 114).

[0021] The central processing unit 111 is a processor including a CPU, MPU, ASIC (Application Specific Integrated Circuit) or the like. The memory 112 includes a ROM or a RAM (that may be a NVRAM).

[0022] The external network I/F 113 is, for example, NIC, HBA or the like. The external network I/F 113 includes one or more network ports 115. Each of the network ports 115 is given an identifier (port ID) (for example, WWN (World Wide Name)).

[0023] The internal network I/F 114 communicates with the other components included in the storage apparatus 10 through the internal switch 16 in accordance with a certain communication standard (for example, PCI bus (PCI: Peripheral Component Interconnect)).

[0024] The communication control processor 11 receives I/O requests sent from the host apparatuses 3. For each I/O request received from any of the host apparatuses 3, the communication control processor 11 performs such as analyzing the protocol level, dis-

tributing the process to one of the main processors 12, storing write data accompanying a write request into the data storage buffer 14, sending read data to the host apparatus 3, and sending a response to the host apparatus 3.

[0025] In accordance with the content of the I/O request sent from the communication control processor 11, the main processor 12 performs an I/O process on a storage drive 171 to be described later (writes write data to the storage drive 171 or reads read data from the storage drive 171).

[0026] Fig. 3 illustrates hardware of the main processor 12. As illustrated in Fig. 3, the main processor 12 includes a core processor 121, a local memory 122, a data transfer device 123, and an internal network interface (hereinafter referred to as internal network I/F 124).

[0027] The core processor 121 is one of the core processors included in a multi-core type processor (hereinafter referred to as a multi-core processor) mounted in the storage apparatus 10. The local memory 122 includes a ROM and a RAM (that may also be a NVRAM).

[0028] The data transfer device 123 includes a high-speed data transfer chip such as a DMA (Direct Memory Access). The data transfer device 123 is incorporated in the core processor 121 in some cases.

[0029] The internal network I/F 124 communicates with the other components included in the storage apparatus 10 through the internal switch 16 in accordance with a certain communication standard (for example, PCI bus, PCI-X bus, or the like).

[0030] In response to the I/O request received by the communication control processor 11, the main processor 12 performs data transfer between the communication control processor 11 and the data storage buffer 14, and performs data transfer between the data storage buffer 14 and the storage device 17. The latter data transfer includes, for example, transfer of data stored in the storage device 17 to the data storage buffer 14 (staging) and transfer of data stored in the data storage buffer 14 to the storage device 17 (destaging).

[0031] The data storage buffer 14 illustrated in Fig. 1 is configured by using, for example, a RAM to which fast access is possible. The data storage buffer 14 stores therein write data received from the host apparatus 3 and read data loaded from the storage device 17.

[0032] The shared memory 15 stores therein various kinds of information to be used to control the storage apparatus 10 (for example, configuration information and operation information of the storage apparatus 10 (an operation rate (%) of each main processor 12 (an operation rate in a given period of time in the past), and the number of I/O requests currently running on each main processor 12 (the total number of I/O requests currently running on the main processor 12 and I/O requests registered in a process

queue of the main processor 12, which is hereinafter referred to as an I/O active count). Here, a user is allowed to access all or part of the information stored in the shared memory 15 through the maintenance device 18 from a management apparatus 7.

[0033] The internal switch 16 illustrated in Fig. 1 is, for example, a high-speed cross bar switch.

[0034] The storage device 17 illustrated in Fig. 1 includes a plurality of storage drives 171 that are physical storage media. The storage drives 171 are, for example, hard disk drives and semiconductor storage devices (SSD) in the formats of SAS (Serial Attached SCSI), SATA (Serial ATA), FC (Fibre Channel), PATA (Parallel ATA), SCSI, and the like. The storage devices 17 are housed in the same chassis as the storage apparatus 10 in some cases, or are housed in a chassis different from the chassis of the storage apparatus 10 in other cases.

[0035] Here, a cache memory for improving the performance of response to the host apparatuses 3, for example, may be provided between the data storage buffer 14 and the storage device 17. Instead, the data storage buffer 14 itself may be configured to function as a cache memory.

[0036] The storage apparatus 10 provides the storage areas of the storage drives 171 to the host apparatuses 3 in units of logical units (LU) each configured by allocating a given unit of storage area (hereinafter referred to as a physical page) of one of the storage drives 171, for example.

[0037] Each of the logical units is associated with a device file or a drive letter (drive name) in the host apparatus 3, for example. In addition, each of the logical units is given a logical unit number (LUN) that is an identifier unique to the logical unit. In this case, the host apparatus 3 specifies the logical unit by using the LUN. This LUN is set in the foregoing I/O request, for example.

[0038] Moreover, the storage apparatus 10 recognizes a physical storage area provided by the storage drive 171 (hereinafter referred to as a physical storage area) in units of logical devices (LDEV) that are storage areas logically set up by using the physical storage area. In this case, each of the logical devices is given a logical device number (LDEV number), as an identifier, which is unique to the logical device.

[0039] The storage areas of the logical devices are configured, for example, as storage areas of a RAID group that is formed by controlling a plurality of hardware devices in accordance with a RAID (Redundant Arrays of Inexpensive (or Independent) Disks) scheme (at least any one of RAIDs 0 to 6).

[0040] The maintenance device 18 (SVP: Service Processor) is configured by using a personal computer, for example. The maintenance device 18 performs such as status monitoring, configuration setting, collection of operation information, and storage of operation information into the shared memory 15 for the components of the storage

apparatus 10 (including at least the communication control processors 11, the main processors 12, the data storage buffer 14, the shared memory 15, and the internal switch 16). In addition, through communications with the management apparatus 7, the maintenance device 18 provides information to the management apparatus 7, controls the storage apparatuses 10 according to instructions from the management apparatus 7, or the like.

[0041] The management apparatus 7 is configured by using a personal computer, for example. The management apparatus 7 is communicatively coupled to the maintenance device 18 through a LAN or the like. The management apparatus 7 provides a user interface (GUI (Graphical User Interface) or CLI (Command Line Interface)) with which a user gives a job execution command to the storage apparatus 10, makes a configuration setting of the storage apparatus 10, makes an operation setting of the storage apparatus 10, or does the like.

[0042] <CCW>

An I/O request sent from the host apparatus 3 to the storage apparatus 10 includes a single command code chain that is a chain of command codes (hereinafter referred to as CCWs (Channel Command Word)) (the command code chain is hereinafter referred to as a CCW chain).

[0043] There are CCWs such as one including a write command forming a write request, one including a read command forming a read request, one including a control command such as a seek command to move the arm of a disk or a search command, and one including a sensing command that requests information on a component in the storage apparatus 10.

[0044] Fig. 4 illustrates a configuration of a CCW. As illustrated in Fig. 4, a CCW 410 includes information pieces of a command code 411, a flag 412, a data length 413, and accompanying information 414 (parameters and data) of the command code.

[0045] Among them, the command code 411 has set therein a command (a write command, a read command, a seek command, a search command, or the like).

[0046] The flag 412 has set therein supplemental information of the command set in the command code 411 or the like.

[0047] The data length 413 has set therein the data length of the data (read data, write data or the like) to be processed for the parameter of the command or the command set in the command code 411.

[0048] The accompanying information 414 of the command code has set therein information (e.g. seek address, search address, write data and the like) required for the command code 411.

[0049] <Basic Operation>

When receiving a CCW chain from any of the host apparatuses 3, the communication

control processor 11 in the storage apparatus 10 analyzes the protocol and thereafter sends the main processor 12 the CCWs 410 included in the CCW chain. Here, for example, when a CCW 410 thus sent includes a write command or a control command, the CCW 410 is accompanied by write data or the like. In this case, the communication control processor 11 stores the data accompanying the CCW into the data storage buffer 14, and informs the main processor 12 of the address of the storage location together with the CCW 410.

[0050] The main processor 12 performs processing according to the CCW 410 sent from the communication control processor 11.

[0051] In the case of a CCW 410 including a write command, for example, the main processor 12 writes the write data stored in the data storage buffer 14 to the storage device 17.

[0052] In the case of a CCW 410 including a read command, for example, the main processor 12 reads the read data from the storage device 17 and stores the data into the data storage buffer 14. Thereafter, the read data stored in the data storage buffer 14 is sent to the host apparatus 3 by the communication control processor 11.

[0053] Figs. 5 and 6 illustrate a basic operation of the storage apparatus 10 for a CCW chain 420 received from the host apparatus 3. Figs. 5 and 6 illustrate the basic operation that the storage apparatus 10 performs when receiving a CCW chain 420 including two CCWs 410 of "CCW1" and "CCW2" from the host apparatus 3.

[0054] Here, the following description is provided on the assumption that a "write command" to write "data1" of write data is set in the command code 411 of the "CCW 1" and that a "read command" instructing to read "data2" of read data is set in the command code 411 of the "CCW2."

[0055] As illustrated in Fig. 5, when receiving the CCW chain 420 (S511), the communication control processor 11 sends information on the CCW chain 420 to the main processor 12.

[0056] Since the write command is set in the "CCW1," for example, the communication control processor 11 sends the content of the "CCW1" to the main processor 12 (S512), and stores the write data "data1" accompanying the "CCW1" into the data storage buffer 14 (S513).

[0057] In addition, since the read command is set in the "CCW2," for example, the communication control processor 11 sends the content of the "CCW2" to the main processor 12 (S512).

[0058] Subsequently, as illustrated in Fig. 6, the main processor 12 stores the write data "data 1" stored in the data storage buffer 14 into the storage drive 171 according to the content of the "CCW1" (S611).

[0059] Then, the main processor 12 reads the read data "data2" from the storage drive 171

and stores (transfers) the read data into the data storage buffer 14 according to the content of the "CCW2" (S612) .

[0060] After that, the communication control processor 11 sends the host apparatus 3 the read data stored in the data storage buffer 14 and a completion notification of the I/O request (S613).

[0061] <CKD format>

The storage apparatus 10 of the present embodiment manages data sent from the mainframe host apparatuses 3 in accordance with the CKD format (CKD: Count Key Data format).

[0062] Fig. 7 illustrates a configuration of the CKD format. As illustrated in Fig. 7, the CKD format includes three information fields of a C field 711 (Count field), a K field 712 (Key field), and a D field 713 (Data field) (hereinafter these fields are also collectively referred to as data fields).

[0063] Among them, the C field 711 stores therein the location of the data (a cylinder number, head number, record number (CCHHR)), a data length of the K field 712, a data length of the D field 713 and the like.

[0064] The K field 712 stores therein information related to an application that runs in the host apparatus 3 (e.g., information on a search key), for example.

[0065] The D field 713 stores therein the entity of the write data or read data.

[0066] Here, the C field 711 is data in a fixed-length format (for example 8 Bytes), whereas both the K field 712 and the D field 713 are data in a variable-length format.

[0067] As illustrated in Fig. 8, the data in CKD format is managed in units of records and tracks.

[0068] Specifically, a piece of data in CKD format (data including a combination of one C field 711, one K field 712, and one D field 713 (the K field 712 and the D field 713, however, are not included in the CKD format in some cases)) is managed as one record 811. A set of a predetermined number of records 811 is managed as one track 812 including a header part 810. The predetermined number is set to an appropriate value depending on the configuration of the storage apparatus 10 or the type of the storage drives 171.

[0069] <Master-Slave relationship between Main Processors>

In the storage apparatus 10 of the present embodiment, the main processors 12 are classified into master processors and slave processors according to their respective roles.

[0070] When receiving the CCWs 410 sent from the communication control processor 11, the master processor sends a plurality of commands (hereinafter referred to as sub-commands) generated according to the contents of the CCWs 410 to a plurality of slave processors.

[0071] The slave processor performs processing according to the sub-command sent from the master processor, and returns a response or a processing result thereof to the master processor or the communication control processor 11.

[0072] Here, information indicating which of the master processor and the slave processor each of the main processors 12 currently serves as is stored and managed in, for example, the shared memory 15, the memory 112 of the communication control processor 11, in the local memory 122 of in the main processor 12, or the like. By use of this information, the communication control processor 11 and the main processor 12 can know which of the master processor and the slave processor each of the main processors 12 serves as.

[0073] Which of the main processors 12 serves as the master processor and which of the main processors 12 serves as the slave processor are determined by the communication control processor 11, the main processor 12, the maintenance device 18 (the management apparatus 7), or the like, statically (e.g., depending on the attribute or type of the storage drive 171 for which the main processor 12 is responsible) or dynamically (e.g., in view of load balancing based on the operation state of the storage apparatus 10).

[0074] <Read Sequence>

Fig. 9 illustrates a sequence performed in the information processing system 1 when the storage apparatus 10 receives a CCW 410 including a read command (hereinafter referred to as a read sequence S900). The read sequence S900 will be described below with reference to Fig. 9.

[0075] When receiving the CCW 410 including the read command from the host apparatus 3 (S911), the communication control processor 11 stores the content of the CCW 410 into the local memory 122 of the master processor (S912).

[0076] The master processor performs processing of determining a main processor 12 in charge of the process for the CCW 410 stored in the local memory 122 (hereinafter referred to as executor determination processing S913). The executor determination processing S913 will be described in details later.

[0077] Upon completion of the executor determination processing S913, the master processor generates information on the process distribution for the read command (hereinafter referred to as distribution information (Fig. 17)) on the basis of a return value of the executor determination processing S913, and sends the distribution information to the communication control processor 11 (S914). The distribution information will be described in details later.

[0078] Moreover, when the return value of the executor determination processing S913 indicates that a slave processor takes charge of the process for the read command, the master processor sends a sub-command to the slave processor in charge of the process,

the sub-command instructing the slave processor to execute the assigned process (S915, S916). Here, Fig. 9 illustrates a case where the return value of the executor determination processing S913 indicates that a "slave processor A" and a "slave processor C" take charge of the process.

[0079] In response to the sub-command sent from the master processor, each of the slave processors reads read data, which is to be transferred by the slave processor itself, from the storage drive 171 and stores the read data into the data storage buffer 14. In addition, the slave processor sends the communication control processor 11 a notification indicating that the read data is stored in the data storage buffer 14 (hereinafter referred to as a read data storage notification) (S917, S919).

[0080] Meanwhile, on the basis of the distribution information received in S914, the communication control processor 11 monitors in real time a storage state into the data storage buffer 14 for the read data to be returned as a response to the read command to the host apparatus 3. Then, when confirming that read data in a transfer unit to the host apparatus 3 is present in the data storage buffer 14, the communication control processor 11 sends the host apparatus 3 the read data in the transfer unit stored in the data storage buffer 14 (S920). The processing of the communication control processor 11 in this operation will be described in detail later (Fig. 18).

[0081] Since the slave processor sends the read data storage notification directly to the communication control processor 11 not via the master processor, the communication control processor 11 provides the read data to the host apparatus 3 quickly.

[0082] Returning Fig. 9, the slave processor sends the master processor a notification indicating the completion of the process for the sub-command (S918, S921).

[0083] When the notifications indicating the completion of the processes for the sub-commands are received from all the slave processors assigned the process for the read command (S918, S921), the master processor notifies the communication control processor 11 of the completion of the process for the read command (S922).

[0084] The communication control processor 11 sends the host apparatus 3 notification indicating the completion of the process for the read command (S923). In addition, the communication control processor 11 receives an acknowledgement of the receipt of the notification from the host apparatus 3 (S924).

[0085] Although the foregoing read sequence S900 is in the case where only two slave processors (the slave processor A and the slave processor C) take charge of the process for the read command, the master processor may take charge of part of the process for the read command together with the slave processors, instead. In this case, when storing read data to be processed by the master processor itself, the master processor sends the communication control processor 11 of the storage notification thereof (the read data storage notification). Then, the communication control processor 11 transfers

the read data to the host apparatus 3 on the basis of the read data storage notifications sent from the slave processors and the read data storage notification sent from the master processor. In addition, when receiving the completion notifications from the respective slave processors (S918, S921) and completing the part of the process for the read command assigned to the master processor itself, the master processor sends the completion notification to the communication control processor 11 (S922).

[0086] <Write Sequence>

Fig. 10 illustrates a sequence performed in the information processing system 1 when the storage apparatus 10 receives a CCW 410 including a write command (hereinafter referred to as a write sequence S1000). The write sequence S1000 will be described below with reference to Fig. 10.

[0087] When receiving the CCW 410 including the write command from the host apparatus 3 (S1011), the communication control processor 11 stores the content of the CCW 410 into the local memory 122 of the master processor (S1012). Moreover, the communication control processor 11 stores the write data accompanying the CCW 410 into the data storage buffer 14 (S1012).

[0088] The master processor performs processing of determining a main processor 12 in charge of the process for the CCW 410 stored in the local memory 122 (hereinafter referred to as executor determination processing S1013). The executor determination processing S1013 will be described in detail later.

[0089] When a return value of the executor determination processing S1013 indicates that a slave processor takes charge of the process for the write command, the master processor sends a sub-command to the slave processor in charge of the process, the sub-command instructing the slave processor to execute the assigned processing (S1014, S1015). Here, Fig. 10 illustrates a case where the result of the executor determination processing S913 indicates that only "slave processor A" and "slave processor C" take charge of the process.

[0090] In response to the sub-command sent from the master processor, each of the slave processors stores the write data, which is stored in the data storage buffer 14, into the storage drive 171 (S1014, S1015).

[0091] In addition, the slave processor sends the master processor a notification indicating that the write data is stored in the storage drive 171 (hereinafter referred to as a write data storage notification) (S1016, S1017).

[0092] When confirming that the write data storage notifications are received from all the slave processors taking charge of the process for the write command, the master processor sends a notification to that effect to the communication control processor 11 (S1018).

[0093] When receiving the notification, the communication control processor 11 sends a no-

tification of the completion of the process for the write command to the host apparatus 3 (S1019), and receives an acknowledgement of the receipt of the notification from the host apparatus 3 (S1020).

[0094] Although the foregoing write sequence S1000 is in the case where only two slave processors (the slave processor A and the slave processor C) take charge of the process for the write command, the master processor may take charge of part of the process for the write command together with the slave processors, instead. In this case, the master processor stores the write data to be processed by the master processor itself into the storage drive 171. Then, when receiving the write data storage notifications from all the slave processors (S1016, S1017) and completing the part of the process for the write command which the master processor itself takes charge of, the master processor sends the completion notification to the communication control processor 11 (S1018).

[0095] <Executor Determination Processing>

Fig. 11 is a flowchart for explaining the executor determination processing S913 in the read sequence S900 in Fig. 9. It should be noted that the steps in the executor determination processing S1013 in the write sequence S1000 in Fig. 10 are the same as those in the executor determination processing S913. This processing is performed by the master processor as a main actor, and will be described below with reference to Fig. 11.

[0096] To begin with, the master processor judges whether process target data of a command (a read command in the case of the read sequence S900 or a write command in the case of the write sequence S1000; the same applies below) set in the command code 411 of the CCW 410 is data managed over a plurality of tracks (S1111). Note that the master processor makes this judgment based on the type of the command set in the command code 411 of the CCW 410, for example.

[0097] If the process target data of the command is managed over the a plurality of tracks (S1111: YES), the master processor acquires the total number of data fields of the process target data of the command (S1121). Specifically, the master processor acquires the total number of data fields by using, for example, information contained in the CCW 410 or the CCW chain 420 (the number of records of the process target data of the command, and the number of tracks of the process target data of the command). Thereafter, the processing advances to S1131.

[0098] If the process target data of the command is not managed over a plurality of tracks (in other words, the data includes a single track) (S1111: NO), then the master processor judges whether the process target data is managed over a plurality of records (S1112). Note that the master processor makes this judgment based on the type of the command set in the command code 411 of the CCW 410, for example.

[0099] If the process target data of the command is managed over the a plurality of records

(S1112: YES), the master processor acquires the total number of data fields of the process target data of the command (S1121). Specifically, the master processor acquires the total number of data fields by using, for example, information contained in the CCW 410 or the CCW chain 420 (the number of records of the process target data of the command). Thereafter, the processing advances to S1131.

[0100] If the process target data of the command is not managed over the a plurality of records (in other words, the data includes a single record) (S1112: NO), then the master processor judges whether or not the process target data of the command includes a plurality of data fields (S1113). This judgment is made based on the CCW 410, for example.

[0101] If the process target data of the command includes multiple data fields (S1113: YES), then the master processor judges whether or not the process target data of the command contains the D field 1413 (S1125). If the process target data of the command contains the D field 1413(S1125: YES), the master processor acquires the total number (= 2 or 3) of data fields of the process target data of the command from the CCW 410 (S1126). Thereafter, the processing advances to S1131.

[0102] On the other hand, if the process target data of the command is judged as containing only a single data field in S1113 (S1113:NO), the master processor sets the return value of the executor determination processing S913 to "no distribution" (S1141). Then, the executor determination processing S913 is terminated.

[0103] Here, in the case where the return value is "no distribution," the master processor takes charge of the process for the command sent from the host apparatus 3, in principle, without sending the sub-commands to the slave processors in the read sequence S900 or the write sequence S1000.

[0104] If the process target data of the command is judged as not containing the D field 1413 in S1125 (S1125: NO), the processing advances to S1141. Here, for the process target data of the command not containing the D field 1413, "no distribution" is set as the return value of the executor determination processing S913. This is because, when the data contains only two data fields (the C field 1411 and the K field 1412) (not containing the D field 1413), the processing load of the main processor 12 is small, so that the process distribution cannot be expected to produce much effect.

[0105] In S1131, the master processor performs processing of selecting a slave processor in charge of the process (hereinafter referred to as executor selection processing S1131). The executor selection processing S1131 will be described in detail later.

[0106] After completion of the executor selection processing S1131, the master processor sets the return value to a result of the executor selection processing S1131 (S1132). Then, the processing returns to the read sequence S900 or the write sequence S1000.

[0107] <Executor Selection Processing S1131>

Fig. 12 is a flowchart for explaining the details of the executor selection processing S1131 in Fig. 11. The executor selection processing S1131 will be described below with reference to Fig. 12.

[0108] From the operation information stored in the shared memory 15, the master processor firstly acquires the I/O active count of each of slave processors that can be selected as a process allocation target (hereinafter referred to as a selectable slave processor) (S1211). Here, information indicating a relationship between a master processor and slave processors available to the master processor is stored and managed in the shared memory 15, for example. The master processor accesses the shared memory 15 when needed, and acquires information on which main processor 12 is a selectable slave processor.

[0109] Fig. 13 illustrates an example of the operation information. In operation information 1300, an operation rate (%) 1312 and an I/O active count (load indicator) 1313 are managed in association with an identifier 1311 (main processor ID) of each main processor 12, as illustrated in Fig. 13.

[0110] Returning to Fig. 12, the master processor judges whether or not there is a selectable slave processor having an I/O active count of "0" (i.e., a processor currently not being running any process) among the selectable slave processors (S1212). If there is a selectable slave processor having an I/O active count of "0" (S1212: YES), the processing advances to S1221. Instead, if there is no selectable slave processor having an I/O active count of "0" (S1212: NO), the processing advances to S1213.

[0111] In S1221, the master processor acquires the latest operation rate of each of the selectable slave processors from the shared memory 15 (S1221). Then, the master processor performs distribution target selection processing S1222 and distribution processing S1223 by using the acquired operation rates, and sets a result of the distribution processing S1223 as the return value of the executor selection processing S1131 (S1224). Thereafter, the executor selection processing S1131 is terminated. The distribution target selection processing S1222 and the distribution processing S1223 will be described in details later.

[0112] On the other hand, in S1213, the master processor compares its own I/O active count with the I/O active count of each of the selectable slave processors. If its own I/O active count is larger than the I/O active counts of all the selectable slave processors (S1213: NO), the processing advances to S1221.

[0113] If its own I/O active count is equal to or smaller than the I/O active count of any of the selectable slave processors as a result of the comparison (S1213: YES), the master processor sets "no distribution" as the return value (S1214). Then, the executor selection processing S1131 is terminated.

[0114] In this way, "no distribution" is set as the return value when the I/O active count of

the master processor is equal to or smaller than the I/O active count of any of the selectable slave processors as a result of the comparison (in other words, set so that only the master processor performs the processing for the command). This is because, if the master processor currently under a low processing load itself performs the I/O process, the performance improvement of the storage apparatus 10 can be expected with appropriate load balancing.

[0115] <Distribution Target Selection Processing S1222>

Fig. 14 is a flowchart for explaining the details of the distribution target selection processing S1222 in Fig. 12.

[0116] Firstly, the master processor judges whether or not there is a slave processor having an I/O active count of "0" (S1411). In the case where there is a slave processor having an I/O active count of "0," the entire selectable slave processors are understood to operate under a low load.

[0117] If there is a slave processor having an I/O active count of "0" among the selectable slave processors (S1411: YES), the master processor prioritizes the selectable slave processors in ascending order of the operation rate (S1421). Since the master processor prioritizes the selectable slave processors in ascending order of the operation rate in this way, the master processor can allocate I/O processes to the slave processors having low operation rates preferentially, which in turn leads to appropriate load balancing among the main processors 12.

[0118] On the other hand, if there is no slave processor having an I/O active count of "0" (S1411: NO), the master processor prioritizes the selectable slave processors in ascending order of the I/O active count (S1412). Thereafter, the processing advances to S1413. Since the master processor prioritizes the selectable slave processors in ascending order of the I/O active count in this way, the master processor can allocate I/O processes to slave processors currently under a low processing load preferentially, which in turn leads to appropriate load balancing among the main processors 12.

[0119] In S1413, the master processor compares the number of data fields of the process target data of the command (the total number of data fields acquired in S1121 or S1126 in Fig. 11) with the total number of selectable slave processors.

[0120] If the total number of data fields of the process target data of the command is equal to or larger than the total number of selectable slave processors as a result of the comparison (S1413: YES), the master processor selects all the selectable slave processors available to the master processor as distribution targets (S1422). Then, the distribution target selection processing S1222 is terminated.

[0121] On the other hand, if the total number of data fields of the process target data of the command is smaller than the total number of selectable slave processors as a result of the comparison (S1413: NO), the master processor selects, as distribution targets, the

same number of selectable slave processors as that of data fields of process target data in descending order of the priority (S1414). Then, the distribution target selection processing S1222 is terminated.

[0122] <Distribution Processing S1223>

Fig. 15 is a flowchart for explaining the details of the distribution processing S1223 in Fig. 12.

[0123] Firstly, the master processor acquires the content of the C field of each record of the process target data of the command (the cylinder number/head number/record number (CCHHR), the data length (KL) of the K field 1412, and the data length (DL) of the D field 1413) (S1511).

[0124] Subsequently, if the command is a command to request a transfer of data in the C field 1411 to the host apparatus 3, the master processor subtracts the number of the C fields 1411 from the total number of data fields of the process target data (the number of data fields acquired in S1211 or S1126 in Fig. 11) (S1512). In this way, if the command is a command to request a transfer of data in the C field 1411 to the host apparatus 3, the number of C fields 1411 is subtracted from the total number of data fields of the process target data for the following reason. Specifically, even when the processes for the C fields 1411 are allocated to the slave processors, the master processor always needs to read the data from the C fields 1411 (needs the information in the C fields 1411 to generate sub-commands to the slave processors). Hence, it is not preferable to distribute the processes for the C fields 1411 to other processors (the effect of load balancing is low even though the processes are distributed to the slave processors).

[0125] Thereafter, the master processor performs processing of determining which slave processor will take charge of the process for each data field (hereinafter referred to as allocation processing S1513). The allocation processing S1513 will be described in details later.

[0126] Upon completion of the allocation processing S1513, the master processor generates the distribution information on the basis of a return value of the allocation processing S1513 (an allocation ratio of data fields to be allocated to the selectable slave processors). Then, the distribution processing S1223 is terminated.

[0127] <Allocation Ratio Determination Processing S1513>

Fig. 16 is a flowchart for explaining the details of the allocation processing S1513 in the distribution processing S1223 in Fig. 15.

[0128] Firstly, the master processor compares the number of tracks of the process target data of the command and the total number of selectable slave processors (S1611).

[0129] If the number of tracks of the process target data of the command is equal to or larger than the total number of selectable slave processors as a result of the comparison

(S1611: YES), the master processor determines that an allocation unit of data fields of the process target data is a result obtained by dividing the total number of data fields of the process target data of the command by the number of tracks to be processed (in other words, divides the total number of data fields to be processed by the number of tracks, and allocates the I/O processes for data fields in units of the resulting number to the selectable slave processors) (S1613). Thereafter, the processing advances to S1621.

[0130] On the other hand, if the number of tracks of the process target data of the command is smaller than the total number of selectable slave processors as a result of the comparison (S1611: NO), then the master processor compares the number of records of the process target data of the command with the total number of selectable slave processors (S1612).

[0131] If the number of records of the process target data of the command is equal to or larger than the total number of selectable slave processors as a result of the comparison (S1612: YES), the master processor determines that the allocation unit of data fields of the process target data is a result obtained by dividing the total number of data fields of the process target data of the command by the number of records to be processed (in other words, divides the total number of data fields to be processed by the number of records to be processed, and allocates the I/O processes for data fields in units of the resulting number to the selectable slave processors) (S1614). Thereafter, the processing advances to S1621.

[0132] Instead, if the number of records of the process target data of the command is smaller than the total number of selectable slave processors as a result of the comparison (S1612: NO), the master processor determines that the distribution unit of data fields of the process target data is each of data fields to be processed (that is, allocates the I/O processes for data fields on a data-field by data-field basis to the selectable slave processors (S1615). Thereafter, the processing advances to S1621.

[0133] Here, a specific example is provided on the assumption that the number of data fields is 42, the number of records is 21, the number of tracks is 2, and the total number of selectable slave processors is 4, for instance. In this case, the number of tracks is smaller than the total number of selectable slave processors (S1611: NO) and the number of records is equal to or greater than the total number of selectable slave processors (S1612: YES) hold true. Hence, the divisor is the number of records, i.e., 21, and the number of data fields, i.e., 42 is divided by 21 to obtain a result of 2. As a result, the I/O processes for the data fields are allocated to the selectable slave processors in units of the resulting number (=2) of data fields, i.e., two data fields.

[0134] In this way, the master processor allocates the selectable slave processors the I/O processes for data fields to be processed in units of the same number of data fields as a result obtained by using the number of tracks as the divisor if the number of tracks of

the process target data is equal to or larger than the total number of selectable slave processors, or a result obtained by using the number of records as the divisor if the number of records of the process target data is equal to or larger than the total number of selectable slave processors. Thus, the master processor can efficiently and appropriately allocate the data fields to the selectable slave processors depending on the number of pieces of the process target data.

[0135] In S1621, the master processor allocates the data fields to the selectable slave processors in units of a result obtained by the divisor determined in the processing in S1611 to S1615, by determining how many units are to be allocated to the selectable processors on the basis of their respective operation rates. Then, the allocation ratio determination processing S1513 is terminated.

[0136] Here, a specific example is provided on the assumption that the total number of selectable slave processors is 4 (hereinafter also referred to as selectable slave processors #1 to #4 in a distinguishable manner), and the operation rates of the slave processors #1 to #4 are 5%, 15%, 10%, and 10%, respectively. In this case, the master processor allocates the selectable slave processors #1 to #4 the I/O processes for the data fields in inverse ratio (6:2:3:3) to the ratio of the operation rates.

[0137] Thus, since the master processor allocates more data fields to a slave processor having a lower operation rate as described above, the master processor can equalize the loads on the respective slave processors and achieve appropriate load balancing.

[0138] <Distribution Information>

In the read sequence S900 in Fig. 9, the master processor notifies the communication control processor 11 of the distribution information (S914) after the completion of the executor determination processing S913. This distribution information is generated by the master processor on the basis of the return value of the executor determination processing S913, and is notified to the communication control processor 11 by the master processor. Then, the communication control processor 11 organizes read data for the read command on the basis of the allocation information notified by the master processor, and transmits the organized read data to the host apparatus 3 when needed.

[0139] Fig. 17 illustrates an example of the distribution information. Distribution information 1700 is generated for each main processor 12 that takes charge of a process for read data.

[0140] A read portion 1711 has set information indicating which one of record or track is handled by the distribution information 1700. In Fig. 17, the allocation information 1700 is generated for a record.

[0141] A read number 1712 has set information indicating how many records or tracks are handled by the information managed in the distribution information 1700. In Fig. 17, the allocation information 1700 is for two records.

[0142] A CCW number 1713 has set information indicating which CCW 410 is handled by the information managed in the allocation information 1700 (hereinafter such information is referred to as a CCW number). In Fig. 17, the distribution information 1700 is for a CCW 410 having a CCW number "x0300."

[0143] A sequence number 1714 has set an identifier assigned to each of sequences generated to process the command in the CCW 410 set in the CCW number 1713 (hereinafter such identifier is referred to as a sequence number). In Fig. 17, the distribution information 1700 is for a sequence with a sequence number "1."

[0144] A storage destination address 1715 has set information (address) indicating a location in the data storage buffer 14 where the read data read from the storage drive 171 by the sequence set in the sequence number 1714 is to be stored. In Fig. 17, the data read from the storage drive 171 by the sequence is stored at an address "x48000000" in the data storage buffer 14.

[0145] A C field information 1716 has stored the information in the C fields 1411 of the records to be processed based on the distribution information 1700. In Fig. 17, the distribution information 1700 is for two records, and the C field information 1716 has set the information in the C field 1411 of each of the two records.

[0146] <Processing by Communication Control Processor for Read Command Process>

As illustrated in Fig. 9, after storing the read data into the data storage buffer 14, the slave processor directly notifies the communication control processor 11 of the completion of the storage (read data storage notification) (S918, S919).

[0147] On the basis of the allocation information 1700 received in S914, the communication control processor 11 monitors in real time the storage state in the data storage buffer 14 for the read data to be returned as a response to the read command to the host apparatus 3. The communication control processor 11 transfers the read data stored in the data storage buffer 14 to the host apparatus 3 sequentially on the basis of the foregoing read data storage notifications.

[0148] Fig. 18 is a flowchart for explaining processing performed by the communication control processor 11 for the above transfer of the read data to the host apparatus 3 (hereinafter referred to as read data transfer processing S1800).

[0149] Upon receipt of the read data storage notification sent from the slave processor (S1811: YES), the communication control processor 11 searches for an entry matched with the CCW number included in the read data storage notification (S1812).

[0150] Fig. 19 illustrates an example of the entry managed by the communication control processor 11. Here, an entry 1900 is stored, for example, in the memory 112 of the communication control processor 11, the shared memory 15, or the like. The communication control processor 11 generates the entry 1900 for each CCW number on the basis of the distribution information 1700.

[0151] In Fig. 19, a CCW number 1911 has set a CCW number. A data length 1912 has set the total data length of read data specified in the CCW.

[0152] A sequence number 1913 has set the total number of sequences described above (sequences generated to process the command in the CCW 410).

[0153] A per-sequence data length 1914 has set a storage state of read data in the data storage buffer 14 for each sequence (each sequence is identified by a sequence number 19141) (i.e., indicates a data length of the read data already stored in the data storage buffer 14).

[0154] A sent data length 1915 has set a transmission state of the read data specified in the CCW 410 to the host apparatus 3 (the total data length of the read data already sent to the host apparatus 3). A data storage buffer address 1916 has set an address of the data storage buffer 14 allocated to the CCW 410 (the address at which the read data of the CCW 410 is stored).

[0155] Returning to Fig. 18, the communication control processor 11 searches out the matched entry 1900, and updates the per-sequence data length 1914 (S1813) on the basis of the data length of the read data already stored in the data storage buffer 14 for each sequence number, the data length contained in the read data storage notification.

[0156] Subsequently, the communication control processor 11 judges whether or not the data length of read data unsent to the host apparatus 3 in the read data already stored in the data storage buffer 14 is longer than the unit for transferring read data to the host apparatus 3 (hereinafter referred to as a transfer length) (S1814).

[0157] Note that the communication control processor 11 obtains the data length of read data unsent to the host apparatus 3 in the read data already stored in the data storage buffer 14, by subtracting the sent data length 1915 from the total sum of the data lengths of data already stored for the sequences in per-sequence data length 1914 in the entry 1900.

[0158] If the data length of unsent read data is longer than the transfer length as a result of the judgment (S1814: YES), the communication control processor 11 sends the host apparatus 3 the unsent read data stored in the data storage buffer 14 of a transfer length (or an integral multiple of the transfer length) (S1815). Then, the communication control processor 11 adds the data length of the data transferred in S1815 to the sent data length 1915 of the entry 1900 (S1816). Then, the processing returns to S1811.

[0159] On the other hand, if the data length of unsent read data is not longer than the transfer length as a result of the judgment (S1814: NO), the communication control processor 11 judges whether or not the read data for all the sequences of the entry 1900 has been received (S1821). If the read data for all the sequences of the entry 1900 has not been received yet (S1821: NO), the processing returns to S1811.

[0160] Instead, if the read data for all the sequences of the entry 1900 has been already

received (S1821: YES), the processing advances to S1822, and the communication control processor 11 sends the remaining unsent read data to the host apparatus 3 (S1822). Thereafter, the read data transfer processing S1800 for the entry 1900 is terminated.

[0161] As described above, the storage apparatus 10 causes the slave processors to share processes in the units of data fields (the C field, the K field, and the D field), which form a single record in CKD format, and thereby achieves load balancing of I/O processes in finer units than in a conventional method. In addition, the distribution of I/O processes in the units of data fields can produce load balancing effects even when an I/O request for only one record is received. In the case where a large number of I/O requests each targeting for only one record are sent from the host apparatuses 3, in particular, the distribution in the foregoing way can make the load balancing function effectively.

[0162] Moreover, this enables the load balancing among core processors (main processors 12) in a multi-core processor, and thus brings the benefit from the use of the multi-core processor.

[0163] Although the preferred embodiments of the present invention have been heretofore described, these embodiments are only for illustrative purposes and are not intended to limit the scope of the present invention to the embodiments. The present invention can be implemented as various other embodiments.

Claims

[Claim 1] A storage apparatus comprising:
a communication control processor that is a processor configured to control communications with an external apparatus; and
a plurality of main processors each configured to perform an I/O process on a storage drive according to an I/O request received from the external apparatus,
wherein the storage apparatus
manages data to be stored or data that has been stored in the storage drive in accordance with CKD format, and
distributes the I/O process to a plurality of the main processors in units of data fields on basis of operation rates of the respective main processors and information on a C field, a K field, and a D field that are data fields forming a record of data targeted by the I/O process and managed in CKD format.

[Claim 2] The storage apparatus according to claim 1, wherein
the storage apparatus distributes the I/O process to the plurality of main processors on basis of the operation rates, the information on the C field, the K field, and the D field that are the data fields forming the record of data targeted by the I/O process and managed in the CKD format, and an I/O load indicator being a load indicator of the I/O process currently running on each of the main processors.

[Claim 3] The storage apparatus according to claim 2, wherein
the storage apparatus sets a destination of the I/O process distribution preferentially in ascending order of the operation rate or the I/O load indicator of the main processor.

[Claim 4] The storage apparatus according to claim 2, wherein the storage apparatus does not perform the distribution when the I/O load indicators of all the main processors other than the main processor receiving the I/O request are not 0, and when the I/O load indicator of the main processor receiving the I/O request is equal to or smaller than those of all the other main processors.

[Claim 5] The storage apparatus according to claim 1, wherein the storage apparatus does not have the I/O process for a record not including a D field as a subject to the distribution .

[Claim 6] The storage apparatus according to claim 1, wherein when the I/O request is a request to read a C field, the storage apparatus does not

have the I/O process for the C field to be read as a subject to the distribution.

[Claim 7] The storage apparatus according to claim 1, wherein the storage apparatus
obtains a total number of the data fields to be processed for the I/O process subjected to the distribution,
the storage apparatus obtains a number of the records and a number of tracks that is a set of a given number of the records based on the total number,
the storage apparatus distributes the I/O process for the data fields in units of a same number of data fields as a result obtained by dividing a total number of the data fields by the number of tracks when the number of tracks is equal to or larger than the total number of the main processors available for distribution of the I/O process,
the storage apparatus distributes the I/O process for the data fields in units of the same number of data fields as a result obtained by dividing the total number of data fields by a number of the records when the number of tracks is smaller than a total number of the main processors available for distribution of the I/O process, and when the number of records is equal to or larger than the total number of the main processors available for the distribution of the I/O process, and
the storage apparatus distributes the I/O process for the data fields in units of each data field when the number of records is smaller than the total number of the main processors available for the allocation of the I/O process.

[Claim 8] The storage apparatus according to claim 1, wherein
when receiving the I/O request, the communication control processor notifies a content of the I/O request to a master processor that is one of the main processors, and
the master processor distributes the I/O process to a plurality of slave processors that are other ones of the main processors.

[Claim 9] The storage apparatus according to claim 7, further comprising
a data storage buffer configured to store therein read data to be read from the storage drive and sent to the external apparatus, wherein
the slave processor after storing the read data in the data storage buffer in response to the I/O process distributed by the master processor, directly sends the communication control processor a notification that the read data is stored, and

the communication control processor transfers the read data to the external apparatus in response to the notification.

[Claim 10] The storage apparatus according to claim 1, wherein each of the main processors is one of core processors configuring a multi-core processor.

[Claim 11] A method of controlling a storage apparatus including a communication control processor that is a processor configured to control communications with an external apparatus, and a plurality of main processors each configured to perform an I/O process on a storage drive according to an I/O request received from the external apparatus, the storage apparatus being configured to manage data to be stored or data that has been stored in the storage drive in accordance with CKD format, the method comprising: distributing the I/O process to a plurality of the main processors by the storage apparatus in units of data fields on the basis of operation rates of the respective main processors and information on a C field, a K field, and a D field that are data fields forming a record of data targeted by the I/O process and managed in CKD format.

[Claim 12] The method of controlling a storage apparatus according to claim 11, wherein the storage apparatus distributes the I/O process to the plurality of main processors on basis of the operation rates, the information on the C field, the K field, and the D field that are the data fields forming the record of data targeted by the I/O process and managed in the CKD format, and an I/O load indicator being a load indicator of the I/O process currently running on each of the main processors.

[Claim 13] The method of controlling a storage apparatus according to claim 11, wherein the storage apparatus does not have the I/O process for a record not including a D field as a subject to the distribution.

[Claim 14] The method of controlling a storage apparatus according to claim 11, wherein when the I/O request is a request to read a C field, the storage apparatus does not have the I/O process for the C field to be read as a subject to the distribution.

[Claim 15] The method of controlling a storage apparatus according to claim 11, wherein

when receiving the I/O request, the communication control processor notifies a content of the I/O request to a master processor that is one of the main processors,

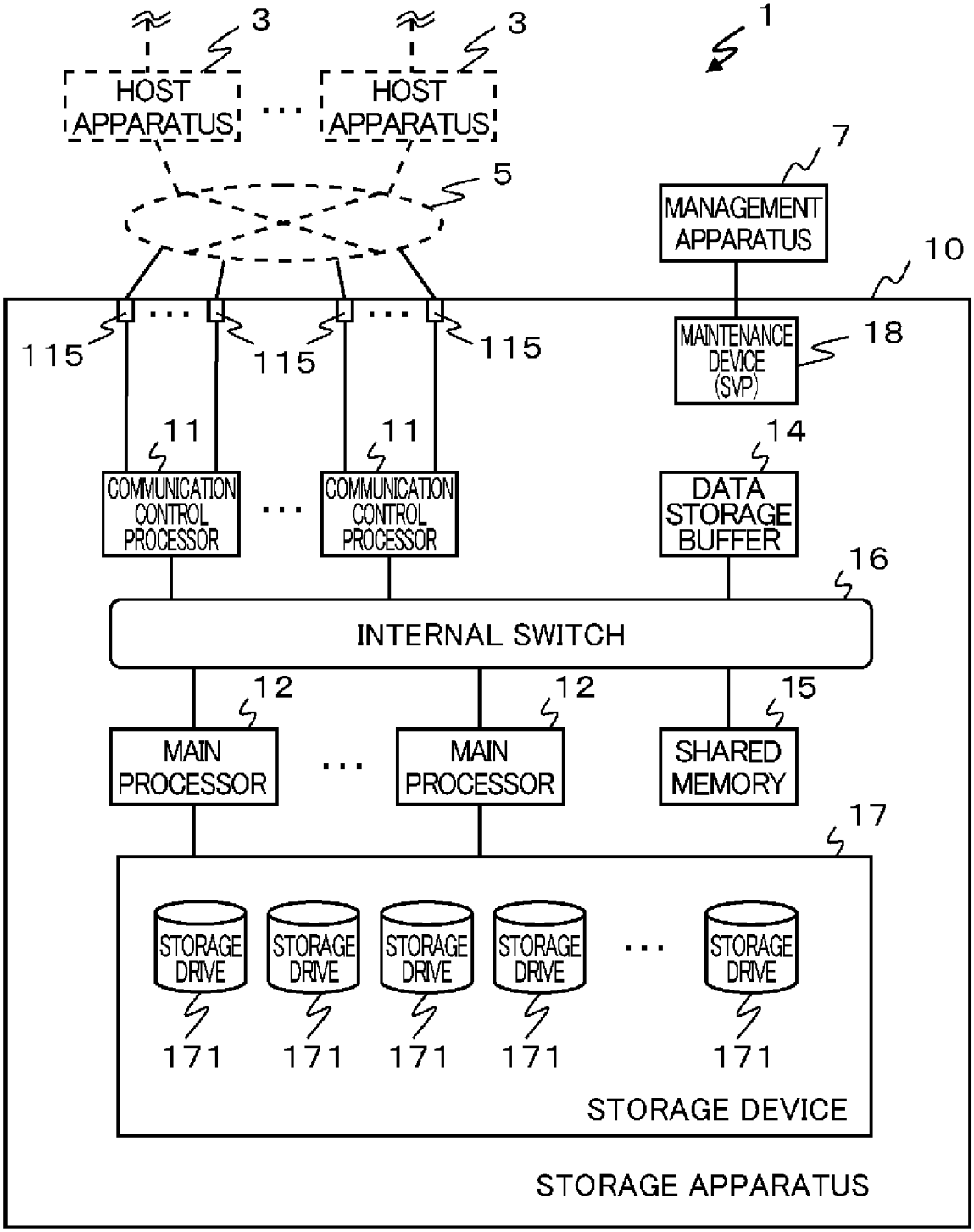
the master processor distributes the I/O process to a plurality of slave processors that are other ones of the main processors,

the storage apparatus includes a data storage buffer configured to store therein read data to be read from the storage drive and sent to the external apparatus,

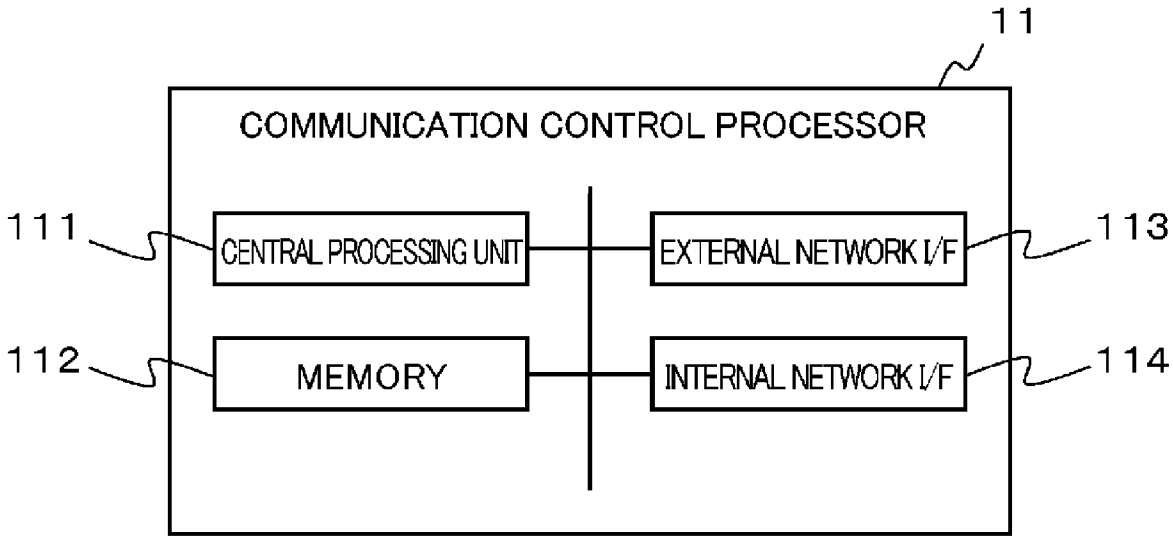
the slave processor after storing the read data in the data storage buffer in response to the I/O process distributed by the master processor, directly sends the communication control processor a notification that the read data is stored, and

the communication control processor transfers the read data to the external apparatus in response to the notification.

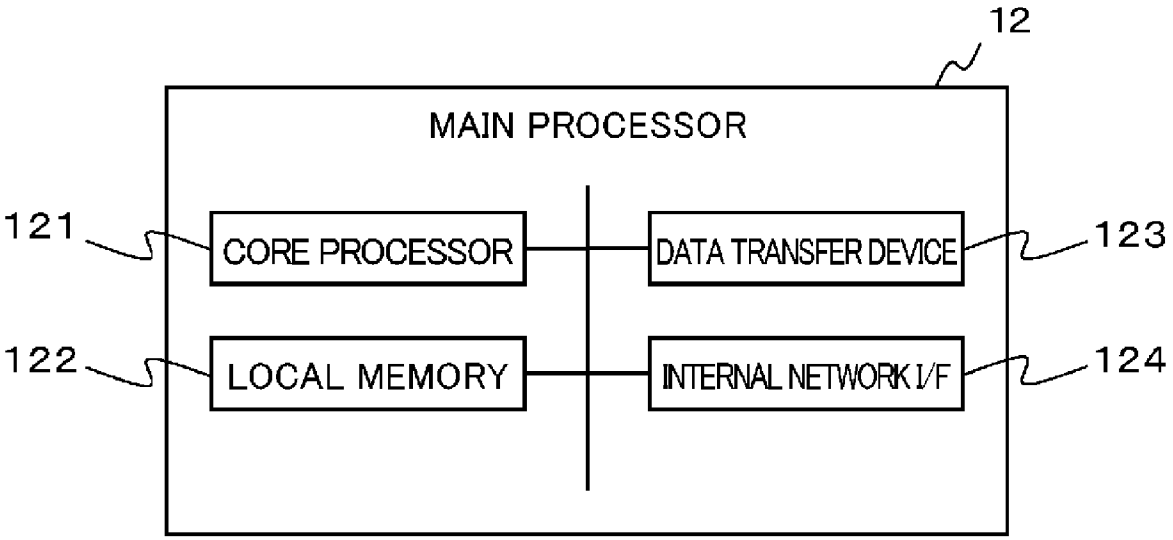
[Fig. 1]


1
3
3
HOST APPARATUS
HOST APPARATUS
5
7
MANAGEMENT APPARATUS
10
115
115
115
MAINTENANCE DEVICE (SVP)
18
11
11
14
COMMUNICATION CONTROL PROCESSOR
COMMUNICATION CONTROL PROCESSOR
DATA STORAGE BUFFER
16
INTERNAL SWITCH
12
12
15
MAIN PROCESSOR
MAIN PROCESSOR
SHARED MEMORY
17
STORAGE DRIVE
STORAGE DRIVE
STORAGE DRIVE
STORAGE DRIVE
STORAGE DRIVE
171
171
171
171
171
STORAGE DEVICE
STORAGE APPARATUS

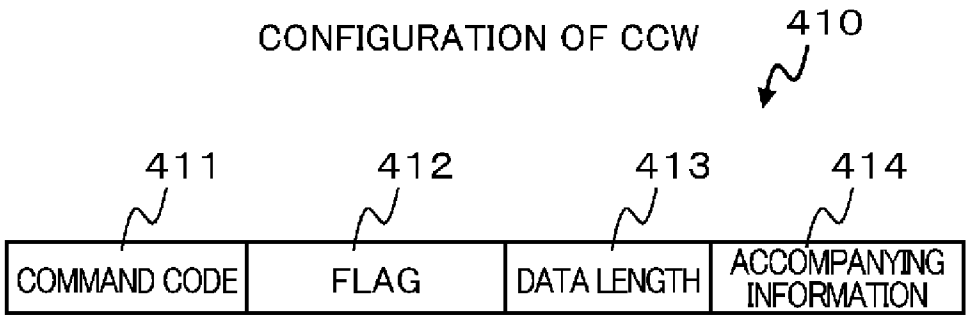
[Fig. 2]



[Fig. 3]

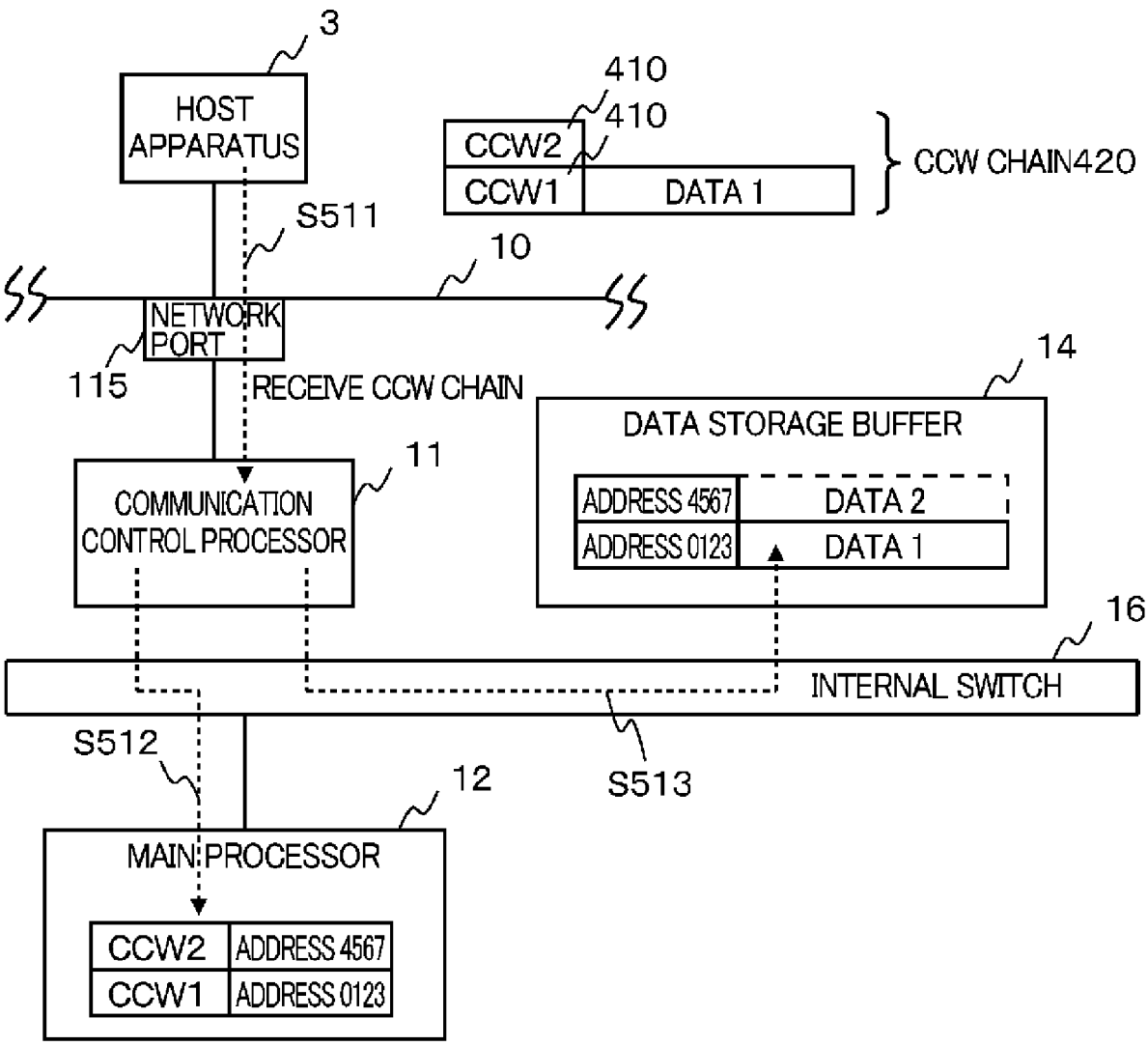


[Fig. 4]



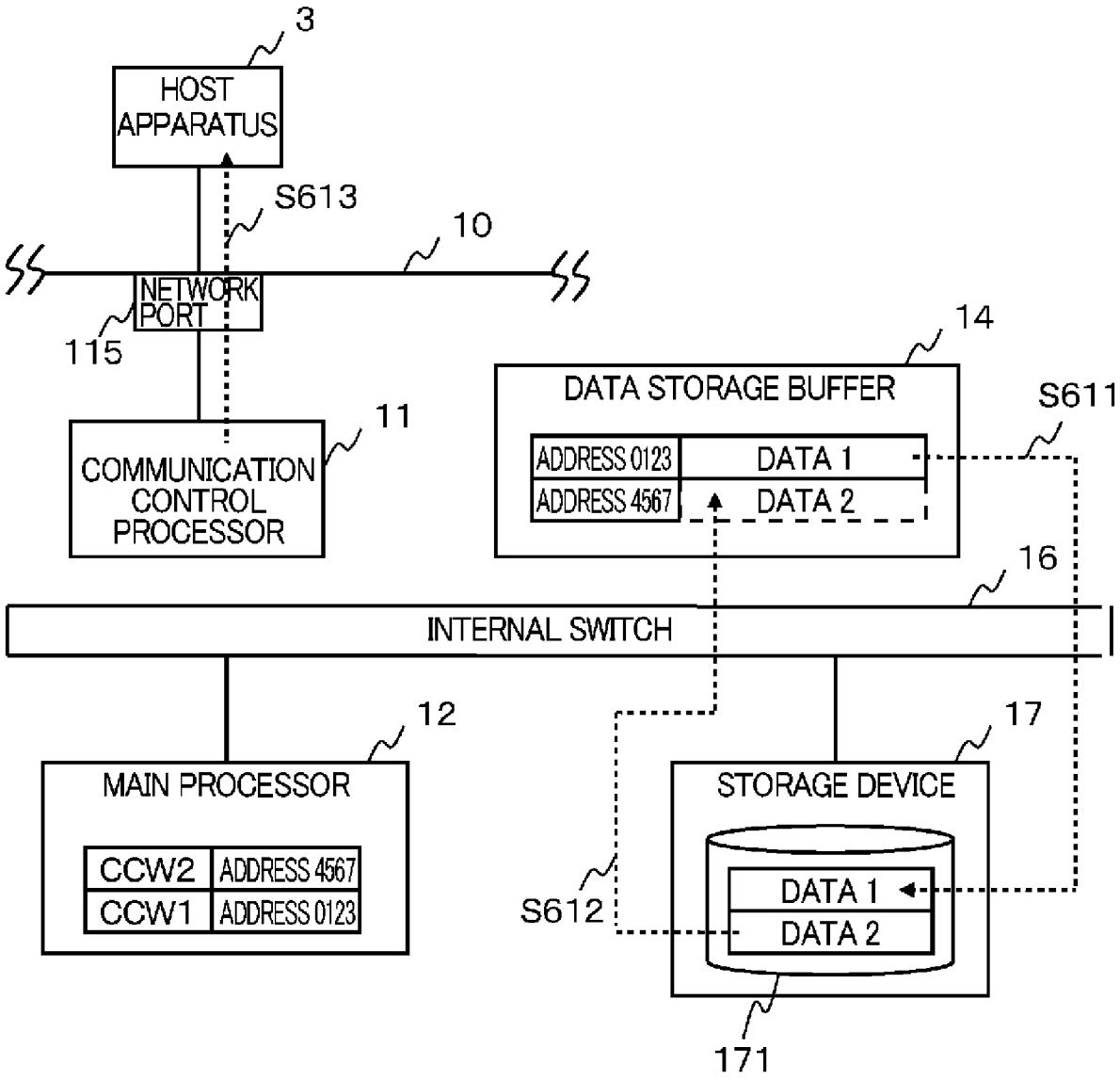
[Fig. 5]

BASIC OPERATION OF STORAGE APPARATUS



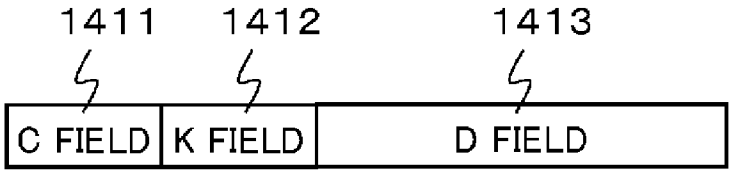
[Fig. 6]

BASIC OPERATION OF STORAGE APPARATUS

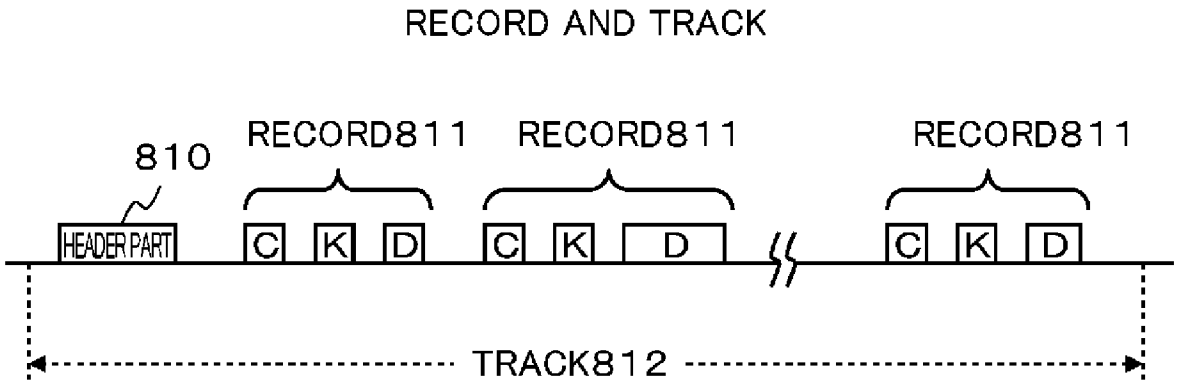


[Fig. 7]

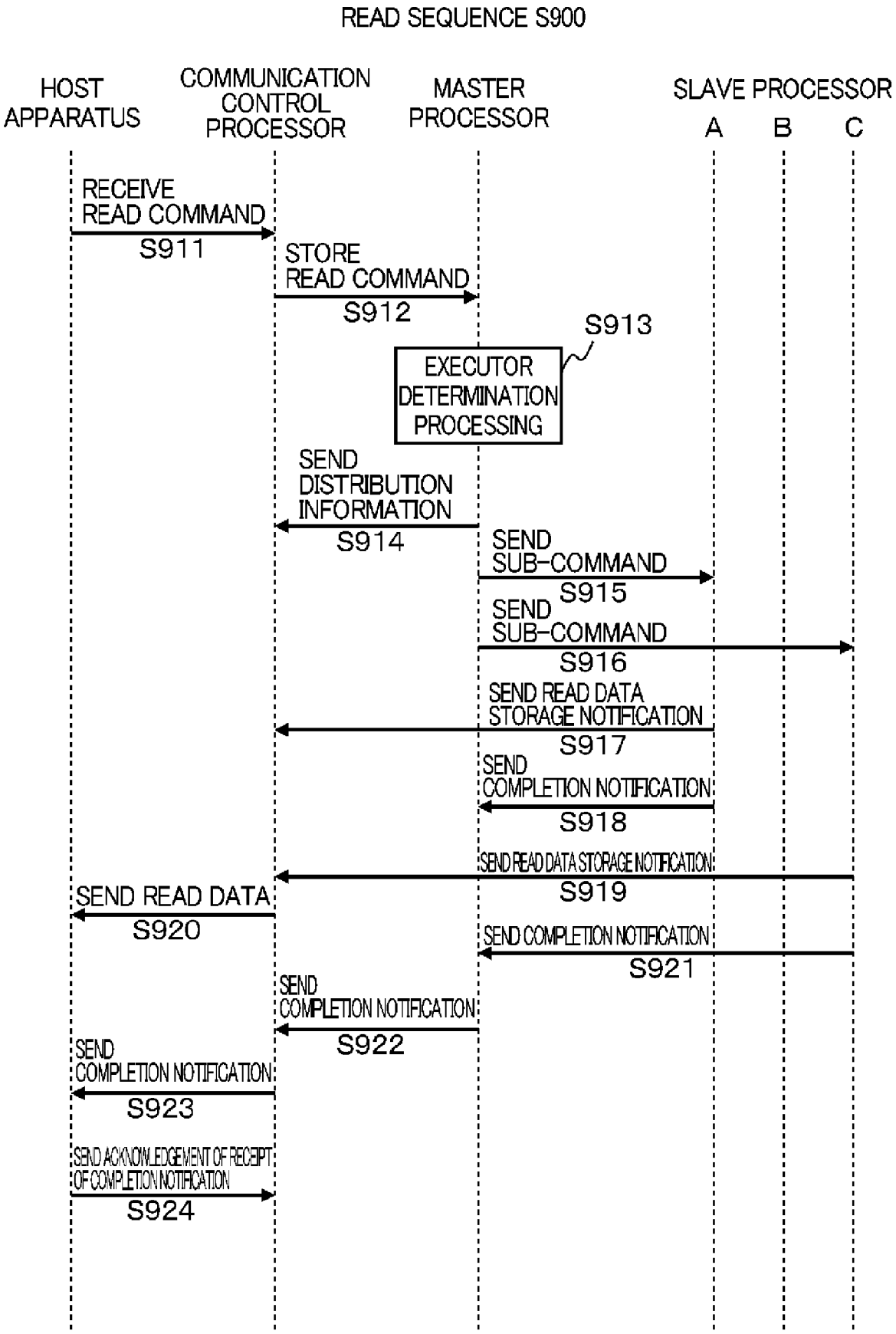
CKD FORMAT



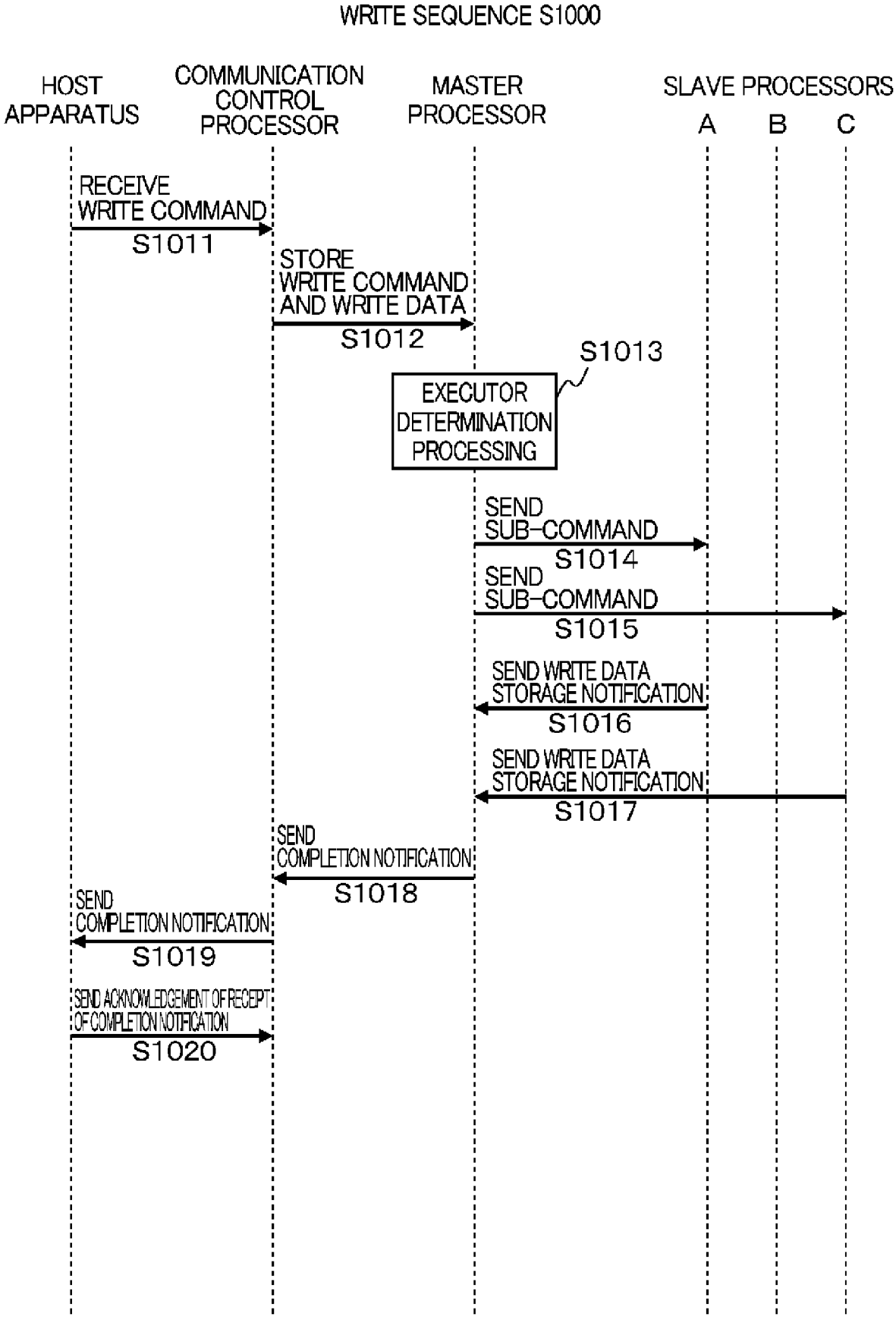
[Fig. 8]


RECORD AND TRACK
810
RECORD811
RECORD811
RECORD811
HEADER PART
C
K
D
C
K
D
C
K
D
TRACK812

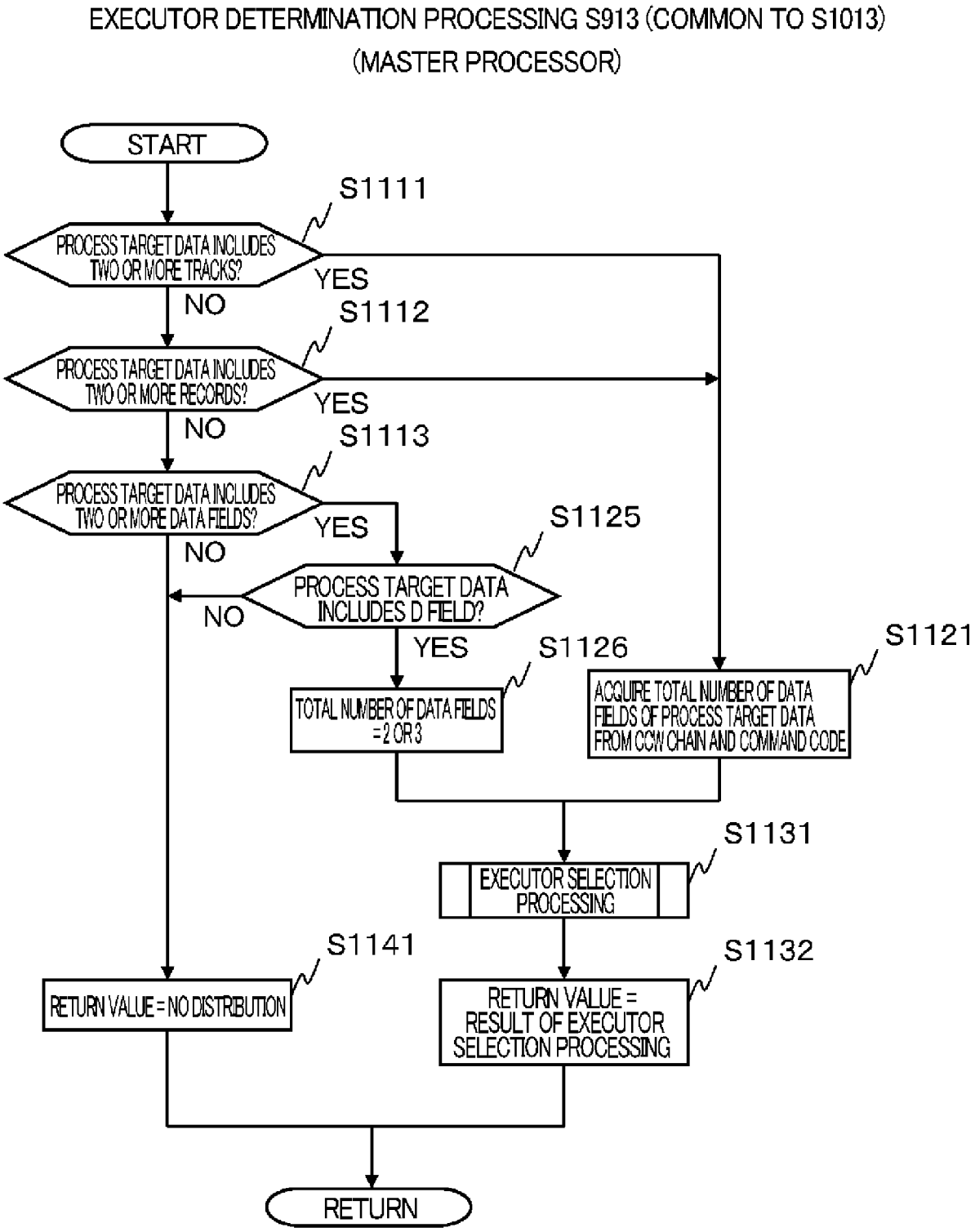
[Fig. 9]


READ SEQUENCE S900
HOST APPARATUS
COMMUNICATION CONTROL PROCESSOR
MASTER PROCESSOR
SLAVE PROCESSOR
A
B
C
RECEIVE READ COMMAND
S911
STORE READ COMMAND
S912
S913
EXECUTOR DETERMINATION PROCESSING
SEND DISTRIBUTION INFORMATION
S914
SEND SUB-COMMAND
S915
SEND SUB-COMMAND
S916
SEND READ DATA STORAGE NOTIFICATION
S917
SEND COMPLETION NOTIFICATION
S918
SEND READ DATA STORAGE NOTIFICATION
S919
SEND READ DATA
S920
SEND COMPLETION NOTIFICATION
S921
SEND COMPLETION NOTIFICATION
S922
SEND COMPLETION NOTIFICATION
S923
SEND ACKNOWLEDGEMENT OF RECEIPT OF COMPLETION NOTIFICATION
S924

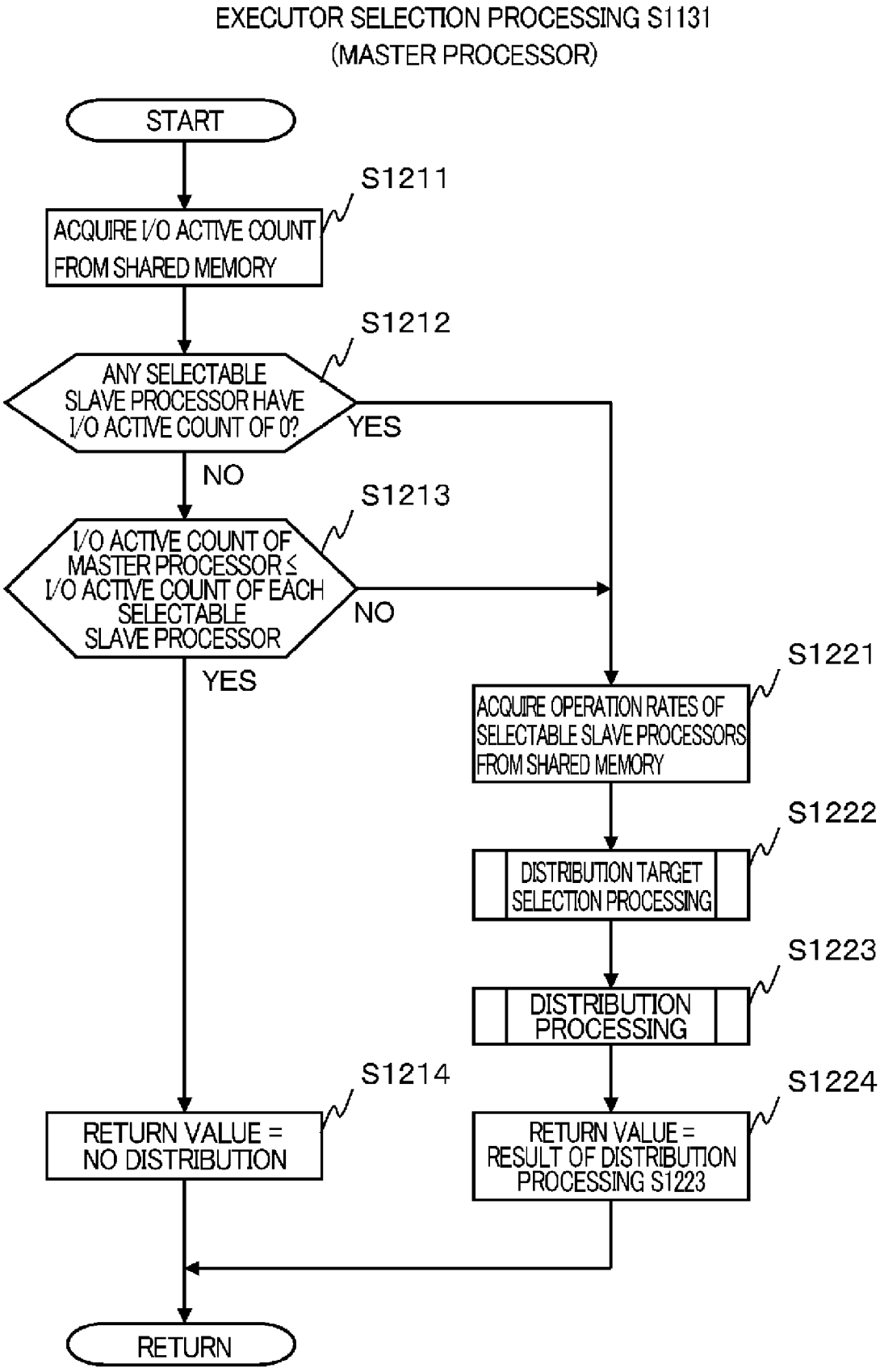
[Fig. 10]


WRITE SEQUENCE S1000
HOST APPARATUS
COMMUNICATION CONTROL PROCESSOR
MASTER PROCESSOR
SLAVE PROCESSORS
A
B
C
RECEIVE WRITE COMMAND
S1011
STORE WRITE COMMAND AND WRITE DATA
S1012
S1013
EXECUTOR DETERMINATION PROCESSING
SEND SUB-COMMAND
S1014
SEND SUB-COMMAND
S1015
SEND WRITE DATA STORAGE NOTIFICATION
S1016
SEND WRITE DATA STORAGE NOTIFICATION
S1017
SEND COMPLETION NOTIFICATION
S1018
SEND COMPLETION NOTIFICATION
S1019
SEND ACKNOWLEDGEMENT OF RECEIPT OF COMPLETION NOTIFICATION
S1020

[Fig. 11]


EXECUTOR DETERMINATION PROCESSING S913 (COMMON TO S1013)
(MASTER PROCESSOR)
START
S1111
PROCESS TARGET DATA INCLUDES TWO OR MORE TRACKS?
YES
NO
S1112
PROCESS TARGET DATA INCLUDES TWO OR MORE RECORDS?
YES
NO
S1113
PROCESS TARGET DATA INCLUDES TWO OR MORE DATA FIELDS?
YES
NO
S1125
PROCESS TARGET DATA INCLUDES D FIELD?
NO
YES
S1126
TOTAL NUMBER OF DATA FIELDS = 2 OR 3
S1121
ACQUIRE TOTAL NUMBER OF DATA FIELDS OF PROCESS TARGET DATA FROM CCW CHAIN AND COMMAND CODE
S1131
EXECUTOR SELECTION PROCESSING
S1141
RETURN VALUE = NO DISTRIBUTION
S1132
RETURN VALUE = RESULT OF EXECUTOR SELECTION PROCESSING
RETURN

[Fig. 12]


EXECUTOR SELECTION PROCESSING S1131
(MASTER PROCESSOR)
START
S1211
ACQUIRE I/O ACTIVE COUNT FROM SHARED MEMORY
S1212
ANY SELECTABLE SLAVE PROCESSOR HAVE I/O ACTIVE COUNT OF 0?
YES
NO
S1213
I/O ACTIVE COUNT OF MASTER PROCESSOR ≤ I/O ACTIVE COUNT OF EACH SELECTABLE SLAVE PROCESSOR
NO
YES
S1221
ACQUIRE OPERATION RATES OF SELECTABLE SLAVE PROCESSORS FROM SHARED MEMORY
S1222
DISTRIBUTION TARGET SELECTION PROCESSING
S1223
DISTRIBUTION PROCESSING
S1214
RETURN VALUE = NO DISTRIBUTION
S1224
RETURN VALUE = RESULT OF DISTRIBUTION PROCESSING S1223
RETURN

[Fig. 13]

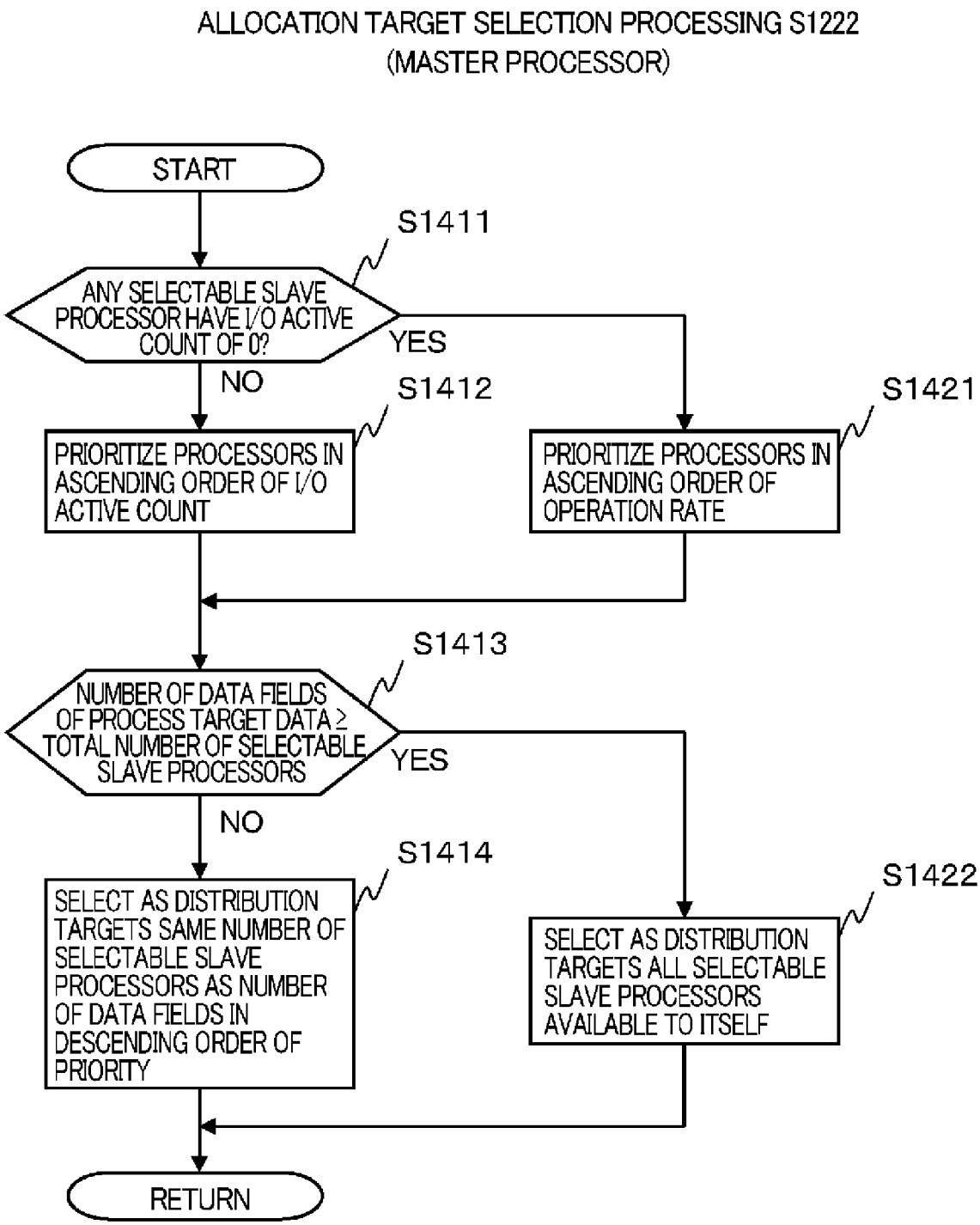
OPERATION INFORMATION 1300

MAIN PROCESSOR ID	ITEM	VALUE
0	OPERATION RATE (%)	5
	I/O ACTIVE COUNT	2
1	OPERATION RATE (%)	1
	I/O ACTIVE COUNT	0
2	OPERATION RATE (%)	8
	I/O ACTIVE COUNT	3
3	OPERATION RATE (%)	8
	I/O ACTIVE COUNT	1
.	.	
.	.	
n		

1311 1311 1311 1311

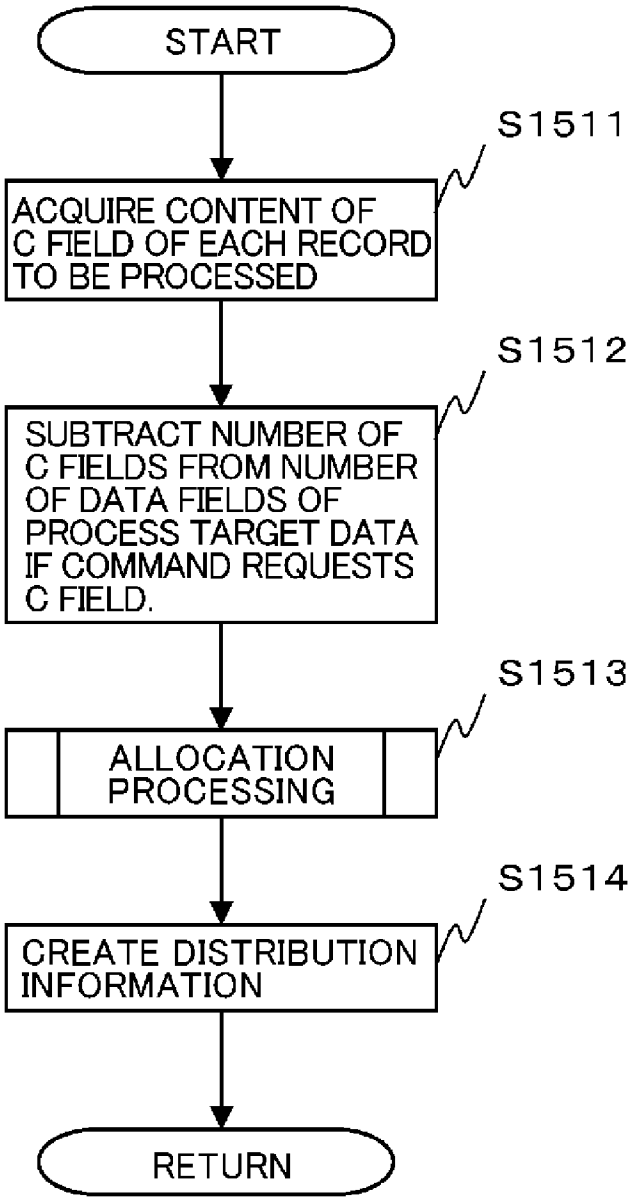
1312 1313 1312 1313 1312 1313 1312 1313

[Fig. 14]

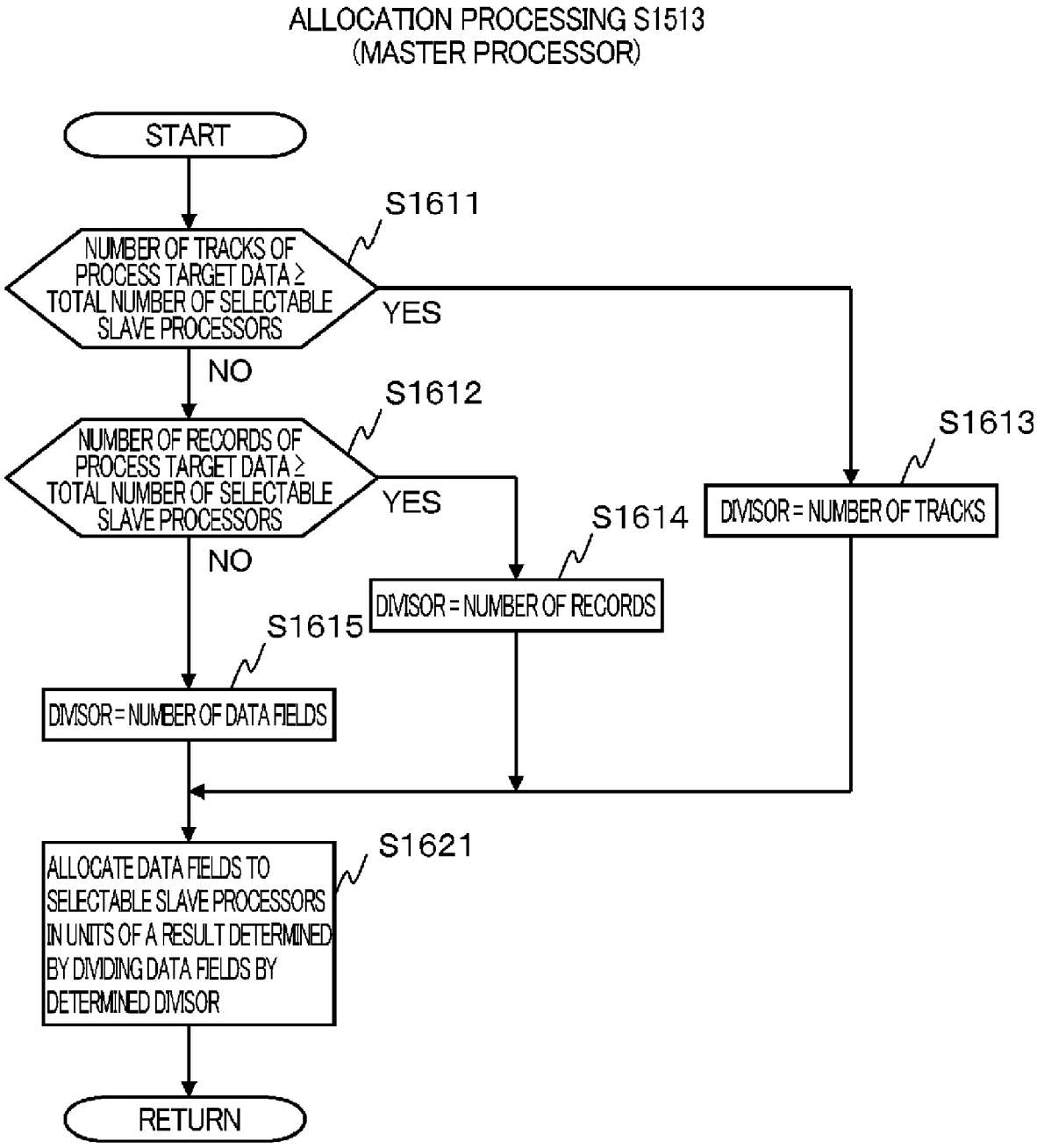

ALLOCATION TARGET SELECTION PROCESSING S1222
(MASTER PROCESSOR)
START
S1411
ANY SELECTABLE SLAVE PROCESSOR HAVE I/O ACTIVE COUNT OF 0?
YES
NO
S1412
PRIORITIZE PROCESSORS IN ASCENDING ORDER OF I/O ACTIVE COUNT
S1421
PRIORITIZE PROCESSORS IN ASCENDING ORDER OF OPERATION RATE
S1413
NUMBER OF DATA FIELDS OF PROCESS TARGET DATA ≥ TOTAL NUMBER OF SELECTABLE SLAVE PROCESSORS
YES
NO
S1414
SELECT AS DISTRIBUTION TARGETS SAME NUMBER OF SELECTABLE SLAVE PROCESSORS AS NUMBER OF DATA FIELDS IN DESCENDING ORDER OF PRIORITY
S1422
SELECT AS DISTRIBUTION TARGETS ALL SELECTABLE SLAVE PROCESSORS AVAILABLE TO ITSELF
RETURN

[Fig. 15]

DISTRIBUTION PROCESSING S1223
(MASTER PROCESSOR)

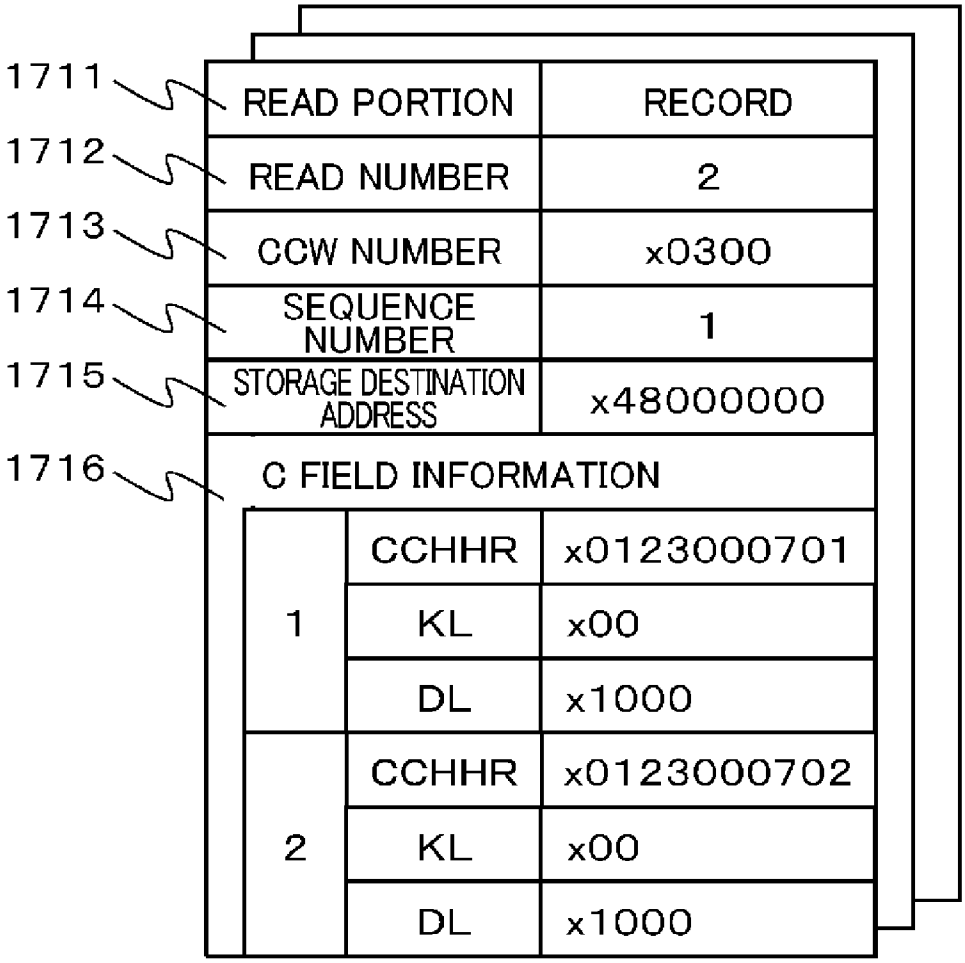


[Fig. 16]

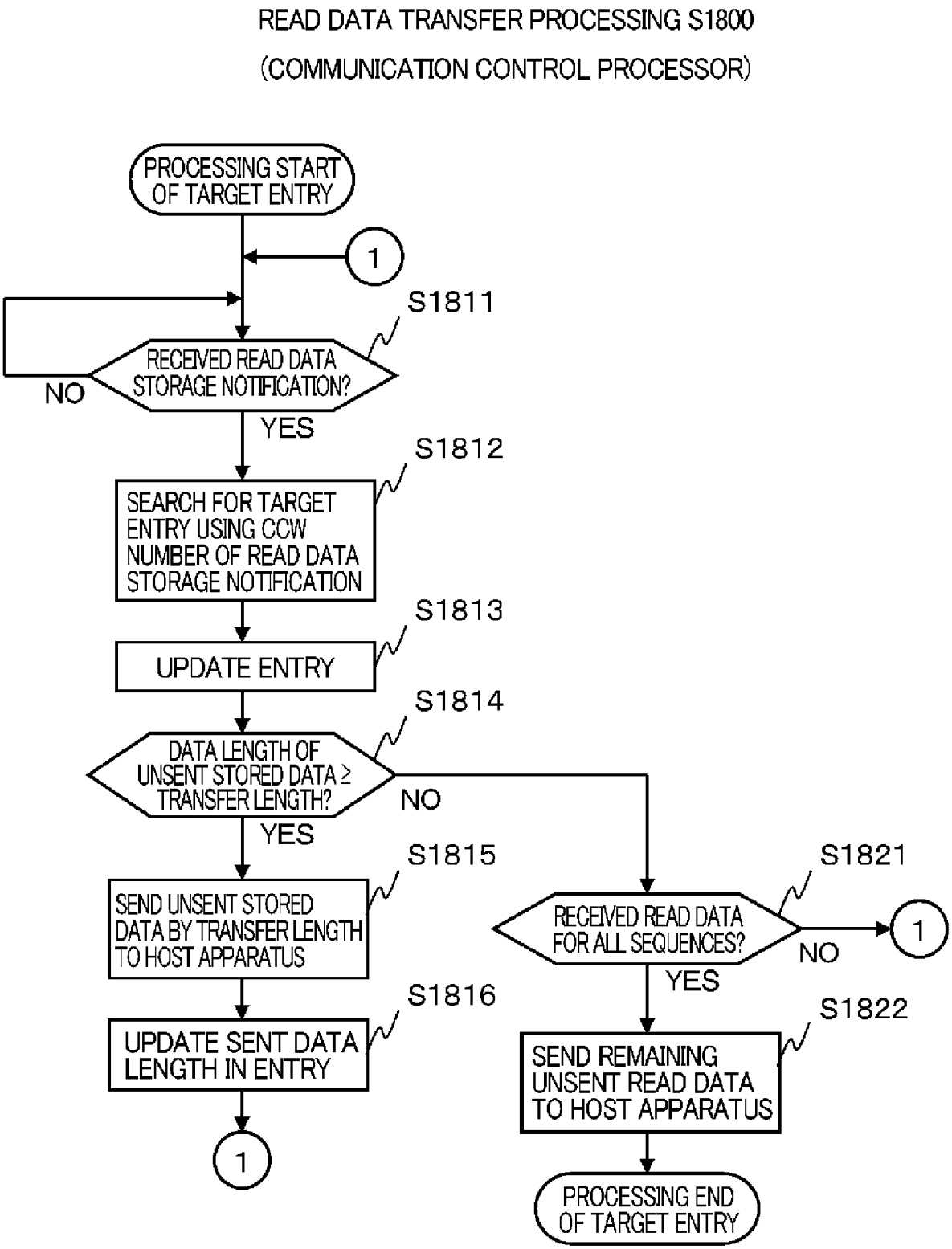

ALLOCATION PROCESSING S1513
(MASTER PROCESSOR)
START
S1611
NUMBER OF TRACKS OF PROCESS TARGET DATA ≥ TOTAL NUMBER OF SELECTABLE SLAVE PROCESSORS
YES
NO
S1612
NUMBER OF RECORDS OF PROCESS TARGET DATA ≥ TOTAL NUMBER OF SELECTABLE SLAVE PROCESSORS
YES
NO
S1613
DIVISOR = NUMBER OF TRACKS
S1614
DIVISOR = NUMBER OF RECORDS
S1615
DIVISOR = NUMBER OF DATA FIELDS
S1621
ALLOCATE DATA FIELDS TO SELECTABLE SLAVE PROCESSORS IN UNITS OF A RESULT DETERMINED BY DIVIDING DATA FIELDS BY DETERMINED DIVISOR
RETURN

[Fig. 17]

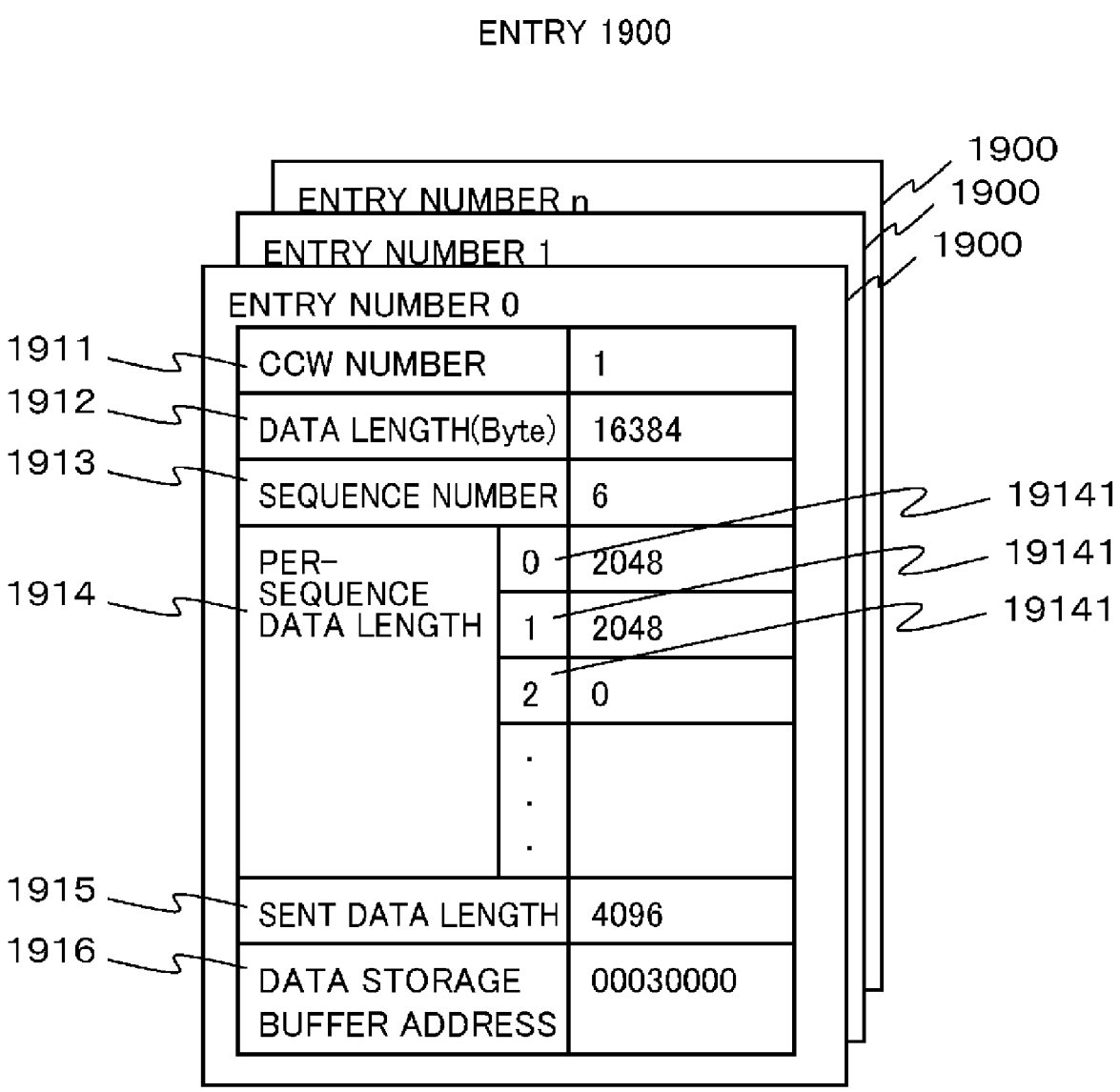
DISTRIBUTION INFORMATION 1700



[Fig. 18]


READ DATA TRANSFER PROCESSING S1800
(COMMUNICATION CONTROL PROCESSOR)
PROCESSING START OF TARGET ENTRY
1
S1811
RECEIVED READ DATA STORAGE NOTIFICATION?
NO
YES
S1812
SEARCH FOR TARGET ENTRY USING CCW NUMBER OF READ DATA STORAGE NOTIFICATION
S1813
UPDATE ENTRY
S1814
DATA LENGTH OF UNSENT STORED DATA ≥ TRANSFER LENGTH?
NO
YES
S1815
SEND UNSENT STORED DATA BY TRANSFER LENGTH TO HOST APPARATUS
S1816
UPDATE SENT DATA LENGTH IN ENTRY
1
S1821
RECEIVED READ DATA FOR ALL SEQUENCES?
NO
1
YES
S1822
SEND REMAINING UNSENT READ DATA TO HOST APPARATUS
PROCESSING END OF TARGET ENTRY

[Fig. 19]


ENTRY 1900
1900
1900
1900
ENTRY NUMBER n
ENTRY NUMBER 1
ENTRY NUMBER 0
1911
1912
1913
1914
1915
1916
CCW NUMBER
1
DATA LENGTH(Byte)
16384
SEQUENCE NUMBER
6
PER-SEQUENCE DATA LENGTH
0
2048
1
2048
2
0
SENT DATA LENGTH
4096
DATA STORAGE BUFFER ADDRESS
00030000
19141
19141
19141

INTERNATIONAL SEARCH REPORT

International application No
PCT/JP2012/006097

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F3/06
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)
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 practicable, 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.
X	US 2009/259924 A1 (AMANN STEFAN [DE] ET AL) 15 October 2009 (2009-10-15) abstract; figures 3,4,5,8 paragraphs [0008], [0012], [0013], [0053], [0054], [0065], [0066], [0069], [0070], [0073], [0074] paragraphs [0079], [0081], [0085], [0086]	1-15
A	EP 1 835 389 A2 (HITACHI LTD [JP]) 19 September 2007 (2007-09-19) the whole document	1-15

☐ 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 application or patent 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	Date of mailing of the international search report
7 March 2013	19/03/2013

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Manfrin, Max

1

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/JP2012/006097

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2009259924 A1	15-10-2009	NONE	
EP 1835389 A2	19-09-2007	EP 1835389 A2	19-09-2007
		JP 5089896 B2	05-12-2012
		JP 2007249729 A	27-09-2007
		US 2007220514 A1	20-09-2007
		US 2012311204 A1	06-12-2012