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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
01 February 2024 (01.02.2024)

WIPO | PCT

(10) International Publication Number
WO 2024/022424 A1

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

(21) International Application Number:
PCT/CN2023/109498

(22) International Filing Date:
27 July 2023 (27.07.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
202210891870.7 27 July 2022 (27.07.2022) CN

(71) Applicant: **ZHEJIANG DAHUA TECHNOLOGY CO., LTD.** [CN/CN]; No. 1187 Binan Road, Binjiang District, Hangzhou, Zhejiang 310053 (CN).

(72) Inventors: **XU, Keyao**; No. 1187 Binan Road, Binjiang District, Hangzhou, Zhejiang 310053 (CN). **LUO, Xin**; No. 1187 Binan Road, Binjiang District, Hangzhou, Zhejiang 310053 (CN). **JIANG, Wenlong**; No. 1187 Binan Road, Binjiang District, Hangzhou, Zhejiang 310053 (CN). **ZHOU, Mingwei**; No. 1187 Binan Road, Binjiang District, Hangzhou, Zhejiang 310053 (CN).

(74) Agent: **METIS IP (CHENGDU) LLC**; Room 401, 4th Floor, Unit 1 Of Building 1, No. 269 North Section Of Hupan Road, Tianfu New District, China Free Trade Zone (Sichuan), Chengdu, Sichuan 610213 (CN).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO,

(54) Title: SYSTEM AND METHODS FOR METADATA SERVICES

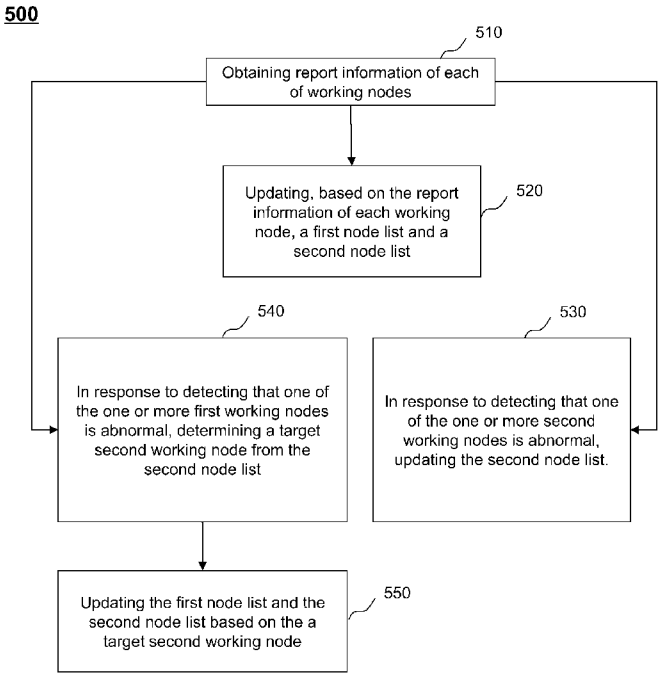


FIG. 5

(57) Abstract: A cluster system may be provided. The cluster system may include a management processor, a main node configured for providing metadata services, and working nodes being communicatively connected to the main node and configured to send report information to the main node. The management processor may be configured to update a first node list relating to the one or more first working nodes configured for metadata backup and a second node list relating to one or more second working nodes other than the one or more first working nodes based on the report information of each working node. In response to detecting that one of the one or more first working nodes is abnormal, the management processor may determine a target second working node from the second node list and update the first node list and the second node list based on the target second working node.

WO 2024/022424 A1

[Continued on next page]

RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) **Designated States** *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, SD, SL, ST, 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, ME, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

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

SYSTEM AND METHODS FOR METADATA SERVICES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims priority to Chinese Patent Application No. 202210891870.7 filed on July 27, 2022, the entire contents of which are hereby incorporated by reference.

TECHNICAL FIELD

[0002] The present disclosure generally relates to metadata services, and more particularly, relates to systems and methods for providing metadata services using a cluster system.

BACKGROUND

[0003] With the development of the big data, private and public cloud storage technologies have been widely used, and the requirements for data security, data reliability, and service availability are getting higher and higher. Metadata services were an object-oriented repository technology that could be integrated with enterprise information systems or with applications that process metadata. A cluster system, which includes multiple metadata nodes, is often used to provide metadata services. The availability of metadata services of a cluster system refers to the ability of the cluster system to provide metadata services when one or more metadata nodes of the cluster system fail (i.e., are abnormal). The availability of the metadata services is vital for the cluster system. Therefore, it is desirable to provide systems and methods to improve the availability of the metadata services of a cluster system.

SUMMARY

[0004] According to yet another aspect of the present disclosure, a cluster system may be provided. The cluster system may include a main node, working nodes, and a management processor. The main node may be configured for providing metadata services. Each working node may be communicatively connected to the main node and configured to send report information to the main node. The working nodes may include one or more first working nodes. The one or more first working nodes may be standby nodes of the main node configured for metadata backup. The management processor may be configured to update a first node list and a second node list based on the report information of each working node. The first node list may relate to the one or more first working nodes, and the second node list may relate to one or more second working nodes other than the one or more first working nodes among the working nodes. In response to detecting that one of the one or more first working nodes is abnormal, the management processor may determine a target second working node from the second node list, designate the target second working node as a new first working node, and update the first node list and the second node list.

[0005] In some embodiments, the management processor may be part of the main node, or the management processor may be independent from the main node and configured to receive the report information of each of the working nodes from the main node.

[0006] In some embodiments, in response to detecting that one of the one or more second working nodes is abnormal, the management processor may be further configured t remove the abnormal second working node from the second node list.

[0007] In some embodiments, the management processor may update the first node list and the second node list based on the report information of each working node by performing the following operations. For each

working node, the management processor may determine, whether the working node is a first working node or a second working node based on the report information of the working node. In response to determining that the working node is a first working node, the management processor may update the first node list based on the report information of the working node. In response to determining that the working node is a second working node, the management processor may update the second node list based on the report information of the working node.

[0008] In some embodiments, the target second working node may be determined from the second node list by performing the following operations. For each second working node in the second node list, the management processor may determine a load of the second working node. Further, the management processor may determine the target second working node based on the load of each second working node.

[0009] In some embodiments, the target second working node may be determined from the second node list by performing the following operations. For each second working node in the second node list, the management processor may determine a probability that the second working node is abnormal based on the report information of the second working node. Further, the management processor may determine the target second working node based on the probability corresponding to each second working node.

[0010] In some embodiments, the target second working node may be determined from the second node list by performing the following operations. The management processor may obtain feature information of each second working node in the second node list. The management processor may further determine the target second working node based on the feature information of each second working node using a target node determination model. The target node determination model may be a trained machine learning model.

[0011] In some embodiments, in response to detecting that one of the one or more first working nodes is abnormal, to determine the target second working node from the second node list, the management processor may be further configured to perform the following operations. In response to determining that detecting that one of the one or more first working nodes is abnormal, the management processor may determine whether the count of remaining first working nodes in the first node list other than the first working node is smaller than a count threshold. In response to determining that the count of remaining first working nodes in the first node list is smaller than a count threshold, the management processor may determine the target second working node from the second node list.

[0012] In some embodiments, in response to detecting that the main node is abnormal, the management processor may be further configured to, determine a target first working node that performs the metadata services of the main node in place of the main node, and update the first node list and the second node list based on the target first working node.

[0013] In some embodiments, the management processor may be communicated with a second management processor. When there is no target second working node in the second node list, the second management processor may be configured to determine a working node from a second cluster system, and designate the working node as the target second working node.

[0014] According to another aspect of the present disclosure, a method implemented on a management processor of a cluster system may be provided. The cluster system may further comprise a main node

configured for providing metadata services and working nodes each of which is communicatively connected to the main node and configured to send report information to the main node. The working nodes may include one or more first working nodes. The one or more first working nodes may be standby nodes of the main node configured for metadata backup. The method comprising updating a first node list and a second node list based on the report information of each working node. The first node list may relate to the one or more first working nodes, and the second node list may relate to one or more second working nodes other than the one or more first working nodes among the working nodes. In response to detecting that one of the one or more first working nodes is abnormal, the method further comprising determining a target second working node from the second node list, designating the target second working node as a new first working node, and updating the first node list and the second node list.

[0015] According to yet another aspect of the present disclosure, a non-transitory computer readable medium may be provided. The non-transitory computer readable medium may comprise a set of instructions. The set of instructions may be executed by a management processor of a cluster system. The cluster system may further comprise a main node configured for providing metadata services and working nodes each of which is communicatively connected to the main node and configured to send report information to the main node. The working nodes may include one or more first working nodes. The one or more first working nodes may be standby nodes of the main node configured for metadata backup. When the set of instructions are executed by the management processor, the set of instructions may cause the management processor to perform a method. The method comprising updating a first node list and a second node list based on the report information of each working node. The first node list may relate to the one or more first working nodes, and the second node list may relate to one or more second working nodes other than the one or more first working nodes among the working nodes. In response to detecting that one of the one or more first working nodes is abnormal, the method further comprising determining a target second working node from the second node list, designating the target second working node as a new first working node, and updating the first node list and the second node list.

[0016] Additional features may be set forth in part in the description which follows, and in part may become apparent to those skilled in the art upon examination of the following and the accompanying drawings or may be learned by production or operation of the examples. The features of the present disclosure may be realized and attained by practice or use of various aspects of the methodologies, instrumentalities, and combinations set forth in the detailed examples discussed below.

BRIEF DESCRIPTION OF THE DRAWINGS

[0017] The present disclosure is further described in terms of exemplary embodiments. These exemplary embodiments are described in detail with reference to the drawings. The drawings are not to scale. These embodiments are non-limiting exemplary embodiments, in which like reference numerals represent similar structures throughout the several views of the drawings, and wherein:

[0018] FIG. 1 is a schematic diagram illustrating an exemplary cluster system according to some embodiments of the present disclosure;

[0019] FIG. 2 is a schematic diagram illustrating an exemplary cluster system 200 according to some embodiments of the present disclosure;

[0020] FIG. 3 is a schematic diagram illustrating exemplary hardware and/or software components of a computing device 300 according to some embodiments of the present disclosure;

[0021] FIG. 4 is a block diagram illustrating an exemplary management processor according to some embodiments of the present disclosure;

[0022] FIG. 5 is a flowchart illustrating an exemplary metadata services process according to some embodiments of the present disclosure;

[0023] FIG. 6 is a schematic diagram illustrating an exemplary updating of a second node list 620 according to some embodiments of the present disclosure;

[0024] FIG. 7 is a schematic diagram illustrating an exemplary updating of a first node list 710 and a second node list 720 according to some embodiments of the present disclosure;

[0025] FIG. 8 is a flowchart illustrating an exemplary process 800 for managing nodes in a cluster system according to some embodiments of the present disclosure; and

[0026] FIG. 9 is a schematic diagram illustrating an exemplary updating of a first node list 910 and a second node list 920 according to some embodiments of the present disclosure.

DETAILED DESCRIPTION

[0027] In the following detailed description, numerous specific details are set forth by way of examples in order to provide a thorough understanding of the relevant disclosure. However, it should be apparent to those skilled in the art that the present disclosure may be practiced without such details. In other instances, well-known methods, procedures, systems, components, and/or circuitry have been described at a relatively high-level, without detail, in order to avoid unnecessarily obscuring aspects of the present disclosure. Various modifications to the disclosed embodiments will be readily apparent to those skilled in the art, and the general principles defined herein may be applied to other embodiments and applications without departing from the spirit and scope of the present disclosure. Thus, the present disclosure is not limited to the embodiments shown, but to be accorded the widest scope consistent with the claims.

[0028] It will be understood that the term "system," "engine," "unit," "module," and/or "block" used herein are one method to distinguish different components, elements, parts, section or assembly of different level in ascending order. However, the terms may be displaced by other expressions if they may achieve the same purpose.

[0029] Generally, the word "module," "unit," or "block," as used herein, refers to logic embodied in hardware or firmware, or to a collection of software instructions. A module, a unit, or a block described herein may be implemented as software and/or hardware and may be stored in any type of non-transitory computer-readable medium or other storage devices. In some embodiments, a software module/unit/block may be compiled and linked into an executable program. It will be appreciated that software modules can be callable from other modules/units/blocks or from themselves, and/or may be invoked in response to detected events or interrupts. Software modules/units/blocks configured for execution on computing devices (e.g., processor 320 as illustrated in FIG. 3) may be provided on a computer readable medium, such as a compact disc, a digital

video disc, a flash drive, a magnetic disc, or any other tangible medium, or as a digital download (and can be originally stored in a compressed or installable format that needs installation, decompression, or decryption prior to execution). Such software code may be stored, partially or fully, on a storage device of the executing computing device, for execution by the computing device. Software instructions may be embedded in firmware, such as an EPROM. It will be further appreciated that hardware modules (or units or blocks) may be included in connected logic components, such as gates and flip-flops, and/or can be included in programmable units, such as programmable gate arrays or processors. The modules (or units or blocks) or computing device functionality described herein may be implemented as software modules (or units or blocks) but may be represented in hardware or firmware. In general, the modules (or units or blocks) described herein refer to logical modules (or units or blocks) that may be combined with other modules (or units or blocks) or divided into sub-modules (or sub-units or sub-blocks) despite their physical organization or storage.

[0030] It will be understood that when a unit, engine, module, or block is referred to as being "on," "connected to," or "coupled to" another unit, engine, module, or block, it may be directly on, connected or coupled to, or communicate with the other unit, engine, module, or block, or an intervening unit, engine, module, or block may be present, unless the context clearly indicates otherwise. As used herein, the term "and/or" includes any and all combinations of one or more of the associated listed items.

[0031] The terminology used herein is for the purposes of describing particular examples and embodiments only and is not intended to be limiting. As used herein, the singular forms "a," "an," and "the" may be intended to include the plural forms as well, unless the context clearly indicates otherwise. It will be further understood that the terms "include" and/or "comprise," when used in this disclosure, specify the presence of integers, devices, behaviors, stated features, steps, elements, operations, and/or components, but do not exclude the presence or addition of one or more other integers, devices, behaviors, features, steps, elements, operations, components, and/or groups thereof.

[0032] In addition, it should be understood that in the description of the present disclosure, the terms "first", "second", or the like, are only used to distinguish the purpose of description, and cannot be interpreted as indicating or implying relative importance, nor can be understood as indicating or implying the order.

[0033] Conventionally, in order to prevent the abnormality of one or more metadata nodes of a cluster system from affecting the operation of the cluster system, one or more standby nodes (also referred to as secondary nodes) are usually set up for a main node (also referred to as a primary node) that provides metadata services. When the main node is abnormal, the one or more standby nodes can replace the main node to provide the metadata services for the cluster system. In this case, the availability of metadata services directly depends on a number of standby nodes that can be used to replace the main node to provide metadata services. The maximum availability of the metadata services depends on the number of the standby nodes. For example, if there are n metadata standby nodes in the cluster system, in order to ensure the normal operation of metadata services, the cluster system can allow up to n metadata nodes to be abnormal, and the maximum availability of the metadata services of the cluster system is limited.

[0034] The present disclosure provides systems and methods for metadata services using a cluster system. The systems may include a main node, working nodes, and a management processor. The main node may be

configured for providing metadata services. Each working node may be communicatively connected to the main node and configured to send report information to the main node. The working nodes may include one or more first working nodes. The one or more first working nodes may be standby nodes of the main node configured for metadata backup. The management processor may be configured to update a first node list and a second node list based on the report information of each working node. The first node list may relate to the one or more first working nodes, and the second node list may relate to one or more second working nodes other than the one or more first working nodes among the working nodes. In response to detecting that one of the one or more first working nodes is abnormal, the management processor may determine a target second working node from the second node list, designate the target second working node as a new first working node, and update the first node list and the second node list.

[0035] According to the present systems and methods, the availability of metadata services depends not only on the number of standby nodes (i.e., the first working nodes) but also on the number of second working nodes, the maximum availability of the metadata services of the cluster system may be determined based on a sum of the number of the first working nodes and the number of second working nodes. Compared with the conventional approach for metadata services, the present systems and methods may greatly improve the availability of metadata services.

[0036] FIG. 1 is a schematic diagram illustrating an exemplary cluster system 100 according to some embodiments of the present disclosure. As shown in FIG.1, the cluster system 100 may include a main node 110, working nodes 120, and a management processor 130. In some embodiments, two or more of the main node 110, the working nodes 120, and the management processor 130 may be connected each other via a network or directly.

[0037] The main node 110 may be configured for providing metadata services. The metadata services may include various management services for metadata such as a metadata storage service, a metadata updating service, a metadata collection service, etc. Each working node may be communicatively connected to the main node 110 and configured to send report information to the main node 110. The main node 110 may record, update, and detect the report information reported by each working node in real time. In some embodiments, each working node may be capable of providing metadata services.

[0038] In some embodiments, the working nodes 120 may include one or more first working nodes 1201 and one or one or more second working nodes 1202 other than the one or more first working nodes 1201. The one or more first working nodes 1201 may be standby nodes of the main node 110 configured for metadata backup. A first working node refers to a node that is providing metadata services and stores metadata of the cluster system 100. The task(s), configuration information, or other information relating to metadata services of the main node 110 and the one or more first working nodes 1201 may be synchronized. For example, after the main node 110 collects metadata, the metadata may be synchronized to the one or more first working nodes 1201 for storage and backup. The one or more first working nodes 1201 may replace the main node 110 to perform the metadata services of the cluster system 100 when, for example, the main node 110 is abnormal.

[0039] As used herein, that a node is abnormal refers to that the node cannot operate in accordance with a normal mode. For example, the node may crash when a computer program, such as a software application or

an operating system of the node stops functioning properly. As another example, the node can not work normally when a hard disk drive has malfunctions. For example, when the main node 110 is abnormal, one of the one or more first working nodes 1201 may take over one or more tasks of the main node 110 and perform the task(s). In some embodiments, when the main node 110 is normal, the one or more first working nodes 1201 may be also configured to work as a working server and perform one or more other tasks (e.g., data computational tasks) of the cluster system 100.

[0040] As used herein, a second working node refers to a node that has not started metadata services and stores data other than the metadata of the cluster system 100. The one or one or more second working nodes 1202 may process information and/or data relating to the cluster system 100 other than metadata to perform one or more tasks of the cluster system 100 other than the metadata services. For example, the one or one or more second working nodes 1202 may be used to perform computational tasks to analyze data. As another example, the one or one or more second working nodes 1202 may be used to process an instruction or data received from a user terminal.

[0041] The management processor 130 may be independent from the main node 110 and configured to receive the report information of each of the working nodes from the main node 110. The management processor 130 may be configured to monitor and/or manage the nodes (e.g., the main node 110, the one or more first working nodes 1201, the one or one or more second working nodes 1202, etc.) of the cluster system 100. In some embodiments, the management processor 130 may be configured to monitor each node of the cluster system 100 and keep the cluster system 100 operating normally. For example, the management processor 130 may monitor the working nodes of the cluster system 100 via process 500 to ensure that the cluster system 100 includes enough first working nodes (i.e., standby nodes of the main node configured for metadata backup). As another example, the management processor 130 may monitor the main node 110 of the cluster system 100 via process 800 to ensure that the cluster system 100 includes a normal main node for providing metadata services.

[0042] In some embodiments, the management processor 130 (e.g., one or more modules illustrated in FIG. 4) may perform the methods of the present disclosure. For example, the process 500 and/or the process 800 may be implemented as a set of instructions (e.g., an application) stored in a storage device. The management processor 130 may execute the set of instructions and may accordingly be directed to perform the process 500 and/or the process 800.

[0043] In some embodiments, a node of the cluster system 100 (such as, the main node 110, the working nodes 120, and the management processor 130) may include one or more processing units (e.g., single-core processing device(s) or multi-core processing device(s)). Merely by way of example, the management processor 130 may include a central processing unit (CPU), an application-specific integrated circuit (ASIC), an application-specific instruction-set processor (ASIP), a graphics processing unit (GPU), a physics processing unit (PPU), a digital signal processor (DSP), a field programmable gate array (FPGA), a programmable logic device (PLD), a controller, a microcontroller unit, a reduced instruction-set computer (RISC), a microprocessor, or the like, or any combination thereof.

[0044] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. In some embodiments, the cluster system 100 may include one or more additional components. For example, the cluster system 100 may include a network that can facilitate the exchange of information and/or data in the cluster system 100. In some embodiments, one or more components in the cluster system 100 (e.g., the main node 110, the working nodes 120, and the management processor 130) may send information and/or data to another component(s) in the cluster system 100 via the network. The network may be any type of wired or wireless network, or a combination thereof. As another example, the cluster system 100 may further include a user terminal that enables user interactions between a user and one or more components of the cluster system 100.

[0045] FIG. 2 is a schematic diagram illustrating an exemplary cluster system 200 according to some embodiments of the present disclosure. As shown in FIG.2, the cluster system 200 may be similar to the cluster system 100 as described in FIG. 1, except that the management processor 130 is part of the main node 110. In such cases, the main node 110 may be configured to monitor each node of the cluster system 100 and keep the cluster system 100 operating normally (e.g., by performing the process 500).

[0046] FIG. 3 is a schematic diagram illustrating exemplary hardware and/or software components of a computing device 300 according to some embodiments of the present disclosure. The computing device 300 may be used to implement any component (e.g., the main node 110, a working node 120, and the management processor 130) of the cluster system 100 as described herein. For example, the management processor 130 may be implemented on the computing device 300, via its hardware, software program, firmware, or a combination thereof. Although only one such computer is shown, for convenience, the computer functions relating to system recovery as described herein may be implemented in a distributed fashion on a number of similar platforms to distribute the processing load.

[0047] The computing device 300, for example, may include COM ports 350 connected to and from a network connected thereto to facilitate data communications. The computing device 300 may also include a processor (e.g., the processor 320), in the form of one or more processors (e.g., logic circuits), for executing program instructions. For example, the processor 320 may include interface circuits and processing circuits therein. The interface circuits may be configured to receive electronic signals from a bus 310, wherein the electronic signals encode structured data and/or instructions for the processing circuits to process. The processing circuits may conduct logic calculations, and then determine a conclusion, a result, and/or an instruction encoded as electronic signals. Then the interface circuits may send out the electronic signals from the processing circuits via the bus 310.

[0048] The computing device 300 may further include program storage and data storage of different forms including, for example, a disk 370, and a read-only memory (ROM) 330, or a random-access memory (RAM) 340, for various data files to be processed and/or transmitted by the computing device 300. The computing device 300 may also include program instructions stored in the ROM 330, RAM 340, and/or another type of non-transitory storage medium to be executed by the processor 320. The methods and/or processes of the

present disclosure may be implemented as the program instructions. The computing device 300 may also include an I/O component 360, supporting input/output between the computer and other components. The computing device 300 may also receive programming and data via network communications.

[0049] Merely for illustration, only one processor is illustrated in FIG. 3. Multiple processors 320 are also contemplated; thus, operations and/or method steps performed by one processor 320 as described in the present disclosure may also be jointly or separately performed by the multiple processors. For example, if in the present disclosure the processor 320 of the computing device 300 executes both step A and step B, it should be understood that step A and step B may also be performed by two different processors 320 jointly or separately in the computing device 300 (e.g., a first processor executes step A and a second processor executes step B or the first and second processors jointly execute steps A and B).

[0050] FIG. 4 is a block diagram illustrating an exemplary management processor according to some embodiments of the present disclosure. As shown in FIG. 4, the management processor 130 may include an acquisition module 410, an updating module 420, and a determination module 430.

[0051] The acquisition module 410 may be configured to obtain information relating to the cluster system 100. For example, the acquisition module 410 may obtain report information of each of working nodes. More descriptions regarding the obtaining of the report information of each of working nodes may be found elsewhere in the present disclosure. See, e.g., operation 510 in FIG. 5, and relevant descriptions thereof. As another example, the acquisition module 410 may obtain information relating to the main node. More descriptions regarding the obtaining of the information relating to the main node may be found elsewhere in the present disclosure. See, e.g., operation 810 in FIG. 8, and relevant descriptions thereof.

[0052] The updating module 420 may be configured to update a first node list and a second node list based on the report information of each working node. In some embodiments, in response to determining that the working node is a first working node, the updating module 42 may update the first node list based on the report information of the first working node. In some embodiments, in response to determining that the working node is a second working node, the updating module 420 may update the second node list based on the report information of the second working node. In some embodiments, in response to detecting that one of the one or more second working nodes is abnormal, the updating module 420 may update the second node list. In some embodiments, in response to detecting that one of the one or more first working nodes is abnormal, the updating module 420 may update the first node list and the second node list. More descriptions regarding the updating of the first node list and the second node list may be found elsewhere in the present disclosure. See, e.g., operations 520, 530, and 550 in FIG. 5, and relevant descriptions thereof.

[0053] The determination module 430 may be configured to determine a target second working node from the second node list. For example, in response to detecting that one of the one or more first working nodes is abnormal, the m determination module 430 may determine a target second working node from the second node list. More descriptions regarding the determination of the target second working node may be found elsewhere in the present disclosure. See, e.g., operation 540 in FIG. 5, and relevant descriptions thereof.

[0054] In some embodiments, the determination module 430 may be also configured to determine whether the main node is abnormal based on the information relating to the main node. In response to determining that

the main node is abnormal, the determination module 430 may be configured to determine a target first working node that performs the metadata services of the main node in place of the main node. More descriptions regarding the determination of the target first working node may be found elsewhere in the present disclosure. See, e.g., operation 830 in FIG. 8, and relevant descriptions thereof.

[0055] In some embodiments, the updating module 420 may be also configured to update the first node list and the second node list based on the target first working node. More descriptions regarding the updating of the first node list and the second node list based on the target first working node may be found elsewhere in the present disclosure. See, e.g., operation 840 in FIG. 8, and relevant descriptions thereof.

[0056] It should be noted that the above description is merely provided for the purposes of illustration, and not intended to limit the scope of the present disclosure. For persons having ordinary skills in the art, multiple variations and modifications may be made under the teachings of the present disclosure. However, those variations and modifications do not depart from the scope of the present disclosure. In some embodiments, the management processor 130 may include one or more additional modules, such as a storage module (not shown) for storing data.

[0057] FIG. 5 is a flowchart illustrating an exemplary metadata services process according to some embodiments of the present disclosure.

[0058] In 510, the management processor 130 (e.g., the acquisition module 410) may obtain report information of each of working nodes.

[0059] As described in FIG. 1, each working node may be communicatively connected to the main node 110 and configured to send report information to the main node 110 periodically or aperiodically. When the management processor 130 is independent from the main node 110, the management processor 130 may obtain the report information of each of working nodes from the main node 110 periodically or aperiodically. When the management processor 130 is part of the main node 110, the management processor 130 may directly obtain the report information from each working node.

[0060] As described in FIG. 1, the working nodes may include one or more first working nodes 1201 and one or one or more second working nodes 1202 other than the one or more first working nodes 1201. In some embodiments, the report information relating to a first working node may include an internet protocol (IP) address of the first working node, a reporting time, a metadata service state of the first working node, a mark indicating that the first working node is a standby node of the main node, a version number of the first working node, or the like, or any combination thereof. As used herein, a metadata service state of a node may include an abnormal state and a normal state. As used herein, the abnormal state of the metadata service state refers to a state in which the node cannot perform metadata services in accordance with a normal mode; the normal state of the metadata service state refers to a state in which the node can perform metadata services in accordance with the normal mode.

[0061] In some embodiments, the report information relating to each of the one or more second working nodes may include an internet protocol (IP) address of the second working node, a reporting time, or the like, or any combination thereof.

[0062] In 520, the management processor 130 (e.g., the updating module 420) may update, based on the report information of each working node, a first node list and a second node list.

[0063] In some embodiments, the first node list may relate to the one or more first working nodes. The first node list may include report information of one or more first working nodes. In some embodiments, the first node list may be stored as a file, such as a first information configuration file. Merly by way of example, the first node list may include report information of M first working nodes as shown in Table 1 below.

Table 1 Exemplary First Node List

Node Index	IP Address	Reporting Time	Metadata service state	Mark
Node 1	192.168.0.1	8：00	Normal State	Standby Node 1
Node 2	192.168.0.2	8：10	Normal State	Standby Node 2
…	…	…	…	…
Node (M-1)	192.168.0.M-1	9：10	Normal State	Standby Node (M-1)
Node M	192.168.0.M	09：20	Normal State	Standby Node M

[0064] In some embodiments, the second node list may relate to the one or more second working nodes. The second node list may include report information of one or more second working nodes. In some embodiments, the second node list may be stored as a file, such as a second information configuration file. Merely by way of example, the second node list may include report information of N second working nodes as shown in Table 2 below.

Table 2 Exemplary Second Node List

Node Index	IP Address	Reporting Time	Metadata service state
Node 1	192.168.1.1	8：10	Normal State
Node 2	192.168.1.2	8：30	Normal State
…	…	…	…
Node (N-1)	192.168.1.N-1	9：30	Normal State
Node N	192.168.1.N	09：40	Normal State

[0065] According to some embodiments of the present application, the report information of the first working nodes and the report information of the second working nodes may be recorded separately in different lists (i.e., the first node list and the second node list), which may be convenient for users to query and manage the report information of different nodes.

[0066] In some embodiments, for each working node, the management processor 130 may determine whether the working node is a first working node or a second working node based on the report information of the working node. For example, the management processor 130 may determine whether the working node is a first working node or a second working node according to the IP address of the working node and IP addresses of different working nodes (which is pre-stored in the management processor 130). As another example, the

management processor 130 may determine whether the report information of the working node includes a mark indicating that the first working node is a standby node of the main node. In response to determining that the report information of the working node includes the mark indicating that the first working node is a standby node of the main node, the management processor 130 may determine that the working node is a first working node. In response to determining that the report information of the working node does not include the mark indicating that the first working node is a standby node of the main node, the management processor 130 may determine that the working node is a second working node.

[0067] In some embodiments, in response to determining that the working node is a first working node, the management processor 130 may update the first node list based on the report information of the first working node. The management processor 130 may update the report information of the first working node in the first node list. For example, the management processor 130 may determine whether there is a record corresponding to the first working node in the first node list according to the IP address of the first working node. In response to determining that there is a record corresponding to the first working node in the first node list, the management processor 130 may update the record of the first working node in the first node list based on the newly received report information of the first working node. For example, the management processor 130 may replace the reporting time in the first node list with the current time. If the metadata service state of the first working node changes, the management processor 130 may update the metadata service state in the first node list. In response to determining that there is no record corresponding to the first working node in the first node list, the management processor 130 may add a record for recording the report information of the first working node into the first node list.

[0068] In some embodiments, in response to determining that the working node is a second working node, the management processor 130 may update the second node list based on the report information of the second working node. In some embodiments, the second node list may be updated in a similar manner as how the first list node is updated. For example, in response to determining that there is a record corresponding to the second working node in the second node list, the management processor 130 may update the record of the second working node in the second node list based on the newly received report information. For example, the management processor 130 may the reporting time in the second node list with the current time. In response to determining that there is no record corresponding to the second working node in the second node list, the management processor 130 may add a record for recording the report information of the second working node in the second node list.

[0069] In some embodiments, the management processor 130 may monitor the states of each first working node and each second working node. When there are one or more working nodes are abnormal, the management processor 130 may also update the first node list and the second node list by performing operation 530 and/or operation 540.

[0070] In 530, in response to detecting that one of the one or more second working nodes is abnormal, the management processor 130 (e.g., the updating module 420) may update the second node list.

[0071] In some embodiments, for each of the one or more second working nodes, the management processor 130 may determine whether the second working node is abnormal based on the reporting time of the

second working node in the second node list. For example, the management processor 130 may determine whether a difference between the reporting time of the second working node in the second node list and the current time is greater than a first time threshold. In response to determining that the difference between the reporting time of the second working node in the second node list and the current time is greater than the first time threshold (i.e., the second working node has not reported for a long time), the management processor 130 may determine that the second working node is abnormal.

[0072] In response to detecting that one of the one or more second working nodes is abnormal, the management processor 130 may update the second node list. Specifically, the management processor 130 may remove the abnormal second working node from the second node list (e.g., by deleting the report information of the abnormal second working node). For example, FIG. 6 is a schematic diagram illustrating an exemplary updating of a second node list 620 according to some embodiments of the present disclosure. As shown in FIG. 6, in response to detecting that a black second working node 1202 in the second node list 620 is abnormal, the management processor 130 may update the second node list 620 by removing the black second working node 1202 and deleting the report information of the black second working node 1202 from the second node list 620.

[0073] In 540, in response to detecting that one of the one or more first working nodes is abnormal, the management processor 130 (e.g., the determination module 430) may determine a target second working node from the second node list.

[0074] In some embodiments, for each of the one or more first working nodes, the management processor 130 may determine whether the first working node is abnormal based on the reporting time of the first working node in the first node list. For example, the management processor 130 may determine whether a difference between the reporting time of the first working node in the first node list and the current time is greater than a second time threshold. In response to determining that the difference between the reporting time of the first working node in the first node list and the current time is greater than the second time threshold, the management processor 130 may determine that the first working node is abnormal. The first time threshold and the second time threshold may be set manually by a user (e.g., an engineer) according to an experience value or be a default setting of the cluster system 100, such as 5 mins, 10 mins, or a larger or smaller value.

[0075] As another example, the management processor 130 may determine whether the first working node is abnormal according to the metadata service state of the first working node. In response to determining that the metadata service state of the first working node is an abnormal state, the management processor 130 may determine that the first working node is abnormal.

[0076] In some embodiments, as described in FIG. 1, the one or one or more second working nodes 1202 may process information and/or data relating to the cluster system 100 other than metadata to perform one or more other tasks of the cluster system 100 other than the metadata services, that is, the second working nodes in the second node list may be performing other tasks other than the metadata services. When a load of a second working node is greater than a load threshold, the second working node is not suitable to provide metadata services. The load of the second working node may reflect an amount of tasks processed by the second working node. Thus, in some embodiments, for each second working node in the second node list, the

management processor 130 may determine the load of the second working node. For example, the management processor 130 may determine the load of the second working node based on a central processing unit (CPU) usage, a memory usage, an input/output (IO) load, a network bandwidth, etc., used by the second working node. For example, the greater the CPU usage used by the second working node is, the greater the load of the second working node may be. Further, the management processor 130 may determine the target second working node based on the load of each second working node. Specifically, the management processor 130 may determine one or more second working nodes with loads smaller than the load threshold, and select one of the one or more second working nides as the target second working node. For example, the management processor 130 may designate a second working node with the minimum load in the one or more second working nides as the target second working node. In this way, the target second working node may have enough load to perform metadata related tasks, and load balancing can be achieved in the cluster system.

[0077] In some embodiments, for each second working node in the second node list, the management processor 130 may determine a probability that the second working node is abnormal based on the report information of the second working node. For brevity, the probability that the second working node is abnormal may be also referred to as the abnormal probability corresponding to the second working node. For example, the management processor 130 may determine a time difference between the reporting time of the second working node in the second node list and the current time. Then, the management processor 130 may determine the probability that the second working node is abnormal according to the time difference. Merely by way of example, the smaller the time difference corresponding to a second working node is, the smaller abnormal probability corresponding to the second working node may be. Further, the management processor 130 may determine the target second working node based on the abnormal probability corresponding to each second working node. For example, the management processor 130 may designate a second working node with the minimum abnormal probability as the target second working node.

[0078] In some embodiments, the management processor 130 may obtain feature information of each second working node in the second node list, and determine the target second working node based on the feature information of each second working node using a target node determination model. Exemplary feature information of a second working node may include the reporting time of the second working node in the second node list, the CPU usage, the memory usage, the input/output (IO) load, the network bandwidth, etc., used by the second working node, or the like, or any combination thereof. Specifically, the feature information of each of the second working nodes in the second node list may be input into the target node determination model, the target node determination model may directly output the target second working node and/or information relating to each second working node. For example, the formation relating to each second working node may be a recommendation score of each second working node. The management processor 130 may designate a second working node with the maximum score as the target second working node.

[0079] In some embodiments, the target node determination model may be a trained machine learning model. For example, the target node determination model may include a deep learning model, such as a Deep Neural Network (DNN) model, a Convolutional Neural Network (CNN) model, a Recurrent Neural Network

(RNN) model, a Feature Pyramid Network (FPN) model, etc. Exemplary CNN models may include a V-Net model, a U-Net model, a Link-Net model, or the like, or any combination thereof.

[0080] In some embodiments, the management processor 130 may obtain the target node determination model from one or more components of the cluster system 100 (e.g., a storage device, or an external source) via a network. For example, the target node determination model may be previously trained by a computing device, and stored in a storage device of the cluster system 100. The management processor 130 may access the storage device and retrieve the target node determination model. In some embodiments, the target node determination model may be generated by training a preliminary model based on a plurality of training samples. For example, each training sample may include sample feature information of a sample second working node and a reference score corresponding to the sample second working node, wherein the reference score can be used as a ground truth (also referred to as a label) for model training. In some embodiments, the reference score may be determined by a user or may be automatically determined by a training device.

[0081] In some embodiments, the management processor 130 may determine the target second working node combining a plurality of manners. For example, the management processor 130 may firstly obtain a plurality of candidate second working nodes using the target node determination model. Then, the management processor 130 may determine the target second working node based on the loads of the plurality of candidate second working nodes. As another example, the management processor 130 may firstly determine a plurality of candidate second working nodes with loads smaller than the load threshold. Further, the management processor 130 may determine the target second working node based on the abnormal probabilities corresponding to the plurality of candidate second working nodes. In this way, the determined target second working node may be more accurate.

[0082] Based on the load, the abnormal probability, and/or a recommendation score of each second working node, the target second working node selected from the second node list may be more suitable to take over the metadata services, which improves the whole operation efficiency of the cluster system.

[0083] In some embodiments, the management processor 130 may be communicated with a second management processor. When there is no target second working node in the second node list (e.g., there is no second working node in the second node list, the load of all second working nodes in the second node list is greater than the load threshold), the second management processor may be configured to determine a working node from a second cluster system, and designate the working node as the target second working node. In other words, when the nodes in the cluster system 100 cannot satisfy the availability of the metadata services, the methods of the present disclosure may obtain working nodes from other clusters to add them to the first node list. In this way, the maximum availability of the metadata services of the cluster system 100 may be determined based on a sum of the number of the first working nodes, the number of second working nodes, and the working nodes of other cluster systems, which may greatly improve the availability of metadata services. In some embodiments, the second management processor may determine the working node from the second cluster system in a similar manner as how to determine the target working node from the second node list, and the descriptions of which are not repeated here.

[0084] In some embodiments, the availability of the metadata services of the cluster system 100 depends on a count (number) of the first working nodes in the first node list. In some cases, the minimum count of first working nodes in the first node list may be set in advance. When some first working nodes in the first node list are abnormal, the count of first working nodes in the first node list may decrease. However, since some abnormal first working nodes are repaired and re-added into the first node list or some second working node are added from the second node list into the first working node list, etc., the count of first working nodes in the first node list may increase, and may even cause the count of first working nodes in the first node list to be greater than the minimum count, thus causing waste of resources.

[0085] Therefore, to avoid the waste of resources, in some embodiments, in response to determining that detecting that one of the one or more first working nodes is abnormal, the management processor 130 may determine whether the count of remaining first working nodes in the first node list other than the first working node is smaller than a count threshold. The count threshold may be set manually by a user (e.g., an engineer) according to an experience value or a default setting of the cluster system 100, such as 3, 5, or a larger or smaller value. In response to determining that the count of remaining first working nodes in the first node list is smaller than the count threshold, the management processor 130 may determine the target second working node from the second node list. In response to determining that the count of remaining first working nodes in the first node list is not smaller than the count threshold, the management processor 130 may does not add a second working node from the second node list into the first node list. In this case, the management processor 130 may only update the first node list by removing the abnormal first working node and deleting the report information of the abnormal first working node from the first node list.

[0086] In 550, the management processor 130 (e.g., the updating module 420) may update the first node list and the second node list based on the target second working node.

[0087] In some embodiments, the management processor 130 may directly designate the target second working node as a new first working node. The metadata of the main node may be synchronized to the new first working node.

[0088] The management processor 130 may update the first node list by performing the following operations. The management processor 130 may remove the abnormal first working node and delete the report information of the abnormal first working node from the first node list. The management processor 130 may add the new first working node and report information of the new first working node into the first node list. The management processor 130 may update the second node list by removing the target second working node and deleting the report information of the target second working node from the second node list.

[0089] For example, FIG. 7 is a schematic diagram illustrating an exemplary updating of a first node list 710 and a second node list 720 according to some embodiments of the present disclosure. As shown in FIG. 7, in response to detecting that a black first working node 1201 in the first node list 710 is abnormal, the management processor 130 may determine a grey second working node 1202 from the second node list 720. Then, the management processor 130 may update the first node list 710 by performing the following operations. The management processor 130 may remove the black first working node 1201 and delete the report information of the black first working node 1201 from the first node list 710. The management

processor 130 may add the grey second working node 1202 and the report information of the grey second working node 120 into the first node list 710. The management processor 130 may update the second node list 720 by removing the grey second working node 1202 and the report information of the grey second working node 1202 from the second node list 720.

[0090] In some embodiments, the abnormal first working node and the report information of the abnormal first working node may be removed from the first node list and added into a first deleting list. The target second working node and the report information of the target second working node may be removed from the second node list and added into a second deleting list. The management processor 130 may delete the abnormal first working node and the report information of the abnormal first working node from the first deleting list. The management processor 130 may delete the target second working node and the report information of the target second working node form the second deleting list. In some embodiments, the management processor 130 may monitor whether the abnormal first working node and the report information of the abnormal first working node are deleted completely. The management processor 130 may also monitor whether the target second working node and the report information of the target second working node are deleted completely.

[0091] In some embodiments, operations 520-550 may be performed in any sequence or simultaneously. When an operation of the operations 520-550 is performed, the management processor 130 may obtain the latest first node list and/or the latest second node list, and perform the operation based on the latest first node list and/or the latest second node list. In some embodiments, operations of the process 500 may be performed multiple times. For example, operations 510 and 520 may be performed each time the main node 110 receives report information from a working node. As another example, operations 540 and 550 may be performed each time the management processor 130 detects an abnormality of a first working node.

[0092] FIG. 8 is a flowchart illustrating an exemplary process 800 for managing nodes in a cluster system according to some embodiments of the present disclosure. In some embodiments, the process 800 may be performed by a management processor 130 independent from the main node.

[0093] In 810, the management processor 130 (e.g., the acquisition module 410) may obtain information relating to the main node.

[0094] In some embodiments, the information relating to the main node may include an internet protocol (IP) address of the main node, a metadata service state of the main node, a mark indicating that it is a main node, a version number of the main node, or the like, or any combination thereof. In some embodiments, the management processor 130 may obtain the information relating to the main node periodically or aperiodically.

[0095] In 820, the management processor 130 (e.g., the determination module 430) may determine whether the main node is abnormal based on the information relating to the main node.

[0096] For example, the management processor 130 may determine whether the main node is abnormal according to the metadata service state of the main node. In response to determining that the metadata service state of the main node is an abnormal state, the management processor 130 may determine that the main node is abnormal. As another example, the management processor 130 may determine that the main node is abnormal if it haven't received information from the main node for more than a predetermined period. In

response to determining that the main node is abnormal, the management processor 130 may perform operations 830 and 840.

[0097] In 830, the management processor 130 (e.g., the determination module 430) may determine a target first working node that performs the metadata services of the main node in place of the main node.

[0098] In some embodiments, when the main node is abnormal, one first working node in the first node list may automatically replace the main node 110 to perform the metadata services of the cluster system 100 according to a preset rule. The management processor 130 may determine the first working node that replaces the main node as the target first working node. For example, when one first working node replaces the main node 110 to perform the metadata services of the cluster system 100, the management processor 130 may update the mark indicating that the first working node is a standby node with a mark indicating that the first working node is a main node. The management processor 130 may determine the target first working node based on the first node list. In some embodiments, when the main node is abnormal, the management processor 130 may select one first working node from the first node list, and designate the first working node as a new main node (i.e., the target first working node). For example, the target first working node may be determined based on the load of each first working node, the probability that each first working node is abnormal, or the like, or any combination thereof, which is similar to how the target second working node is selected from the second node list.

[0099] In 840, the management processor 130 (e.g., the updating module 420) may update the first node list and the second node list based on the target first working node.

[0100] In some embodiments, the management processor 130 may update the first node list by performing the following operations. The management processor 130 may remove the target first working node and delete the report information of the target first working node from the first node list. The management processor 130 may determine a reference second working node from the second node list. In some embodiments, the determination of the reference second working node may be performed in a similar manner as that of the target second working node, and the descriptions thereof are not repeated here. The management processor 130 may designate the reference second node as a new first working node, and add the new first working node and the report information of the new first working node into the first node list. The management processor 130 may update the second node list by removing the reference second working node and deleting the report information of the reference second working node from the second node list.

[0101] For example, FIG. 9 is a schematic diagram illustrating an exemplary updating of a first node list 910 and a second node list 920 according to some embodiments of the present disclosure. As shown in FIG. 9, in response to detecting that the main node 110 is abnormal, the management processor 130 may determine a black first working node 1201 in the first node list 910 that performs the metadata services of the main node 110 in place of the main node 110. The management processor 130 may update the first node list 910 by performing the following operations. The management processor 130 may remove the black first working node 1201 and delete the report information of the black first working node 1201 from the first node list 910. The management processor 130 may determine a gray second working node 1202 from the second node list 920. The management processor 130 may add the gray second working node 1202 and the report information

of the gray second working node 1202 into the first node list 910. The management processor 130 may update the second node list 920 by removing the gray second working node 1202 and deleting the report information of the gray second working node 1202 from the second node list 920.

[0102] As described elsewhere in the present disclosure, in the conventional approach for metadata services, the availability of metadata services depends on a number of standby nodes that can be used to replace the main node to provide metadata services.

[0103] According to the present systems and methods, the availability of metadata services depends not only on the number of standby nodes (i.e., the first working nodes) but also on the number of the second working nodes, and the maximum availability of the metadata services of the cluster system may be determined based on a sum of the number of the first working nodes and the number of the second working nodes. Compared with the conventional approach, the present systems and methods may greatly improve the availability of metadata services.

[0104] Having thus described the basic concepts, it may be rather apparent to those skilled in the art after reading this detailed disclosure that the foregoing detailed disclosure is intended to be presented by way of example only and is not limiting. Various alterations, improvements, and modifications may occur and are intended to those skilled in the art, though not expressly stated herein. These alterations, improvements, and modifications are intended to be suggested by this disclosure, and are within the spirit and scope of the exemplary embodiments of this disclosure.

[0105] Moreover, certain terminology has been used to describe embodiments of the present disclosure. For example, the terms "one embodiment," "an embodiment," and/or "some embodiments" may mean that a particular feature, structure or characteristic described in connection with the embodiment is included in at least one embodiment of the present disclosure. Therefore, it is emphasized and should be appreciated that two or more references to "an embodiment" or "one embodiment" or "an alternative embodiment" in various portions of this specification are not necessarily all referring to the same embodiment. Furthermore, the particular features, structures or characteristics may be combined as suitable in one or more embodiments of the present disclosure.

[0106] Further, it will be appreciated by one skilled in the art, aspects of the present disclosure may be illustrated and described herein in any of a number of patentable classes or context including any new and useful process, machine, manufacture, or composition of matter, or any new and useful improvement thereof. Accordingly, aspects of the present disclosure may be implemented entirely hardware, entirely software (including firmware, resident software, micro-code, etc.) or combining software and hardware implementation that may all generally be referred to herein as a "unit," "module," or "system." Furthermore, aspects of the present disclosure may take the form of a computer program product embodied in one or more computer readable media having computer readable program code embodied thereon.

[0107] A computer readable signal medium may include a propagated data signal with computer readable program code embodied therein, for example, in baseband or as part of carrier wave. Such a propagated signal may take any of a variety of forms, including electro-magnetic, optical, or the like, or any suitable combination thereof. A computer readable signal medium may be any computer readable medium that is not a computer

readable storage medium and that may communicate, propagate, or transport a program for use by or in connection with an instruction execution system, apparatus, or device. Program code embodied on a computer readable signal medium may be transmitted using any appropriate medium, including wireless, wireline, optical fiber cable, RF, or the like, or any suitable combination of the foregoing.

[0108] Computer program code for carrying out operations for aspects of the present disclosure may be written in any combination of one or more programming languages, including an object oriented programming language such as Java, Scala, Smalltalk, Eiffel, JADE, Emerald, C++, C#, VB. NET, Python or the like, conventional procedural programming languages, such as the "C" programming language, Visual Basic, Fortran 2103, Perl, COBOL 2102, PHP, ABAP, dynamic programming languages such as Python, Ruby and Groovy, or other programming languages. The program code may execute entirely on the user's computer, partly on the user's computer, as a stand-alone software package, partly on the user's computer and partly on a remote computer or entirely on the remote computer or server. In the latter scenario, the remote computer may be connected to the user's computer through any type of network, including a local area network (LAN) or a wide area network (WAN), or the connection may be made to an external computer (for example, through the Internet using an Internet Service Provider) or in a cloud computing environment or offered as a service such as a Software as a Service (SaaS).

[0109] Furthermore, the recited order of processing elements or sequences, or the use of numbers, letters, or other designations therefore, is not intended to limit the claimed processes and methods to any order except as may be specified in the claims. Although the above disclosure discusses through various examples what is currently considered to be a variety of useful embodiments of the disclosure, it is to be understood that such detail is solely for that purpose, and that the appended claims are not limited to the disclosed embodiments, but, on the contrary, are intended to cover modifications and equivalent arrangements that are within the spirit and scope of the disclosed embodiments. For example, although the implementation of various components described above may be embodied in a hardware device, it may also be implemented as a software only solution, for example, an installation on an existing server or mobile device.

[0110] Similarly, it should be appreciated that in the foregoing description of embodiments of the present disclosure, various features are sometimes grouped together in a single embodiment, figure, or description thereof for the purpose of streamlining the disclosure aiding in the understanding of one or more of the various inventive embodiments. This method of disclosure, however, is not to be interpreted as reflecting an intention that the claimed object matter requires more features than are expressly recited in each claim. Rather, inventive embodiments lie in less than all features of a single foregoing disclosed embodiment.

[0111] In some embodiments, the numbers expressing quantities or properties used to describe and claim certain embodiments of the application are to be understood as being modified in some instances by the term "about," "approximate," or "substantially." For example, "about," "approximate," or "substantially" may indicate ±1%, ±5%, ±10%, or ±20% variation of the value it describes, unless otherwise stated. Accordingly, in some embodiments, the numerical parameters set forth in the written description and attached claims are approximations that may vary depending upon the desired properties sought to be obtained by a particular embodiment. In some embodiments, the numerical parameters should be construed in light of the number of

reported significant digits and by applying ordinary rounding techniques. Notwithstanding that the numerical ranges and parameters setting forth the broad scope of some embodiments of the application are approximations, the numerical values set forth in the specific examples are reported as precisely as practicable.

[0112] Each of the patents, patent applications, publications of patent applications, and other material, such as articles, books, specifications, publications, documents, things, and/or the like, referenced herein is hereby incorporated herein by this reference in its entirety for all purposes, excepting any prosecution file history associated with same, any of same that is inconsistent with or in conflict with the present document, or any of same that may have a limiting effect as to the broadest scope of the claims now or later associated with the present document. By way of example, should there be any inconsistency or conflict between the description, definition, and/or the use of a term associated with any of the incorporated material and that associated with the present document, the description, definition, and/or the use of the term in the present document shall prevail.

[0113] In closing, it is to be understood that the embodiments of the application disclosed herein are illustrative of the principles of the embodiments of the application. Other modifications that may be employed may be within the scope of the application. Thus, by way of example, but not of limitation, alternative configurations of the embodiments of the application may be utilized in accordance with the teachings herein. Accordingly, embodiments of the present application are not limited to that precisely as shown and described.

WHAT IS CLAIMED IS:

1. A cluster system, comprising:

a main node configured for providing metadata services;

working nodes each of which is communicatively connected to the main node and configured to send report information to the main node, the working nodes including one or more first working nodes, the one or more first working nodes being standby nodes of the main node configured for metadata backup; and

a management processor configured to:

update, based on the report information of each working node, a first node list and a second node list, the first node list being relating to the one or more first working nodes, and the second node list being relating to one or more second working nodes other than the one or more first working nodes among the working nodes; and

in response to detecting that one of the one or more first working nodes is abnormal, determine a target second working node from the second node list, designate the target second working node as a new first working node, and update the first node list and the second node list.

2. The cluster system of claim 1, wherein

the management processor is part of the main node, or

the management processor is independent from the main node and configured to receive the report information of each of the working nodes from the main node.

3. The cluster system of claim 1, wherein the management processor may be further configured to:

in response to detecting that one of the one or more second working nodes is abnormal, remove the abnormal second working node from the second node list.

4. The cluster system of any one of claims 1-3, wherein the management processor configured to update, based on the report information of each working node, a first node list and a second node list by:

for each working node,

determining, based on the report information of the working node, whether the working node is a first working node or a second working node; and

in response to determining that the working node is a first working node, updating the first node list based on the report information of the working node; or

in response to determining that the working node is a second working node, updating the second node list based on the report information of the working node.

5. The cluster system of any one of claims 1-4, wherein the target second working node is determined from the second node list by:

for each second working node in the second node list, determining a load of the second working node; and

determining, based on the load of each second working node, the target second working node.

6. The cluster system of any one of claims 1-4, wherein the target second working node is determined from the second node list by:

for each second working node in the second node list, determining a probability that the second working node is abnormal based on the report information of the second working node; and

determining, based on the probability corresponding to each second working node, the target second working node.

7. The cluster system of any one of claims 1-4, wherein the target second working node is determined from the second node list by:

obtaining feature information of each second working node in the second node list; and

determining, based on the feature information of each second working node, the target second working node using a target node determination model, the target node determination model being a trained machine learning model.

8. The cluster system of any one of claims 1-7, wherein in response to detecting that one of the one or more first working nodes is abnormal, to determine a target second working node from the second node list, the management processor may be further configured to:

in response to determining that detecting that one of the one or more first working nodes is abnormal, determining whether the count of remaining first working nodes in the first node list other than the first

working node is smaller than a count threshold;

in response to determining that the count of remaining first working nodes in the first node list is smaller than a count threshold, determining the target second working node from the second node list.

9. The cluster system of any one of claims 1-8, wherein the management processor may be further configured to,

in response to detecting that the main node is abnormal,

determine a target first working node that performs the metadata services of the main node in place of the main node; and

update the first node list and the second node list based on the target first working node.

10. The cluster system of any one of claims 1-9, wherein

the management processor is communicated with a second management processor,

when there is no target second working node in the second node list, the second management processor is configured to determine a working node from a second cluster system, and designate the working node as the target second working node.

11. A method implemented on a management processor of a cluster system, the cluster system further comprising a main node configured for providing metadata services and

working nodes each of which is communicatively connected to the main node and configured to send report information to the main node, the working nodes including one or more first working nodes, the one or more first working nodes being standby nodes of the main node configured for metadata backup, and the method comprising:

updating, based on the report information of each working node, a first node list and a second node list, the first node list being relating to the one or more first working nodes, and the second node list being relating to one or more second working nodes other than the one or more first working nodes among the working nodes; and

in response to detecting that one of the one or more first working nodes is abnormal, determining a target second working node from the second node list, designating the target second working node as a new first working node, and updating the first node list and the second node list.

12. The method of claim 11, wherein

the management processor is part of the main node, or

the management processor is independent from the main node and configured to receive the report information of each of the working nodes from the main node.

13. The method of claim 11, wherein the method further comprising:

in response to detecting that one of the one or more second working nodes is abnormal, removing the abnormal second working node from the second node list.

14. The method of any one of claims 11-13, wherein the updating, based on the report information of each working node, a first node list and a second node list including:

for each working node,

determining, based on the report information of the working node, whether the working node is a first working node or a second working node; and

in response to determining that the working node is a first working node, updating the first node list based on the report information of the working node; or

in response to determining that the working node is a second working node, updating the second node list based on the report information of the working node.

15. The method of any one of claims 11-14, wherein the determining a target second working node from the second node list including:

for each second working node in the second node list, determining a load of the second working node; and

determining, based on the load of each second working node, the target second working node.

16. The method of any one of claims 11-14, wherein the determining a target second working node from the second node list including:

for each second working node in the second node list, determining a probability that the second working node is abnormal based on the report information of the second working node; and

determining, based on the probability corresponding to each second working node, the target second working node.

17. The method of any one of claims 11-14, wherein the determining a target second working node from the second node list including:

obtaining feature information of each second working node in the second node list; and

determining, based on the feature information of each second working node, the target second working node using a target node determination model, the target node determination model being a trained machine learning model.

18. The method of any one of claims 11-17, wherein in response to detecting that one of the one or more first working nodes is abnormal, to determine a target second working node from the second node list, the method further comprising:

in response to determining that detecting that one of the one or more first working nodes is abnormal, determining whether the count of remaining first working nodes in the first node list other than the first working node is smaller than a count threshold;

in response to determining that the count of remaining first working nodes in the first node list is smaller than a count threshold, determining the target second working node from the second node list.

19. The method of any one of claims 11-18, wherein the method further comprising,

in response to detecting that the main node is abnormal,

determining a target first working node that performs the metadata services of the main node in place of the main node; and

updating the first node list and the second node list based on the target first working node.

20. The method of any one of claims 11-19, wherein

the management processor is communicated with a second management processor,

when there is no target second working node in the second node list, the second management processor is configured to determine a working node from a second cluster system, and designate the working node as the target second working node.

21. A non-transitory computer readable medium, comprising a set of instructions, the set of instructions being executed by a management processor of a cluster system, the cluster system further comprising a main node configured for providing metadata services and

working nodes each of which is communicatively connected to the main node and configured to send report information to the main node, the working nodes including one or more first working nodes, the one or more first working nodes being standby nodes of the main node configured for metadata backup, wherein when the set of instructions are executed by the management processor, the set of instructions causes the management processor to perform a method, and the method comprising:

updating, based on the report information of each working node, a first node list and a second node list, the first node list being relating to the one or more first working nodes, and the second node list being relating to one or more second working nodes other than the one or more first working nodes among the working nodes; and

in response to detecting that one of the one or more first working nodes is abnormal, determining a target second working node from the second node list, designating the target second working node as a new first working node, and updating the first node list and the second node list.

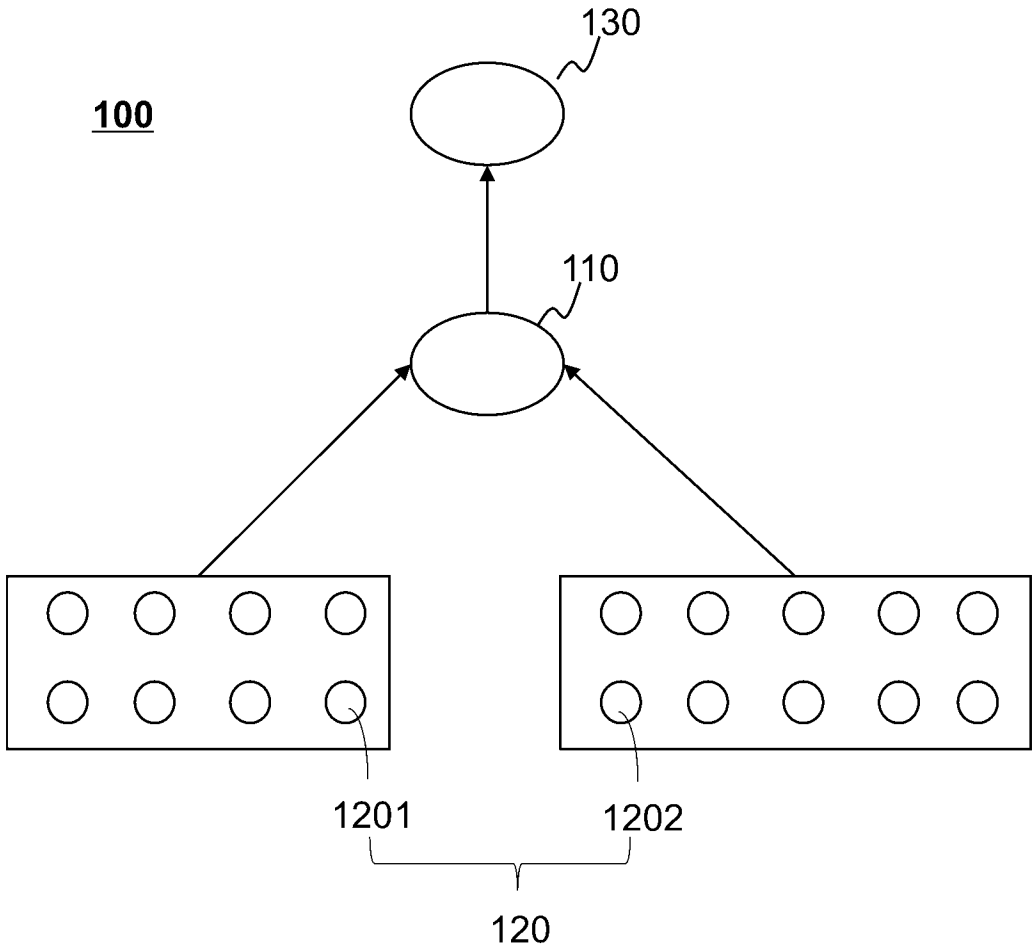


FIG. 1

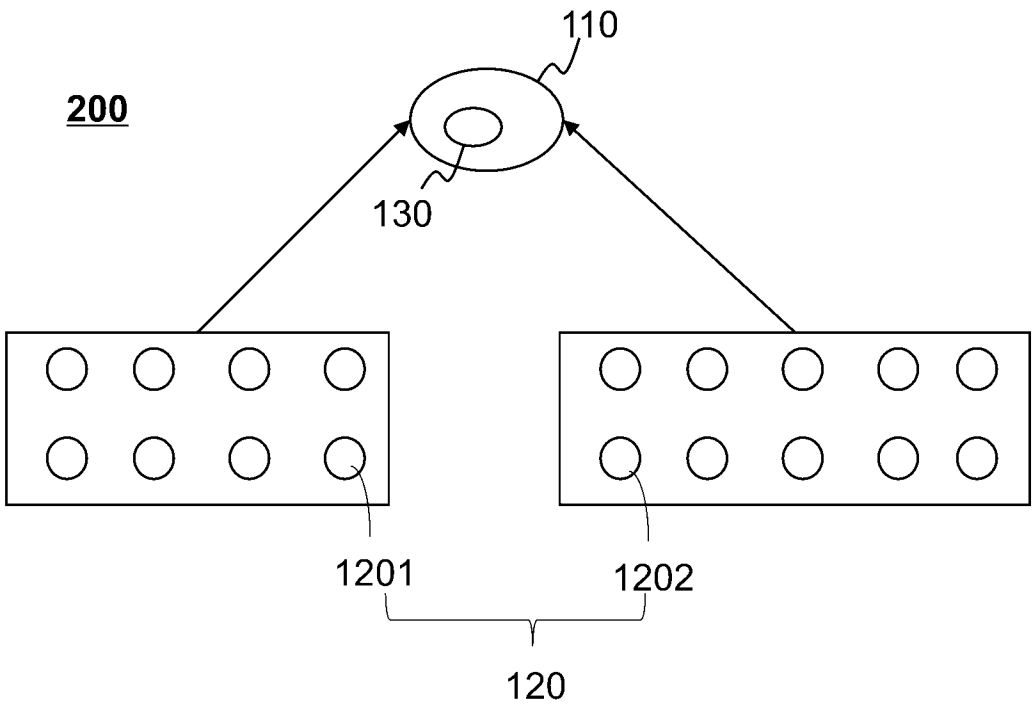


FIG. 2

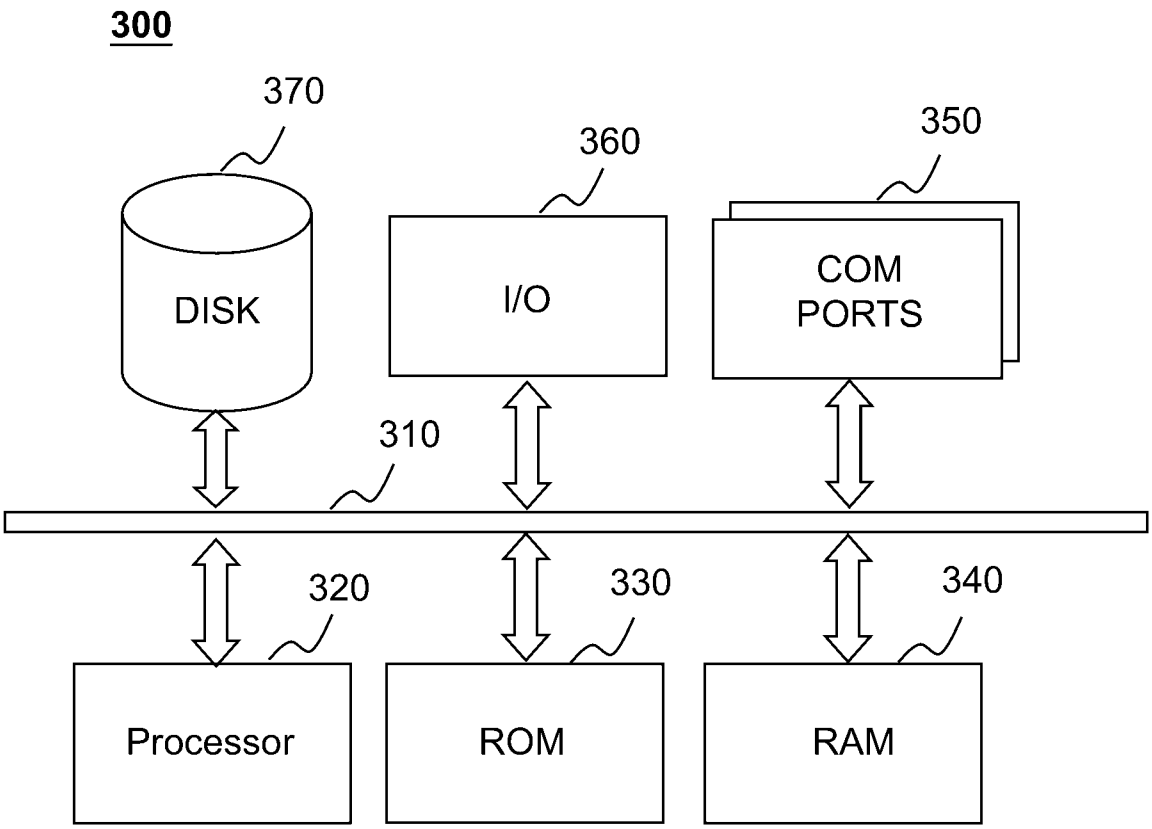

300
370
360
350
DISK
I/O
COM
PORTS
310
320
330
340
Processor
ROM
RAM

FIG. 3

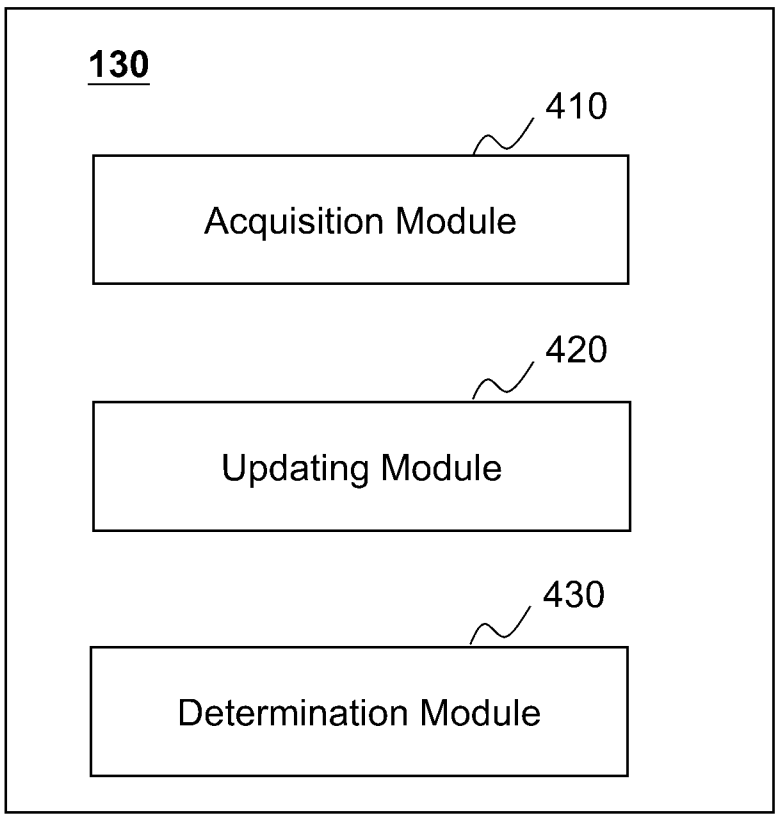

130
410
Acquisition Module
420
Updating Module
430
Determination Module

FIG. 4

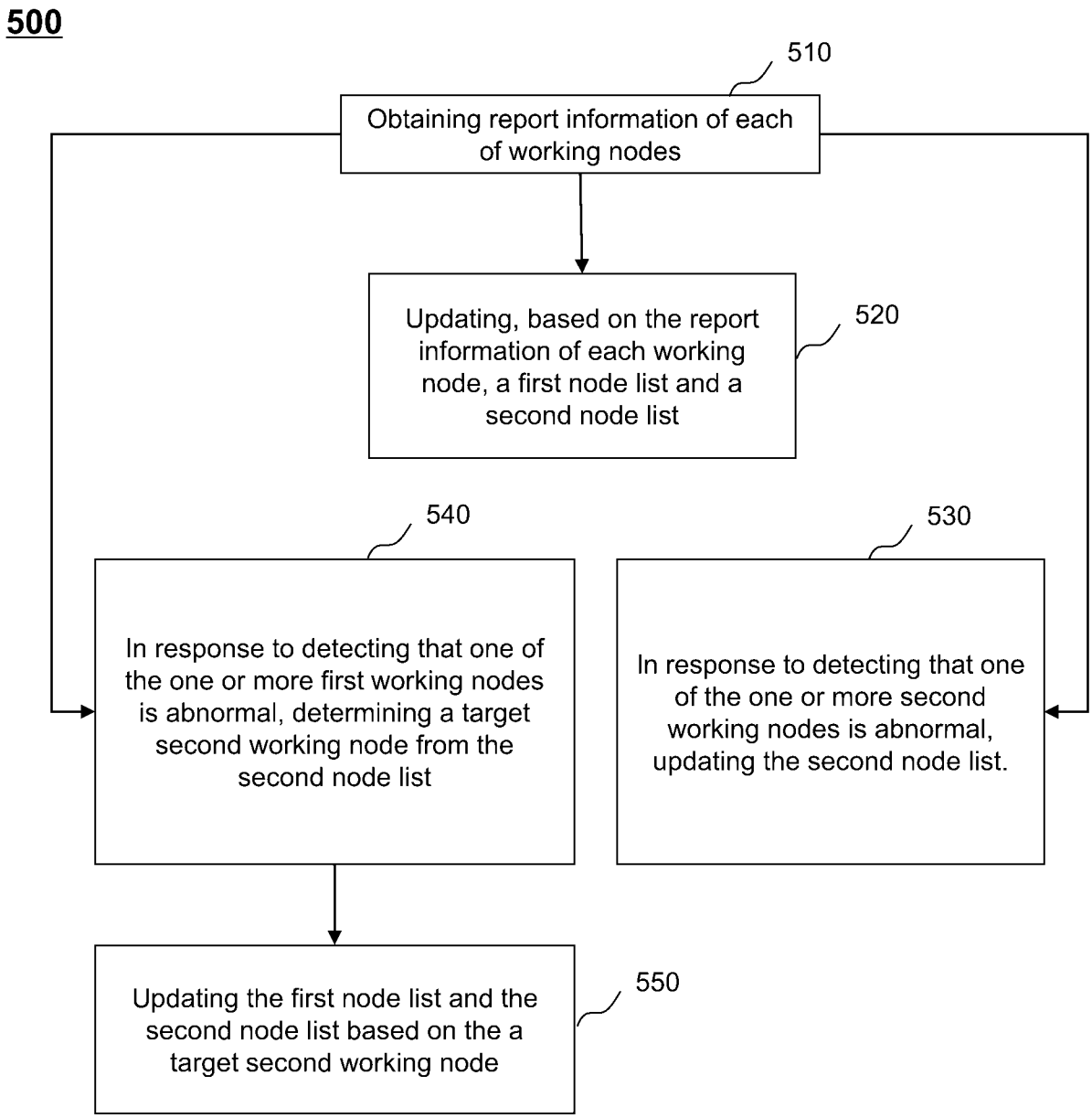

500
510
Obtaining report information of each of working nodes
Updating, based on the report information of each working node, a first node list and a second node list
520
540
530
In response to detecting that one of the one or more first working nodes is abnormal, determining a target second working node from the second node list
In response to detecting that one of the one or more second working nodes is abnormal, updating the second node list.
Updating the first node list and the second node list based on the a target second working node
550

FIG. 5

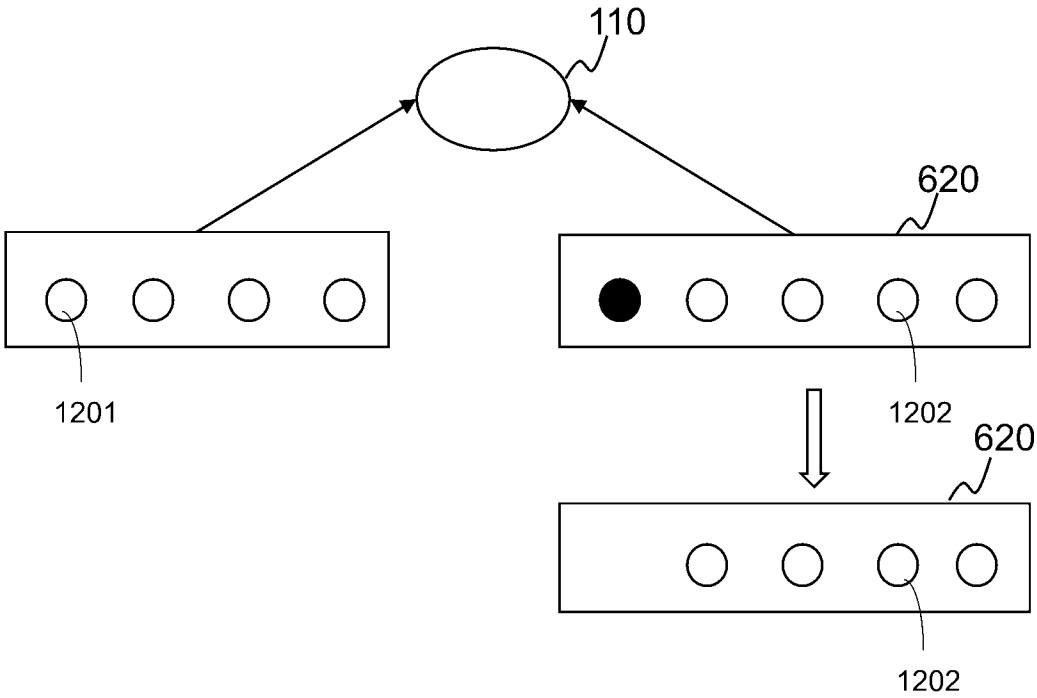

110
620
1201
1202
620
1202

FIG. 6

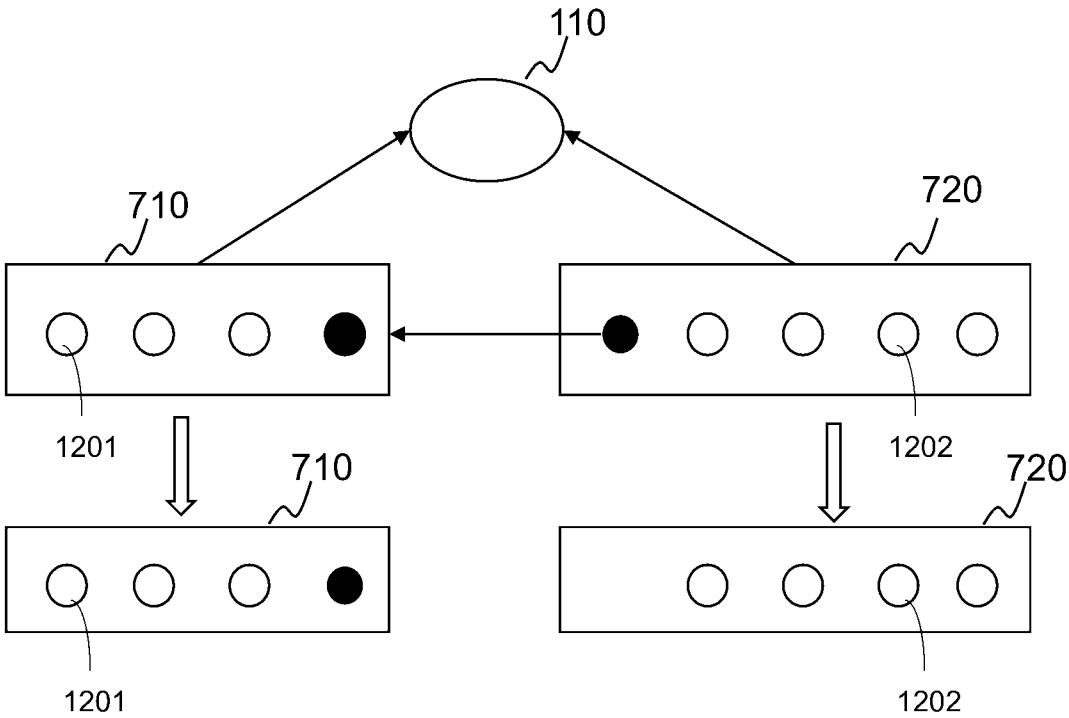

110
710
720
1201
1202
710
720
1201
1202

FIG. 7

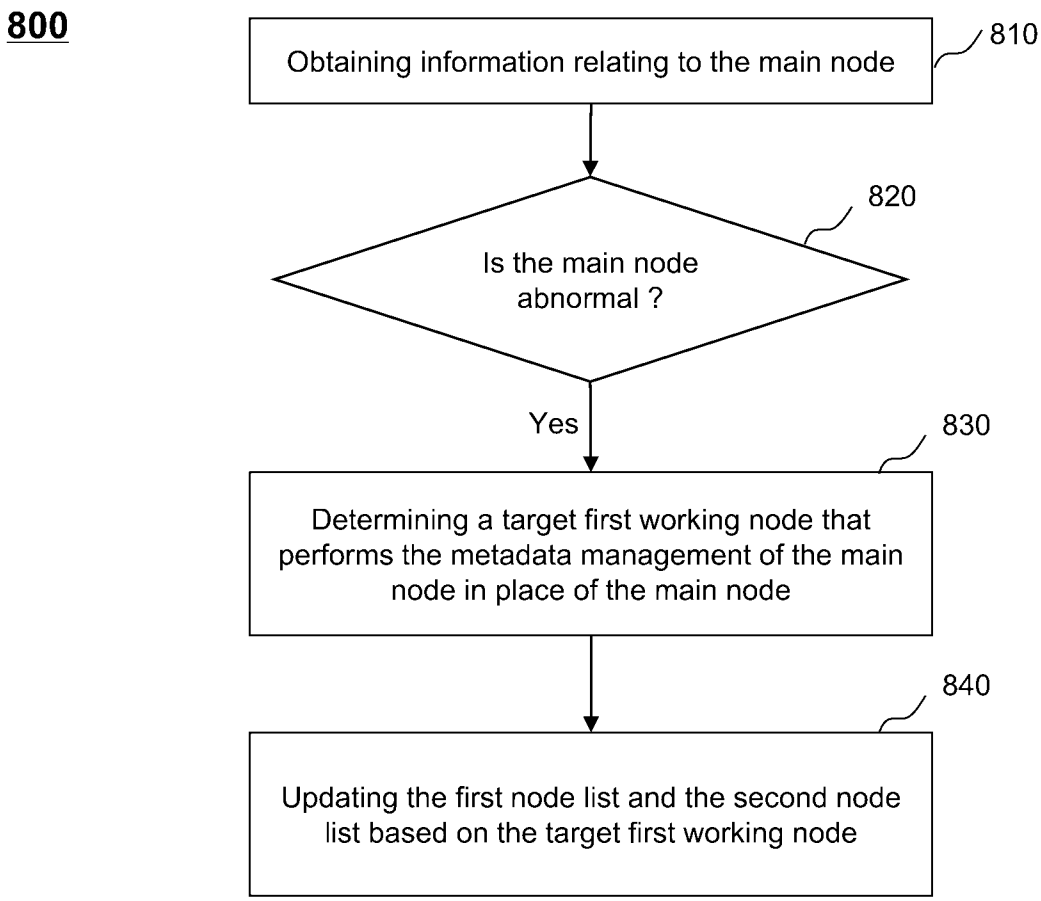

800
Obtaining information relating to the main node
810
Is the main node abnormal ?
820
Yes
Determining a target first working node that performs the metadata management of the main node in place of the main node
830
Updating the first node list and the second node list based on the target first working node
840

FIG. 8

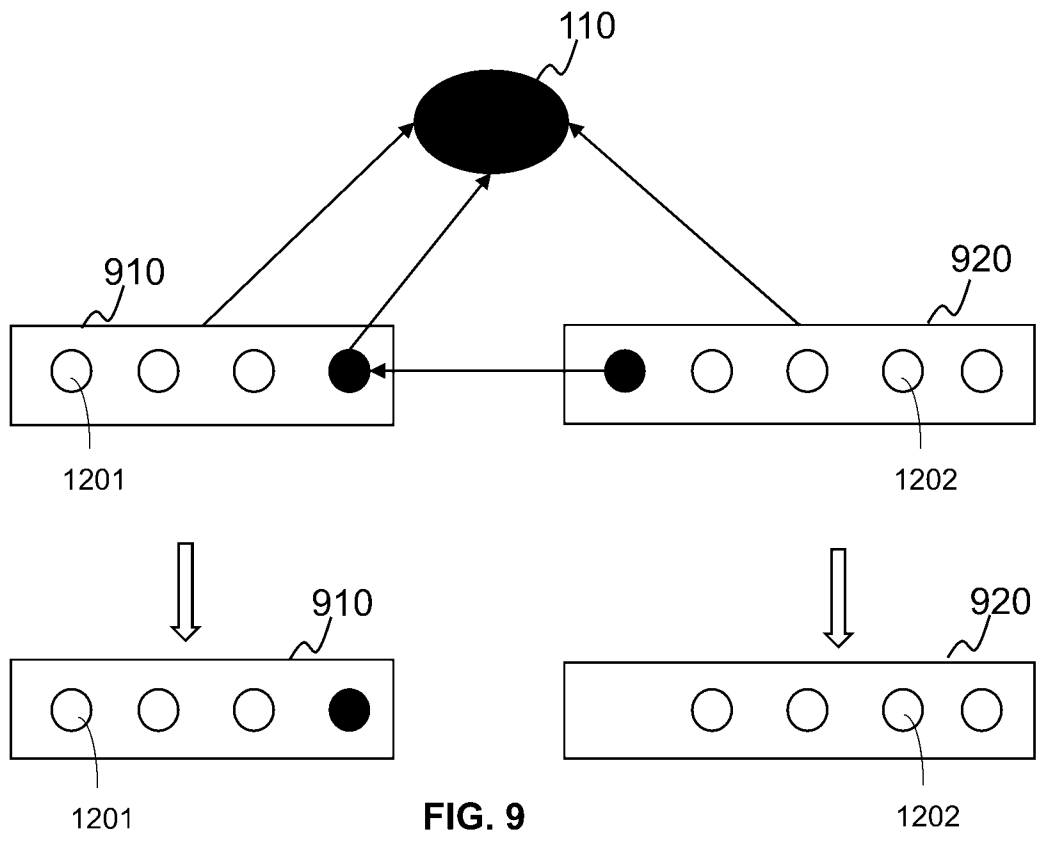

110
910
920
1201
1202
910
920
1201
1202

FIG. 9

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2023/109498

A. CLASSIFICATION OF SUBJECT MATTER

G06F 3/06(2006.01)i

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)

IPC: 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)

CNABS,CNTXT,ENTXT,ENTXTC,DWPI,WPABS,CNKI:node, main, slave, standby, metadata, abnormal, fault, replace, update, list, availability.

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	CN 112190924 A (TENCENT TECHNOLOGY SHENZHEN CO LTD) 08 January 2021 (2021-01-08) The whole document	1-21
A	CN 112463448 A (SUZHOU INSPUR INTELLIGENT TECHNOLOGY CO) 09 March 2021 (2021-03-09) The whole document	1-21
A	CN 114070739 A (HANGZHOU HOLLYSYS AUTOMATION CO LTD) 18 February 2022 (2022-02-18) The whole document	1-21
A	US 2020174665 A1 (EMC IP HOLDING CO LLC) 04 June 2020 (2020-06-04) The whole document	1-21
PX	CN 115268785 A (ZHEJIANG DAHUA TECHNOLOGY CO LTD) 01 November 2022 (2022-11-01) Description, paragraphs 47-114	1-21

☐ 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

"D" document cited by the applicant in the international application

"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
17 October 2023	**23 October 2023**
Name and mailing address of the ISA/CN **CHINA NATIONAL INTELLECTUAL PROPERTY ADMINISTRATION** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing 100088, China**	Authorized officer **ZHAO,Shang** Telephone No. **(+86) 010-62411281**

Form PCT/ISA/210 (second sheet) (July 2022)

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2023/109498

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
CN 112190924 A	08 January 2021	None	
CN 112463448 A	09 March 2021	CN 112463448 B	07 June 2022
CN 114070739 A	18 February 2022	None	
US 2020174665 A1	04 June 2020	US 10845991 B2	24 November 2020
CN 115268785 A	01 November 2022	None	