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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 June 2017 (29.06.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/112032 A1

(51) International Patent Classification:
H04L 29/08 (2006.01)

(21) International Application Number:
PCT/US2016/056349

(22) International Filing Date:
11 October 2016 (11.10.2016)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
14/979,268 22 December 2015 (22.12.2015) US

(71) Applicant: **DROPBOX, INC.** [US/US]; 333 Brannan Street, San Francisco, California 94107 (US).

(72) Inventors: **KOORAPATI, Nipunn**; 333 Brannan Street, San Francisco, California 94107 (US). **RUDE, Christopher**; 333 Brannan Street, San Francisco, California 94107 (US). **VON MUHLEN, Marcio**; 333 Brannan Street, San Francisco, California 94107 (US). **BUNGER, Nils**; 333 Brannan Street, San Francisco, California 94107 (US).

(74) Agents: **STONE, Adam C.** et al.; Hickman Palermo Becker Bingham LLP, 1 Almaden Boulevard, San Jose, California 95113 (US).

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

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, 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, 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))*

(54) Title: CONTENT ITEM BLOCK REPLICATION PROTOCOL FOR MULTI-PREMISES HOSTING OF DIGITAL CONTENT ITEMS

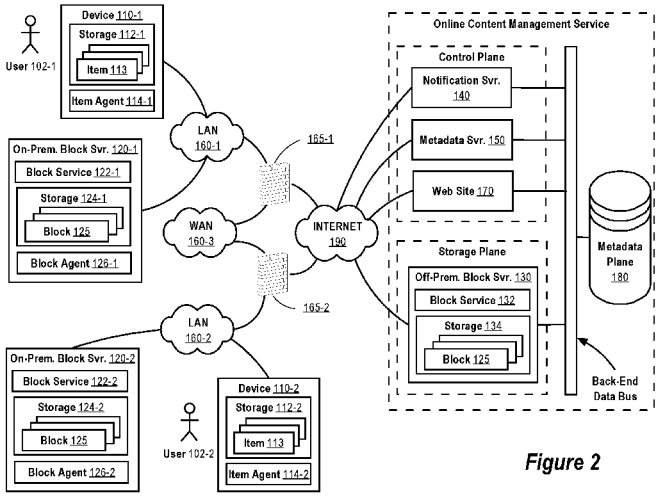

Figure 2

(57) Abstract: A content item block replication protocol for multi-premises hosting of digital content items. In one embodiment, for example, a method comprises: receiving, from a server, a server journal entry identifying one or more content item blocks of a content item represented by the server journal entry; storing a replication task log entry corresponding to the server journal entry in a replication task log, the replication task log entry identifying the one or more content item blocks of the content item represented by the server journal entry and identifying a block server; and either offering to send the one or more content item blocks identified in the replication task log entry to the block server identified in the replication task log entry, or downloading the one or more content item blocks identified in the replication task log entry from the block server identified in the replication task log entry.

WO 2017/112032 A1

CONTENT ITEM BLOCK REPLICATION PROTOCOL FOR MULTI-PREMISES HOSTING OF DIGITAL CONTENT ITEMS

TECHNICAL FIELD

[0001] The present Application relates to management of digital content items. More specifically, the example embodiment(s) of the present invention described below relate to the management of digital content items hosted with an online content management service.

BACKGROUND

[0002] Traditionally, businesses have stored their digital content items (e.g., documents, files, and other digital information) on network file servers they own and operate. Such file servers are typically located on-premises behind a network firewall that prevent unauthorized network access to the file servers. This arrangement works well when most or all of the network access to the file server is by computers that are also behind the network firewall such as, for example, connected to the same Local Area Network (LAN) as the file server. In some cases, network access to the file server from outside the firewall (e.g., over the Internet) is facilitated by a Virtual Private Network (VPN). The VPN, in effect, makes a computer outside the firewall appear to the file server as if it is behind the firewall.

[0003] Today, however, the workforce is more global and more mobile. This is spurred, in large part, by the wide availability of broadband Internet connectivity and also the availability of relatively inexpensive, yet powerful, portable personal computing devices such as, for example, mobile phones, laptop computers, and tablet computers. The result is employees can work virtually anywhere and do not need to be physically present in the office to get their work done (e.g., they can work remotely).

[0004] Recently, online content management services have become available for storing content items "online" where they are accessible on the Internet or other network. A business can use an online content management service to "host" their content items on servers operated by the service. One example of an online content management service is the "Dropbox" service provided by Dropbox, Inc. of San Francisco, California.

[0005] Online storage of content items can provide a number of benefits to businesses and their employees alike. Dropbox, for instance, offers the ability to synchronize and share hosted content items among multiple devices and users. This flexibility, which stems from storing content items both at end-user devices and on Dropbox servers, supports a variety of different on-site and remote working arrangements, providing convenience to employees and increased employee productivity for employers.

[0006] Given the increasing amount of digital information generated by businesses, hosting of content items by businesses with online content management services can only be expected to increase. However, due to the sensitive nature of some of the hosted information, users of such services, including business users, would appreciate improvements to the services that provide a greater level of control over the content items they host with the services. In addition, while broadband Internet connectivity is pervasive among businesses today, uploading and downloading content items over the Internet can still take longer than users expect or desire.

[0007] The approaches described in this section are approaches that could be pursued, but not necessarily approaches that have been previously conceived or pursued. Therefore, unless otherwise indicated, it should not be assumed that any of the approaches described in this section qualify as prior art merely by virtue of their inclusion in this section.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] The example embodiment(s) of the present invention are illustrated by way of example, and not in way by limitation, in the figures of the accompanying drawings and in which like reference numerals refer to similar elements and in which:

[0009] Figure 1 is a block diagram of an example system environment in which some example embodiments of the present invention are implemented.

[0010] Figure 2 is a block diagram of an example system environment in which some example embodiments of the present invention are implemented.

[0011] Figure 3 is a block diagram of content item block replication metadata, according to some example embodiments of the present invention.

[0012] Figure 4 is a block diagram of a content item server journal, according to some example embodiments of the present invention.

[0013] Figure 5 is a block diagram of a content item block replication log, according to some example embodiments of the present invention.

[0014] Figure 6 is a flow diagram of a process for providing content item block replication tasks to an on-premises block server, according to some example embodiments of the present invention.

[0015] Figures 7A-C comprise a single flowchart illustrating operation of the block agent at an on-premises block server in processing replication tasks specified in a replication log stored at the on-premises block server, according to some example embodiments of the present invention.

[0016] Figure 8 is a very general block diagram of a computing device in which the example embodiment(s) of the present invention can be embodied.

[0017] Figure 9 is a block diagram of a basic software system for controlling the operation of the computing device.

DESCRIPTION OF THE EXAMPLE EMBODIMENT(S)

[0018] In the following description, for the purposes of explanation, numerous specific details are set forth in order to provide a thorough understanding of the example embodiment(s) the present invention. It will be apparent, however, that the example embodiment(s) can be practiced without these specific details. In other instances, well-known structures and devices are shown in block diagram form in order to avoid unnecessarily obscuring the example embodiment(s).

[0019] The example embodiments are described according to the following outline:

1.0 ILLUSTRATIVE EXAMPLES

[0020] Example embodiments of the present invention provide users of online content management services with greater control over where their content items are hosted with such services. In the following, a number of computer-implemented processes and network interactions are described. To help in describing those process and network interactions, some illustrative example users will now be introduced. The examples will be used to illustrate features of some example embodiments of the present invention and to aid in describing certain features of some example embodiments of the present invention. The examples are not intended to be limiting and are merely provided for illustration.

[0021] A first example user is referred to herein as “Corporation Alpha.” As a first example, Corporation Alpha has a number of employees that use an online content management service to

synchronize content items stored on their work computers with content items stored on servers on the Internet operated by the service. Corporation Alpha likes that, because the content items are stored at their work computers, the employees have access to the content items when their work computers are not connected to the Internet. At the same time, Corporation Alpha also likes that the content items are synced to online content management service servers for backup and sharing purposes. Corporation Alpha also prefers, when possible, to reduce the time needed to synchronize content item changes between employee work computers.

[0022] As a second example, the employees of Corporation Alpha may be distributed geographically. For example, Corporation Alpha's headquarters may be in San Francisco but may also have satellite offices in New York, Austin, and Seattle. Teams within Corporation Alpha may also be distributed geographically. For example, employees in the San Francisco and Austin offices may be collaborating on a project together. Corporation Alpha would prefer that changes to content items that the team collaborates on are quickly synchronized between computers at the San Francisco and Austin offices.

[0023] As a third example, the project the employees of Corporation Alpha in San Francisco and Austin are collaborating on together may be highly sensitive (e.g., confidential). In this case, Corporation Alpha may prefer to retain more control and oversight over the content items associated with the project. For example, Corporation Alpha may prefer that the project content items be stored on-premises only and not on servers operated by the online content management service.

[0024] Using features of the present invention, users such as Corporation Alpha and other users can control where their content items managed by an online content management service are hosted. In particular, example embodiments allow users to host their content items on-premises only, off-premises only, or both on-premises and off-premises. When hosting content items on-premises, users can use their own content item storage hardware (although some example embodiments involve the online content management service providing on-premises content item storage hardware). Example embodiments allow end-user devices to synchronize content item changes made at the end-user devices to on-premises storage, to off-premises storage, or to both on-premises storage and off-premises storage. Example embodiments also allow end-user devices to synchronize content item changes made by other end-user devices from on-premises storage, from off-premises storage, or from both on-premises storage and off-premises storage.

[0025] These illustrative examples are used in conjunction with some of the following description to aid in describing features of some example embodiments of the present invention.

2.0 EXAMPLE SYSTEM ENVIRONMENT

[0026] Features of some example embodiments of the present invention will now be described by reference to Figure 1, which is a block diagram of an example system environment 100 in which some example embodiments of the present invention may be implemented. As shown, a user (e.g., 102-1) can have or use a personal computing device (e.g., 110-1). A personal computing device can have a local storage (e.g., 112-1) and a content item synchronization agent (e.g., 114-1). A local storage of a personal computing device can store one or more content items 113. An on-premises block server 120 can be composed of a block service 122, block storage 124, and a block management agent 126. The block storage 124 can store one or more content item blocks 125. An online content management service can be composed of a control plane, a storage plane, and a data plane. The control plane can include a notification server 140, a metadata server 150, and a web site 170. The storage plane can include an off-premises block server 130. The off-premises block server 130 can be composed of a block service 132 and block storage 134. The block storage 124 of the off-premises block server 130 can store one or more content item blocks 125. A back-end data bus can be composed of a collection of computing devices, networks, and network devices that facilitate network communications and movement of data within and between the control plane and the storage plane, including the servers 130, 140, 150, and 170 thereof. The back-end data bus can also facilitate access to the metadata plane 180 by the control plane and the storage plane, including the servers 130, 140, 150, and 170 thereof. The data plane 180 can be composed of one or more volatile or non-volatile memory-based, possible distributed, database systems for retrieving and storing data (e.g., memcache, a RDBMS, a distributed key-value store, etc.)

[0027] In the example of Figure 1, the system environment 100 includes user 102-1 and user 102-2 having or using personal computing device 110-1 and personal computing device 110-2, respectively. The personal computing device 110-1 has a local storage 112-1 and a content item synchronization agent 114-1. The personal computing device 110-2 also has a local storage 112-2 and a content item synchronization agent 114-2. It should be understood that while Figure 1 depicts only two users (102-1 and 102-2) and two personal computing devices (110-1 and 110-2) in the system environment 100 for purposes of providing a clear example, the system environment 100 may contain more than two users and more than two personal computing devices, each having a local storage and a content item synchronization agent. For example, system environment 100 may have hundreds or thousands or millions of users or more and hundreds or thousands or millions of personal computing devices or more. Further, there is no requirement of a one-to-one correspondence between users and personal computing devices. For example, a single personal computing device may be used by multiple users and a single user may have or use multiple personal computing devices.

[0028] Pursuant to some example embodiments of the present invention, the user 102-1 can use the personal computing device 110-1 to add or modify a content item 113 in the local storage 112-1 of the personal computing device 110-1. The content item synchronization agent 114-1 at the personal computing device 110-1 then automatically detects the addition or modification of the content item 113 to the local storage 112-1 and uploads the content item 113 to on-premises block server 120 or to off-premises block server 130. According to some of the example embodiments, the upload is facilitated by network communications between the content item synchronization agent 114-1 and a metadata server 150 in accordance with a content item synchronization protocol, example embodiments of which are described in greater detail below.

[0029] Pursuant to some example embodiments of the present invention, another personal computing device 110-2 automatically downloads the added or modified content item 113 to the local storage 112-2 of the personal computing device 112-2. Initially, a notification server 140 notifies the content item synchronization agent 114-2 at the personal computing device 112-2 that a new version of a content item 113 is available. Then, as directed by network communications between the synchronization agent 114-2 and the metadata server 150 in accordance with the content item synchronization protocol, the content item synchronization agent 114-2 downloads the new version of the content item 113 from on-premises block server 120 or from off-premises block server 130. After downloading, the new version of the content item 113 is available to the user 102-2 in local storage 112-2.

[0030] A personal computing device (e.g., 110-1 or 110-2) can be a stationary or portable personal computing device. For example, a personal computing device can be a desktop computer, a workstation computer, a mobile telephone, a laptop computer, a tablet computer, or other stationary or portable personal computing device. A personal computing device may be composed of one or more basic hardware components such as, for example, those of basic computing device 800 described below with respect to Figure 8. A personal computing device may also be configured with a basic software system such as, for example, software system 900 described below with respect to Figure 9.

[0031] A local storage (e.g., 112-1 or 112-2) can be an electronic, magnetic, or optical data storage mechanism that is connected to the bus of a personal computing device 110 by a physical host interface (e.g., Serial Attached SCIS, Serial ATA, PCI Express, Fibre Channel, USB, or the like). For example, the data storage mechanism can be a hard disk, a solid state drive, or an optical drive of a personal computing device. Alternatively, a local storage can be a network drive, such as, for example, a network data storage device on a local area network (LAN) that is "mounted" on a personal computing device. By mounting the network drive, data stored in the network drive appears to a user and applications executing on a personal computing device to be stored at the personal computing device (i.e., on a physical drive of the device), even though the

network drive is physically located across the network separate from the device. Once mounted, a personal computing device typically reads and writes data from and to the network drive in accordance with a network file system protocol such as, for example, the network file system (NFS) protocol for UNIX or the server message block (SMB) protocol for WINDOWS.

[0032] A local storage of a personal computing device may store content items 113. A "content item" is a collection of digital information. When stored in a local storage, a content item can correspond to a file in a file system. For example, a content item can be a document file (e.g., a word processing document, a presentation document, a spreadsheet document, or other type of document), an image file (e.g., a.jpg,.tiff,.gif, or other type of image file), an audio file (e.g., a.mp3,.aiff,.m4a,.wav, or other type of audio file), a video file (e.g., a.mov,.mp4,.m4v, or other type of video file), a web page file (e.g., a.htm,.html, or other type of web page file), a text or rich-text file (e.g., a.txt,.rtf, or other type of text or rich-text file), or other type of file. When stored in a local storage, a content item can have a file system path within a file-folder hierarchy of the file system. For example, the file system path for a content item stored in a local storage might be expressed as the character string "C:\folder1\folder2\my.file" where "C:\" refers to a root of the file system, "folder1" refers to a file system folder at the root of the file system, "folder2" refers to a file system folder with the "folder1" file system folder, and "my.file" refers to a file in the "folder2" file system folder corresponding to the content item. The "\" character is used in the character string expression to delineate the different file system folder and file references in the character string expression.

[0033] Personal computing devices 110-1 and 110-2 and on-premises block server 120 can be connected to a local area network (LAN) 160. For example, LAN 160 can be an IEEE 802-based network including, but not limited to, an IEEE 802.3 or IEEE 802.11-based network, or combination of multiple such networks. The LAN 160 may be protected from a wide area network (WAN) 190 by a network firewall. In an embodiment, the WAN 190 is the Internet or other public network. In particular, the network firewall can prohibit devices connected to WAN 190, including servers 130, 140, 150, and 170, from initiating establishment of a network connection with devices connected to LAN 160, including personal computing devices 110-1 and 110-2 and on-premises block server 120. However, the network firewall can be configured to allow certain types of network connections originating from devices connected to LAN 160, including personal computing devices 110-1 and 110-2 and on-premises block server 120, to be established with devices connected to WAN 190, including servers 130, 140, 150, and 170. Typically, LAN 160 has lower network latency and higher network bandwidth when compared to WAN 190 but there is no requirement that this be the case in a given implementation.

[0034] Any and all of on-premises block server 120, off-premises block server 130, notification server 140, metadata server 150, web site 170, and metadata plane 180 may be

implemented by one or more server computing devices, each of which may be composed of one or more basic hardware components such as, for example, those of basic computing device 800 described below with respect to Figure 8, and each of which may also be configured with a basic software system such as, for example, software system 900 described below with respect to Figure 9. If implemented by more than one server computing device, the server computing devices may be configured in a load balanced, clustered, or other distributed computing arrangement.

[0035] The functionality described herein of each of the content item synchronization agent, the block service 122, the block agent 126, the block service 132, the notification server 140, the metadata server 150, the web site 170, and the metadata plane 180 may be implemented as one or more computer programs configured with instructions for performing the functionality when executed by one or more computing devices. However, the functionality can be implemented in hardware (e.g., as one or more application specific integrated circuits (ASICS) or one or more field programmable gate arrays (FPGAs)) or a combination of hardware and software according to the requirements of the particular implement at hand.

[0036] While the example of Figure 1 depicts only a single on-premises block server 120 for the purpose of providing a clear example, the system environment 100 may include tens or hundreds or thousands or millions or more on-premises block servers, depending on the number of related groups of users of the online content management service. For example, the online content management service may support millions of users or more that belong to various different organizations, businesses, corporations, schools, universities, and other groups. Each one of those organizations, business, corporations, schools, universities, and groups may have or use one or more on-premises block servers.

[0037] The term “on-premises” as used herein is intended to be relative to one or more personal computing devices and the online content management service and, in particular, the off-premises block server 130 of the online content management service. While an on-premises block server (e.g., 120) may be located in the same facility or the same building as a personal computing device, there is no requirement that this be the case. Nor is there a requirement that an on-premises block server be connected to the same local area network (e.g., 160) as a personal computing device, although they may be. Accordingly, reference to an “on-premises” block server herein means that the block server is closer in terms of geography and/or the network to a given personal computing device than the off-premises block server 130 is to the given personal computing device. A personal computing device may be closer to an on-premises block server than the off-premises block server 130 on a network if the network connecting the personal computing device to the on-premises block server generally provides lower network latency

and/or higher network bandwidth capability than the network connecting the personal computing device to the off-premises block server 130.

[0038] A personal computing device may make a network request, or just "request", of various servers including, for example, on-premises block server 120, off-premises block server 130, metadata server 150, and web site 170. And servers 120,130, 150, and 170 may return a network response, or just "response", to a request from a personal computing device. The request typically includes a header and a payload. The request header typically provides context for the request payload to the server receiving the request. The response to a request typically also includes a header and a payload. The header of a response typically provides context for the response payload to the personal computing device receiving the response. A request from a personal computing device and a response returned thereto by a server may be sent over one or more networks (e.g., 160 and 190) and made in accordance with a request-response networking protocol such as, for example, the HyperText Transfer Protocol (HTTP). A request and a response thereto may be sent over a network connection established by a personal computing device and a server according to a connection-oriented networking protocol such as, for example, the Transmission Control Protocol (TCP). The network connection may be long-lived in the sense that more than one request and response pair is sent over the network connection. The network connection may also be encrypted according to a cryptographic networking protocol such as, for example, Transport Layer Security (TLS) or Secure Sockets Layer (SSL). However, no particular networking protocol or particular set of networking protocols is required by the example embodiments and protocols other than HTTP, TCP, TLS, or SSL may be used according to the requirements of the particular implementation at hand.

3.0 CONTENT ITEM BLOCK REPLICATION

[0039] According to various example embodiments of the present invention, a content item namespace can be assigned to more than one block server. For example, a content item namespace may be assigned to one or more on-premises block servers and the off-premises block server or two or more on-premises block servers. When a new content item belonging to such a content item namespace is committed to the online content management service, the new content item block(s) of the new content item are uploaded to one or more of the block servers to which the content item namespace is assigned. However, a content item synchronization agent at a personal computing device may select to download missing content item block(s) of the new content item from an assigned block server that is different from one the new content item block(s) were uploaded to. For example, the content item synchronization agent may select an assigned block server that is closer on the network than a block server to which the missing content item block(s) were uploaded to.

[0040] According to some example embodiments of the present invention, to increase the availability of content item blocks at block servers to which a content item namespace is assigned, the block agents (e.g., 126) of on-premises block servers (e.g., 120), in co-operation with the online content management service, implement a content item block replication protocol. According to the content item block replication protocol, when new content item blocks of a new content item are uploaded to an on-premises block server, the on-premises block server can replicate the new content item blocks to other block servers that are assigned to the content item namespace to which the new content item belongs. Also according to the content item block replication protocol, when new content item blocks of a new content item are uploaded to the off-premises block server, the off-premises block server can replicate the new content item blocks to any on-premises block servers that are also assigned to the content item namespace to which the new content item belongs. In this way, new content item blocks of a new content item that are uploaded to a block server are made available at all block servers assigned to the content item namespace to which the new content item belongs.

[0041] For example, referring now to Figure 2, assume a certain content item namespace 'ABC123' is assigned to on-premises block server 120-1, on premises block server 120-2, and off-premises block server 130. On-premises block server 120-1 and on-premises block server 120-2 may be connected to the same local area network or different local area networks. In the example of Figure 2, on premises block server 120-1 and on-premises block server 120-2 are connected to different local area networks, in particular, LAN 160-1 and LAN 160-2, respectively. On-premises block servers 120-1 and 120-2 and associated LANS 160-1 and 160-2 may be geographically distributed and do not necessarily reside in the same geographic area. For example, on-premises block server 120-1 and LAN 160-1 can be located in a company's San Francisco headquarters while on-premises block server 120-2 and LAN 160-2 can be located in the company's New York offices. Alternatively, on-premises block servers 120-1 and 120-2 and LANS 160-1 and 160-2 may be owned and operated by different businesses or organizations. For example, the content item namespace 'ABC123' may represent a shared folder that Company Alpha and Company Beta use for collaboration. Whether on-premises block servers 120-1 and 120-2 and LANS 160-1 and 160-2 are owned and operated by the same or different companies, with the content item block replication protocol, new content item blocks for a new content item in content item namespace 'ABC123' uploaded to on-premises block server 120-1 by personal computing device 110-1 are automatically made available for download by personal computing device 110-2 from on-premises block server 120-2 and off-premises block server 130. Similarly, new content item blocks for a new content item in content item namespace 'ABC123' uploaded to on-premises block server 120-2 by personal computing 110-2 are automatically made available for download by personal computing device 110-1 from on-premises block server 120-

1 and off-premises block server 130. Additionally, new content item blocks for a new content item in content item namespace 'ABC123' uploaded to the off-premises block server 130 by either personal computing device 110-1 or personal computing device 110-2 are automatically made available for download from on-premises block server 120-1 and on-premises block server 120-2.

[0042] At a high-level, operation of the content item block replication protocol proceeds as follows. The block agent at each on-premises block server maintains a current client cursor value for each content item namespace assigned to the on-premises block server. The current client cursor value for a content namespace represents which server journal entries for the content item namespace the on-premises block server already knows about. The current client cursor value for a content item namespace is used by the block agent at an on-premises block server to determine which content item blocks of content items in the content item namespace stored at the on-premises block server should be offered to other block servers and which content item blocks of content items in the content item namespace stored at the off-premises block server should be downloaded to the on-premises block server.

[0043] The block agent at each on-premises block server may maintain a long-polling connection to the notification server (e.g., 140) of the online content management service. When a new content item in a content item namespace assigned to an on-premises block server is committed to the online content management service, a ping message may be sent to the on-premises block server over the long-polling connection, if the new content item blocks of the new content item were uploaded to the on-premises block server or the off-premises block server. For example, on-premises block server 120-1 may receive a ping message from the notification server 140 if new content item blocks of a new content item belonging to content item namespace 'ABC123' were uploaded to on-premises block server 120-1 or the off-premises block server 130.

[0044] In response to receiving a ping message from the notification server, an on-premises block server may make a "block server list" request of the metadata server (e.g., 150) of the online content management service. The block server list request may specify the current client cursor value for each content item namespace assigned to the on-premises block server. The block server list request may include other information such as a block server identifier of the on-premises block server sending the block server list request.

[0045] In response to receiving a block server list request from an on-premises block server, the metadata server may determine one or more newer server journal entries to send to the on-premises block server in a response to the block server list request. Each newer server journal entry corresponds to one of the content item namespaces assigned to the on-premises block server and specified in the block server list request. Each newer server journal entry for a content

item namespace has a server journal cursor value that is greater than the current client cursor value specified by the on-premises block server for the content item namespace in the block server list request. Each newer server journal entry for a content item namespace corresponds to either a) content item blocks uploaded to the on-premises block server that the on-premises block server can offer to other block servers assigned to the content item namespace or b) content item blocks uploaded to the off-premises block server that the on-premises block server can download from the off-premises block server. The former type a) of newer server journal entry is referred to hereinafter as an "offer" newer server journal entry. The later type b) of newer server journal entry is referred to hereinafter as a "download" newer server journal entry. Processing of offer newer server journal entries and download newer server journal entries by an on-premises block server is described in greater detail below.

3.1 REPLICATION METADATA

[0046] Turning now to Figure 3, it is a block diagram of a content item block replication metadata 300 that may be maintained by a block agent at an on-premises block server. The metadata 300 may contain one or more assigned content item namespace entries 302-1, 302-2, …, 302-N. Each entry 302 represents a content item namespace assigned to the on-premises block server.

[0047] An entry 302 may include an identifier 314 of the content item namespace assigned to the on-premises block server. The entry 302 may also include a current client cursor value 317 for the content item namespace assigned to the on-premises block server. The current client cursor value 317 represents the changes to content items in the content item namespace committed to the online content management service that the on-premises block server already knows about for content item block replication purposes. The entry 302 may also include one or more assigned block server identifiers 313-1, 313-2, … , 313-N. An assigned block server identifier 313 identifies a block server assigned to the content item namespace.

[0048] The block agent at an on-premises block server may use maintained metadata 300 as part of the content item block replication protocol as described in greater detail below.

3.2 SERVER JOURNAL

[0049] Turning now to Figure 4, it is a block diagram of a content item server journal 410 that may be maintained in the metadata plane of the online content management service by the metadata server. The server journal 410 contains one or more server journal entries 412-1, 412-2, … , 412-N. Each server journal entry 412 represents a new content item committed to the online content management service. For example, a server journal entry 412 may be added to the content item server journal 410 by the metadata server 150 in response to receiving a second commit request as described above with respect to upload processes 200 and 1000.

[0050] A server journal entry 412 may contain an identifier 414 of a content item namespace to which the new content item represented by the server journal entry 412 belongs. The server journal entry 412 may also contain an identifier 413 of the target block server the new content item blocks of the new content item were uploaded to. The target block server identifier 413 may be specified in the second commit request from the content item synchronization agent as part of step 226 of upload process 200 or step 1028 of upload process 1000, for example. The server journal entry 412 may also contain a relative path 415 for the new content item. The server journal entry 412 may also contain a content item block list for the new content item identifying the content item block(s) that make up the new content item. The server journal entry 412 may also contain a server journal cursor value 417. The server journal cursor value 417 can be specific to the content item namespace identified 414 in the entry 412. Alternatively, the server journal cursor value 417 can be specific to the combination of the content item namespace and the target block server identified (2214 and 413, respectively) in the entry 412.

3.3 REPLICATION LOG

[0051] According to some example embodiments of the present invention, an on-premises block server maintains a content item block replication log in local storage (e.g., 124-1) at the on-premises block server. The replication log stores one or more replication log entries. Each log entry represents a replication task for the block server of the on-premises block server. A replication task can involve either a) offering to send one or more content item blocks to one or more other block servers and sending one or more content item blocks to the block servers that accept the offer, or b) downloading one or more content item blocks from the off-premises block server.

[0052] Turning now to Figure 5, it is a block diagram of a content item block replication log 500 that may be stored locally at an on-premises block server. The replication log 500 contains one or more replication log entries 502. Each replication log entry 502 corresponds to either an offer newer server journal entry or a download newer server journal entry determined by the metadata server 150. The block agent may add a replication log entry 502 to the replication log 500 for a newer server journal entry and each download newer server journal entry received from the metadata server 150.

[0053] A replication log entry 502 may contain a replication log entry type identifier 518, a content item namespace identifier 514, a content item block list 516, and one or more block server task entries 519. The replication log entry type identifier 518 indicates whether the corresponding newer server journal entry 412 is an offer-type newer server journal entry or a download-type newer server journal entry. The content item namespace identifier 514 corresponds to the content item namespace identifier 414 of the corresponding newer server journal entry 412. The content item block list 516 corresponds to the content item block list 416

of the corresponding newer server journal entry 412. Each block server task entry 519 represents a replication task to be performed by the block agent at the on-premises block server with another block server that is assigned to the content item namespace.

[0054] A block server task entry 519 may identify 513 another block server assigned to the content item namespace and contains task metadata 520 related to performance of the replication task represented by the task entry 519. The task metadata 520 indicates whether the replication task has been completed or not. The task metadata 520 may include other information such as the number of unsuccessful attempts to complete the replication task that have already been made, error codes and error messages related to unsuccessful attempts, and log messages reflecting replication task execution.

[0055] While in some example embodiments only on-premises block servers maintain a replication log, the off-premises block server 130 maintains a replication log in other embodiments in addition to or instead of an on-premises block server maintaining a replication log. In this case, a log entry in the replication log maintained by the off-premises block server can represent either a) a replication task for the off-premises block server 130 of offering to send one or more content item blocks to one or more off-premises block servers and sending one or more content item blocks to the off-premises block servers that accept the offer, or b) downloading one or more content item blocks from an off-premises block server.

3.4 PROVIDING REPLICATION TASKS

[0056] Figure 6 is a flow diagram of a process 600 for providing replication tasks to an on-premises block server (e.g., 120-1 or 120-2). The steps of the process 600 are as follows. At step 602, the block agent (e.g., 126-1 or 126-2) of the on-premises block server determines the current client cursor value (e.g., 317) for each content item namespace assigned to the on-premises block server. As previously stated, this information may be stored as part of content item block replication metadata (e.g., 300) stored at the on-premises block server. At step 604, the block agent of the on-premises block server sends a block server list request to the metadata server (e.g., 150) of the online content management service. The block server list request contains the current client cursor values determined at step 602 in association with the identifiers (e.g., 314) of the content item namespaces to which they pertain. The block server list request may also contain a user account identifier (e.g., 312) identifying a user account that has been successfully authenticated against and a block server identifier identifying the on-premises block server making the block server list request.

[0057] At step 606, the metadata server of the online content management service receives the block server list request and authenticates it. This may include accessing data in the metadata plane (e.g., 180) to verify that the user account identified in the block server list request is authorized to make block server list requests for the on-premises block server identified in the

block server list request. If not, the metadata server may deny the block server list request thereby ending the process 600. Authenticating the request may also include verifying that the content item namespace(s) identified in the block server list request are ones assigned to the block server identified in the request. For the remainder of the process 600, the metadata server may ignore any content item namespaces identified in the block server list request that are not currently assigned to the on-premises block server.

[0058] At step 608, the metadata server accesses a sever journal (e.g., 410) in the metadata plane to determine any newer server journal entries for each content item namespace identified in the block server list request. This determination involves scanning the server journal starting at the newest server journal entry and scanning back through the server journal until all "qualifying" newer server journal entries have been collected for each content item namespace. According to some example embodiments, a qualifying newer server journal entry is one that has all of the following properties, or a subset or a superset thereof:

[0059] The content item namespace identified (e.g., 414) in the server journal entry (e.g., 412-2) is one of the content item namespaces identified in the block server list request;

[0060] More than one block server is assigned to the content item namespace;

[0061] The block server identified (e.g., 413) in the server journal entry identifies either the block server making the block server list request or the off-premises block server (e.g., 130) of the online content management server; and

[0062] The server journal cursor value (e.g., 417) of the server journal entry is greater than the current client cursor value for the content item namespace specified in the block server list request.

[0063] At step 610, the metadata server returns a response to the block server list request to the block agent of the on-premises block server. The response may include information from each qualifying newer server journal entry identified at step 608. In particular, the information returned for each qualifying newer server journal entry may include all of the following information, or a subset or a superset thereof:

[0064] The content item namespace identifier of the qualifying newer server journal entry;

[0065] A replication task type indicator for the entry that varies depending on the target block server identifier of the qualifying newer server journal entry. In particular, if the target block server identifier of the qualifying entry identifies the block server that made the block server list request, then the replication task type indicator indicates that the replication task for the entry is the offer-type replication task. On the other hand, if the target block server identifier of the qualifying entry identifies the off-premises block server, then the replication task type indicator indicates that the replication task for the entry is the download-type replication task;

[0066] The content item block list (e.g., 416) of the qualifying entry; or

[0067] The server journal cursor value of the qualifying entry.

[0068] At step 612, the block agent of the on-premises block server receives the response to the block server list request from the metadata server and stores one or more replication log entries (e.g., 502-2) in a replication log (e.g., 500) at the on-premises block server. In particular, information for each qualifying newer server journal entry returned in the response is used to store a corresponding replication log entry in the replication log. For each content item namespace, the information for the qualifying newer server journal entries may be processed in increasing order of their server journal cursor values. For each qualifying newer server journal entry, a log entry type (e.g., 518) based on the replication task type entry for the qualifying newer server journal entry, a content item namespace identifier (e.g., 514) based on the content item namespace identifier of the qualifying entry, a content item block list (e.g., 516) that is the content item block list of the qualifying entry, and one or more block server task entries (e.g., 519-2).

[0069] A block server task entry may be created for each other block server assigned to the content item namespace identified in the log entry. For download replication task-type log entries, there may be just one block server task entry for the off-premises block server. For offer replication task-type log entries, a block server task entry may be created for each other block server assigned to the content item namespace.

[0070] The task metadata (e.g., 520) of the block server task entry is initially set to indicate that the replication task is not yet complete. When a replication log entry is added to the replication log, the current client cursor value (e.g., 317) for the content item namespace of the log entry in the corresponding assigned content item namespace entry (e.g., 302-2) at the on-premises block server is set to equal the server journal cursor value of the corresponding qualifying entry. By doing so, the block agent will not receive information for the qualifying entry again in response to the next block server list request made by the block agent.

3.5 PROCESSING REPLICATION TASKS

[0071] Figures 7A-C comprise a flow diagram 700 illustrating operation of the block agent (e.g., 126-1) at an on-premises block server (e.g., 120-1) in processing replication tasks specified in a replication log (e.g., 530) stored (e.g., in storage 124-1) at the on-premises block server. At step 702, the block agent obtains a replication log entry (e.g., 502-2) from the replication log. For example, the block agent may periodically scan the replication log for replication log entries that are pending.

[0072] A replication log entry may be pending if at least one of the block server task entries (e.g., 519-2) of the replication log entry is pending. A block server task entry may be pending if indicated so by its task metadata (e.g., 520). The task metadata of a block server task entry may indicate that the block server task entry is pending in a number of different ways. For example,

the task metadata may contain a value or set of values that indicate that the block server task entry is pending or not complete. According to some example embodiments, the task metadata contains a value reflecting a number of attempts that the block agent has already made to complete the block server task entry. If the number is below or at a threshold, then the block server task entry is pending. If the number is above the threshold, then the block server task entry is not pending. By attempting to complete a block server task entry multiple times in the event of prior failures, greater resiliency and fault tolerance is provided.

[0073] At step 704, the block agent determines the replication task type (e.g., 518) of the pending log entry. The replication task type can be one of "download" or "offer". A download replication task type is performed by the block agent to download content item blocks uploaded to the off-premises block server (e.g., 130) that are not stored at the on-premises block server (i.e., are missing at the on-premises block server). An offer replication task type is performed by the block agent to offer to send content item blocks uploaded to the "offeror" on-premises block server to one or more other "offeree" block servers and send them to the other block servers that accept the offer.

[0074] An on-premises block server can only perform an offer replication task with another on-premises block server that it has a peering relationship with. A peering relationship between two on-premises block servers may exist if it is possible to establish a network connection between the block agents of the two on-premises block servers. It may not be possible to establish a network connection between two on-premises block servers because of a network firewall interposed on the network between the two on-premises block servers, or simply because there is no network that connects the two on-premises block servers. For example, referring briefly to Figure 2, network firewall 165-1 or firewall 165-2 may prevent on-premises block server 120-1 and on-premises block server 120-2 from establishing a network connection between them over WAN 160-3 or Internet 190. Alternatively, network firewalls 165-1 and 165-2 may allow the block agents 126-1 and 126-2 to establish a network connection between them over WAN 160-3 but not Internet 190. Other network firewall configurations are possible. For example, network firewalls 165-1 and 165-2 may allow the block agents 126-1 and 126-2 to establish a network connection between them over WAN 160-3 or Internet 190. It should be noted that it is not necessary for a peering relationship to exist between a pair of on-premises block servers that both on-premises block servers of the pair be able to initiate establishment of a network connection between the on-premises block servers. For example, network firewalls 165-1 and 165-2 may allow block agent 126-1 at on-premises block server 102-1 to initiate establishment of a HTTPS connection over WAN 160-3 with block agent 126-2 at on-premises block server 102-1 but not vice versa.

[0075] According to some example embodiments, if an on-premises block server A does not have a peering relationship with on-premises block server B, then on-premises block server A may not store a block server task entry for on-premises block server B in its replication log when storing an offer-type replication log entry in the replication log. This is because, in the absence of a peering relationship, on-premises block server A cannot offer any content item blocks to on-premises block server B. The offer-type replication log entry may be omitted altogether from the replication log if all of the block server task entries of the replication log entry are for on-premises block servers that on-premises block server A does not have peering relationships with. This is because, in the absence of peering relationship, on-premises block server A cannot offer any of its content item blocks to any other on-premises block servers it does not have a peering relationship with. Alternatively, instead of omitting a block server task entry or a replication log entry in the absence of a peering relationship, a block server task entry for an on-premises block server can be stored as part of a replication log entry with task metadata that indicates that there is no peering relationship with the on-premises block server identified in the block server task entry.

[0076] If, at step 704, the block agent determines that the pending replication log entry is a download type replication log entry, then the process 700 proceeds to step 706 (Figure 7B). A download type replication log entry may contain a single block server task entry that identifies the off-premises block server of the online content management service. At step 706, the block agent determines any missing content item blocks identified in the download-type replication log entry. This determination may be based on the content item block list (e.g., 516) of the entry. At step 708, the block agent downloads any missing content item blocks from the off-premises block server. At step 710, depending on whether the download of missing content item blocks is successful, the block agent updates the task metadata of the block server task entry. For example, if not all of the content item blocks could be downloaded, then the block agent may update the task metadata to indicate so. For example, the block agent may increment an attempt counter that tracks the number of attempts that have been made to successfully download all of the missing content item blocks. On the other hand, if the block agent was successful at downloading all of the missing content item blocks, then task metadata may be updated to indicate that the block server task entry is no longer pending. After step 710, the process 700 may return to step 702 to process the next pending replication log entry.

[0077] On the other hand, if, at step 704, the block agent determines that the pending replication log entry is an offer type replication log entry, then the process 700 proceeds to step 712 (Figure 7C). At step 712, the block agent determines which of the content item blocks identified in the pending replication log entry that the on-premises block server is in possession of (i.e., are stored at the on-premises block server). This determination may be based on the

content item block list of the log entry. All content item blocks identified in the content item block list of the log entry should be stored at the on-premises block serer, unless some or all of the content item blocks have been deleted or removed from the on-premises block server. For example, content item blocks may be deleted or removed from the on-premises block server according to a least recently used scheme (e.g., least recently downloaded or least recently uploaded).

[0078] At step 714, the block agent sends an offer request to each other block server identified in a pending block server task entry of the pending replication log entry that the offeror on-premises block server has a peering relationship with. It may be assumed in some implementations that the offeror on-premises block server has a peering relationship with the off-premises block server. The offer request may identify the content item block(s) that are offered. In particular, the offer request may include the content item block hashes of the content item blocks identified in the pending replication log entry that the offeror on-premises block server is in possession of. The offer request may be sent over a network connection (e.g., a HTTPS connection) established between the block agent at the offeror on-premises block server and the block agent at an offeree block server.

[0079] At step 716, the block agent at the offeror on-premises block server receives any acceptance responses sent by the offeree block server(s) in response to receiving an offer request from the offeror on-premises block server. Each acceptance response from an offeree block server may identify one or more of the offered content item blocks that are currently not stored at the offeree block server (i.e., are missing at the offeree block server). Any missing content item block(s) can be identified in the acceptance response by the content item block hash(es) of the missing content item block(s). An acceptance response from an offeree block server may also indicate that none of the offered content item blocks are missing at the offeree block server.

[0080] At step 718, the block agent at the offeror on-premises block server sends (uploads) any missing content item block(s) at the offeree block server(s). In particular, for a given acceptance response from an offeree block server, the block agent at the offeror on-premises block server sends (uploads) any missing content item block(s) identified in the acceptance response from the offeree block server.

[0081] At step 720, the block agent at the offeror on-premises block server updates the task metadata of the pending block server task entries of the pending log entry. In particular, if an acceptance response was received from an offeree block server corresponding to a block server task entry, then the task metadata of the block server task entry is updated depending on whether all missing content item block(s) were successfully sent to the offeree block server or whether the acceptance response indicated the no content item blocks are missing at the offeree block server. In either case, the task metadata may be updated to indicate that the block server task

entry is no longer pending. On the other hand, if an acceptance response was not received or there was a failure in sending (uploading) a missing content item block to the offeree block server, then an attempt counter of the task metadata may be incremented, in which case the block server task entry may remain pending. After step 720, the process 700 may return to step 702 to process the next pending replication log entry.

4.0 DELETING CONTENT ITEM BLOCKS

[0082] Typically, it is expected, but not required, that an on-premises block server (e.g., 120) will have significantly less local data storage space in its local storage (e.g., 124) than the off-premises block server (e.g., 130) has in its local storage (e.g., 134). For example, the total local storage at an on-premises block server may be on the order of one to a few terabytes while the total local storage at the off-premises block server may on the order of eight (8) zettabytes. Thus, the off-premises block server may have up to a billion times more storage space than a given on-premises block server. Even though a given on-premises block server may have many fewer content item namespaces assigned to it than the off-premises block server, the on-premises block server may still not have enough local storage space to store all content item blocks of all of the content items in all of the content item namespaces assigned to the on-premises block server.

[0083] According to some example embodiments, content item blocks locally stored at an on-premises block server are deleted or removed from the local storage to make local storage space at the on-premises block server available for other content item blocks. For example, the other content item blocks might be content item blocks that are being uploaded or are about to be uploaded or that will be uploaded to the on-premises block server.

[0084] Various different approaches may be employed to determine which content item blocks to delete or remove. According to some example embodiments, a least recently used (LRU) approach is employed. According to the LRU approach, if the amount of local storage space at an on-premises block server consumed by content item blocks does not satisfy a threshold, then one or more least recently used content item blocks are deleted or removed from local storage. The threshold can be based on a percentage of the total local storage space for content item blocks at the on-premises block server. Here, total local storage space refers to the current maximum total amount of storage space available for storing content item blocks irrespective of whether some or all of that storage space is currently used for storing content item blocks. If the current consumption amount is greater than the percentage, then the current consumption amount does not satisfy the threshold. The threshold can instead be based on the current maximum local storage space amount that remains after subtracting the amount of storage space currently used by content item blocks stored in the local storage. In this case, if the local storage space remaining after accounting for the current consumption amount is less than a threshold amount, then the current consumption amount does not satisfy the threshold.

[0085] A determination of whether the current consumption amount does or does not satisfy the threshold can be made at various different times. One possible time is when one or more content item blocks are uploaded to the on-premises block server. In particular, if the current consumption amount with the uploaded content item block(s) stored in local storage does not satisfy the threshold, then one or more content item blocks may be deleted or removed from the local storage according to the LRU approach.

[0086] A content item block stored in local storage may be considered to be least recently used based on its most recent upload time and/or its most recent download time. The most recent upload time for a content item block reflects a time at which the content item block was most recently uploaded to the on-premises block server. For example, the content item block stored in local storage that, according to the most recent upload time for the content item block, was least recently uploaded to the on-premises block server may be deleted or removed. The most recent download time for a content item block reflects a time at which the content item block was most recently downloaded from the on-premises block server. For example, the content item block stored in local storage that, according to the most recent download time for the content item block, was least recently download from the on-premises block server may be deleted or removed. The on-premises block server may store and maintain metadata that reflects the most recent upload and download times for content item blocks stored at the on-premises block server.

[0087] According to some example embodiments, a qualified LRU approach is used. According the qualified LRU approach, a content item block that qualifies for deletion or removal according to the LRU approach discussed above, is not deleted or removed unless additional conditions are met.

[0088] One possible additional condition is that the content item block that is a candidate for deletion or removal be stored at one or more other block servers, either the off-premises block server, one or more other on-premises block servers, or one or more other on-premises block server and the off-premises block server. If the candidate content item block is stored only at the on-premises block server at which it is a candidate for deletion or removal, then the content item block may not be deleted or removed in order to preserve the potentially only copy of the content item block.

[0089] Another possible condition is that the content item block that is a candidate for deletion or removal belongs to a content item that belongs to a content item namespace that is assigned to at least one other block server in addition to the on-premises block server at which the content item block is a candidate for deletion. If the content item namespace is assigned only to the on-premises block server at which the content item block is a candidate for deletion or removal, then the content item block may not be deleted or removed in order to preserve the potentially only copy of the content item block.

[0090] Another possible condition is that the content item block is specially marked as a "sticky" content item block. A sticky content item block is a content item block that is not deleted or removed until all non-sticky content item blocks have been deleted or removed. Designating content item blocks as sticky can help prevent deletion or removal of a content item block that is important or relatively more likely to be downloaded in the future. A content item block may be designated as sticky based on a content item namespace with which it is associated. In particular, a content item namespace may be designated as sticky by a user of the online content management service through a graphical user interface provided by the online content management service (e.g., via web site 170). Alternatively, a content item namespace may be automatically designated as sticky based on characteristics and usage of the content item namespace. For example, if a content item namespace is shared among a large number of users (e.g., ten or more), then the content item namespace may be automatically designated as sticky. This is useful to prevent the deletion or removal of content item blocks that are shared among a large number of users. In some example embodiments, a shared content item namespace is designated sticky only if there has been recent user activity in the content item namespace. This is useful to prevent retaining content item blocks that have not recently been used. Recent activity may include recently (e.g., within the past day, week, month, or year) downloading, uploading, or accessing a content item belonging to the content item namespace. A content item namespace designated as sticky as a first time may automatically no longer be designated as sticky at a later second time if the condition for designating the content item namespace as sticky no longer exists. For example, if a content item namespace designated as sticky based on recent activity at a first time no longer has recent activity at a later second time, then, at or after the second time, the content item namespace may no longer be designated as sticky. An on-premises block server may store and maintain metadata that indicates which locally stored content item blocks and/or assigned content item namespaces are designated as sticky.

[0091] According to some embodiments, there are multiple levels of stickiness and content item blocks are deleted or removed according to the qualified LRU approach in order of their level of stickiness. For example, there may be three levels of stickiness A, B, and C where level C is lower than level B and level B is lower than level A. In this case, the content item blocks that are not designated at any level of stickiness are deleted or removed before the first content item block at level C is deleted or removed. Content item blocks at level C are deleted or removed before the first content item block at level B is deleted or removed. Content item blocks at level B are deleted or removed before the first content item block at level A.

5.0 CONTENT ITEM BLOCK REPLICATION WHEN ASSIGNMENT OCCURS

[0092] A content item namespace can be assigned to a block server to which it is not currently assigned. For example, a content item namespace may be assigned to just the off-

premises block server 130. Then, at a later time, the content item namespace may be assigned to the off-premises block server 130 and on-premises block server 120-1. Then, at a later time still, the content item namespace may be assigned to the off-premises block server 130, on-premises block server 120-1, and on-premises block server 120-2. As another example, a content item namespace may be assigned to on-premises block server 120-1 and then, later, assigned to the off-premises block server 120-1 and the off-premises block sever 130. Thus, the set of block servers to which a content item namespace is assigned at one time may be different than the set of block servers to which the content item namespace is assignment at another time.

[0093] When a content item namespace is assigned to a new block server, the block server may store only some or none of the content item blocks that make up the content items that belong to the content item namespace. As a result, a content item synchronization agent at a personal computing device may not be able to download from the block server a content item block of a content item in the content item namespace.

[0094] According to some example embodiments of the present invention, when a content item namespace is assigned to a new block server in a situation where the content item namespace is currently assigned to one or more other block servers, content item blocks of content items in the content item namespace are replicated to the new block server from the other block server(s). For example, if content item namespace 'ABC123' is assigned to off-premises block server 130 and then later is assigned to on-premises block server 120-1, then on-premises block server 120-1 can download from on-premises block server 130 content item blocks that make up content items in the 'ABC123' content item namespace. As another example, if content item namespace 'DEF456' is assigned to on-premises block server 120-1 and then later assigned to on-premises block server 120-2, then on-premises block server 120-1 can offer and send to on-premises block server 120-2 content item blocks that make up content items in the 'DEF456' content item namespace.

[0095] According to some example embodiments, when a content item namespace that is currently assigned to one or more block servers is assigned to a new block server, content item blocks are replicated from the currently assigned block server(s) to the new block server in accordance with the content item block replication protocol described above. In particular, the block server identifier of the new block server is added as an assigned block server identifier 313 to the assigned content item namespace entry 302 for the content item namespace in the content item block replication metadata 300 maintained at each of the currently assigned on-premises block server(s) and the new block server. For example, the block agent at each of the currently assigned on-premises block server(s) and the new block server may add this information to the locally stored content item block replication metadata 300 based on information it receives from the metadata server 150. For example, the block agent at a currently assigned on-premises block

server and the new block server may receive this information from the metadata server 150 in the response to a block server list request. For example, after the content item namespace is assigned to the new block server, the notification server 140 may send a ping message to the block agent at each of the currently assigned on-premises block servers and the new block server. Responsive to receiving the ping message, the block agent at a currently assigned on-premises block server and the new block server may send a block server list request to the metadata server 150 and receive from the metadata server 150 in the response to the block server list request that the content item namespace is now assigned to the new block server.

[0096] The block agent at each of the currently assigned on-premises block server(s) may then scan its replication log 500 for offer-type log entries 502 with a content item namespace identifier 514 that matches the identifier of the content item namespace newly assigned to the new block server and with a log entry type 518 that indicates that the log entry is an offer-type log entry as opposed to a download type log-entry. A new block server task entry 519 is added to each such offer-type log entry. The new block server task entry has a block server identifier 513 identifying the new block server and task metadata 520 indicating that the replication task is not yet complete. The block agent at each of the currently assigned on-premises block server(s) can then offer the content item blocks in the content item namespace that it has in its possession to the new block server in accordance with the content item block replication process 700 described above.

[0097] Also, the block agent at the new block server may perform the content item block replication processes 600 and 700 described above to download from the off-premises block server any content item blocks in the content item namespace assigned to the new block server.

6.0 BASIC COMPUTING HARDWARE AND SOFTWARE

6.1 BASIC COMPUTING DEVICE

[0098] Referring now to Figure 8, it is a block diagram that illustrates a basic computing device 800 in which the example embodiment(s) of the present invention can be embodied. Computing device 800 and its components, including their connections, relationships, and functions, is meant to be exemplary only, and not meant to limit implementations of the example embodiment(s). Other computing devices suitable for implementing the example embodiment(s) can have different components, including components with different connections, relationships, and functions.

[0099] Computing device 800 can include a bus 802 or other communication mechanism for addressing main memory 806 and for transferring data between and among the various components of device 800.

[0100] Computing device 800 can also include one or more hardware processors 804 coupled with bus 802 for processing information. A hardware processor 804 can be a general purpose microprocessor, a system on a chip (SoC), or other processor.

[0101] Main memory 806, such as a random access memory (RAM) or other dynamic storage device, also can be coupled to bus 802 for storing information and software instructions to be executed by processor(s) 804. Main memory 806 also can be used for storing temporary variables or other intermediate information during execution of software instructions to be executed by processor(s) 804.

[0102] Software instructions, when stored in storage media accessible to processor(s) 804, render computing device 800 into a special-purpose computing device that is customized to perform the operations specified in the software instructions. The terms "software", "software instructions", "computer program", "computer-executable instructions", and "processor-executable instructions" are to be broadly construed to cover any machine-readable information, whether or not human-readable, for instructing a computing device to perform specific operations, and including, but not limited to, application software, desktop applications, scripts, binaries, operating systems, device drivers, boot loaders, shells, utilities, system software, JAVASCRIPT, web pages, web applications, plugins, embedded software, microcode, compilers, debuggers, interpreters, virtual machines, linkers, and text editors.

[0103] Computing device 800 also can include read only memory (ROM) 808 or other static storage device coupled to bus 802 for storing static information and software instructions for processor(s) 804.

[0104] One or more mass storage devices 810 can be coupled to bus 802 for persistently storing information and software instructions on fixed or removable media, such as magnetic, optical, solid-state, magnetic-optical, flash memory, or any other available mass storage technology. The mass storage can be shared on a network, or it can be dedicated mass storage. Typically, at least one of the mass storage devices 810 (e.g., the main hard disk for the device) stores a body of program and data for directing operation of the computing device, including an operating system, user application programs, driver and other support files, as well as other data files of all sorts.

[0105] Computing device 800 can be coupled via bus 802 to display 812, such as a liquid crystal display (LCD) or other electronic visual display, for displaying information to a computer user. In some configurations, a touch sensitive surface incorporating touch detection technology (e.g., resistive, capacitive, etc.) can be overlaid on display 812 to form a touch sensitive display for communicating touch gesture (e.g., finger or stylus) input to processor(s) 804.

[0106] An input device 814, including alphanumeric and other keys, can be coupled to bus 802 for communicating information and command selections to processor 804. In addition to or

instead of alphanumeric and other keys, input device 814 can include one or more physical buttons or switches such as, for example, a power (on/off) button, a "home" button, volume control buttons, or the like.

[0107] Another type of user input device can be a cursor control 816, such as a mouse, a trackball, or cursor direction keys for communicating direction information and command selections to processor 804 and for controlling cursor movement on display 812. This input device typically has two degrees of freedom in two axes, a first axis (e.g., x) and a second axis (e.g., y), that allows the device to specify positions in a plane.

[0108] While in some configurations, such as the configuration depicted in Figure 8, one or more of display 812, input device 814, and cursor control 816 are external components (i.e., peripheral devices) of computing device 800, some or all of display 812, input device 814, and cursor control 816 are integrated as part of the form factor of computing device 800 in other configurations.

[0109] Functions of the disclosed systems, methods, and modules can be performed by computing device 800 in response to processor(s) 804 executing one or more programs of software instructions contained in main memory 806. Such software instructions can be read into main memory 806 from another storage medium, such as storage device(s) 810. Execution of the software instructions contained in main memory 806 cause processor(s) 804 to perform the functions of the example embodiment(s).

[0110] While functions and operations of the example embodiment(s) can be implemented entirely with software instructions, hard-wired or programmable circuitry of computing device 800 (e.g., an ASIC, a FPGA, or the like) can be used in other embodiments in place of or in combination with software instructions to perform the functions, according to the requirements of the particular implementation at hand.

[0111] The term "storage media" as used herein refers to any non-transitory media that store data and/or software instructions that cause a computing device to operate in a specific fashion. Such storage media can comprise non-volatile media and/or volatile media. Non-volatile media includes, for example, non-volatile random access memory (NVRAM), flash memory, optical disks, magnetic disks, or solid-state drives, such as storage device 810. Volatile media includes dynamic memory, such as main memory 806. Common forms of storage media include, for example, a floppy disk, a flexible disk, hard disk, solid-state drive, magnetic tape, or any other magnetic data storage medium, a CD-ROM, any other optical data storage medium, any physical medium with patterns of holes, a RAM, a PROM, and EPROM, a FLASH-EPROM, NVRAM, flash memory, any other memory chip or cartridge.

[0112] Storage media is distinct from but can be used in conjunction with transmission media. Transmission media participates in transferring information between storage media. For

example, transmission media includes coaxial cables, copper wire and fiber optics, including the wires that comprise bus 802. Transmission media can also take the form of acoustic or light waves, such as those generated during radio-wave and infra-red data communications.

[0113] Various forms of media can be involved in carrying one or more sequences of one or more software instructions to processor(s) 804 for execution. For example, the software instructions can initially be carried on a magnetic disk or solid-state drive of a remote computer. The remote computer can load the software instructions into its dynamic memory and send the software instructions over a telephone line using a modem. A modem local to computing device 800 can receive the data on the telephone line and use an infra-red transmitter to convert the data to an infra-red signal. An infra-red detector can receive the data carried in the infra-red signal and appropriate circuitry can place the data on bus 802. Bus 802 carries the data to main memory 806, from which processor(s) 804 retrieves and executes the software instructions. The software instructions received by main memory 806 can optionally be stored on storage device(s) 810 either before or after execution by processor(s) 804.

[0114] Computing device 800 also can include one or more communication interface(s) 818 coupled to bus 802. A communication interface 818 provides a two-way data communication coupling to a wired or wireless network link 820 that is connected to a local network 822 (e.g., Ethernet network, Wireless Local Area Network, cellular phone network, Bluetooth wireless network, or the like). Communication interface 818 sends and receives electrical, electromagnetic, or optical signals that carry digital data streams representing various types of information. For example, communication interface 818 can be a wired network interface card, a wireless network interface card with an integrated radio antenna, or a modem (e.g., ISDN, DSL, or cable modem).

[0115] Network link(s) 820 typically provide data communication through one or more networks to other data devices. For example, a network link 820 can provide a connection through a local network 822 to a host computer 824 or to data equipment operated by an Internet Service Provider (ISP) 826. ISP 826 in turn provides data communication services through the world wide packet data communication network now commonly referred to as the “Internet” 828. Local network(s) 822 and Internet 828 use electrical, electromagnetic or optical signals that carry digital data streams. The signals through the various networks and the signals on network link(s) 820 and through communication interface(s) 818, which carry the digital data to and from computing device 800, are example forms of transmission media.

[0116] Computing device 800 can send messages and receive data, including program code, through the network(s), network link(s) 820 and communication interface(s) 818. In the Internet example, a server 830 might transmit a requested code for an application program through Internet 828, ISP 826, local network(s) 822 and communication interface(s) 818.

[0117] The received code can be executed by processor 804 as it is received, and/or stored in storage device 810, or other non-volatile storage for later execution.

6.2 BASIC SOFTWARE SYSTEM

[0118] Figure 9 is a block diagram of a basic software system 900 that can be employed for controlling the operation of computing device 800. Software system 900 and its components, including their connections, relationships, and functions, is meant to be exemplary only, and not meant to limit implementations of the example embodiment(s). Other software systems suitable for implementing the example embodiment(s) can have different components, including components with different connections, relationships, and functions.

[0119] Software system 900 is provided for directing the operation of computing device 800. Software system 900, which can be stored in system memory (RAM) 806 and on fixed storage (e.g., hard disk or flash memory) 810, includes a kernel or operating system (OS) 910.

[0120] The OS 910 manages low-level aspects of computer operation, including managing execution of processes, memory allocation, file input and output (I/O), and device I/O. One or more application programs, represented as 902A, 902B, 902C ... 902N, can be "loaded" (e.g., transferred from fixed storage 810 into memory 806) for execution by the system 900. The applications or other software intended for use on device 900 can also be stored as a set of downloadable computer-executable instructions, for example, for downloading and installation from an Internet location (e.g., a Web server, an app store, or other online service).

[0121] Software system 900 includes a graphical user interface (GUI) 915, for receiving user commands and data in a graphical (e.g., "point-and-click" or "touch gesture") fashion. These inputs, in turn, can be acted upon by the system 900 in accordance with instructions from operating system 910 and/or application(s) 902. The GUI 915 also serves to display the results of operation from the OS 910 and application(s) 902, whereupon the user can supply additional inputs or terminate the session (e.g., log off).

[0122] OS 910 can execute directly on the bare hardware 920 (e.g., processor(s) 804) of device 800. Alternatively, a hypervisor or virtual machine monitor (VMM) 930 can be interposed between the bare hardware 920 and the OS 910. In this configuration, VMM 930 acts as a software "cushion" or virtualization layer between the OS 910 and the bare hardware 920 of the device 800.

[0123] VMM 930 instantiates and runs one or more virtual machine instances ("guest machines"). Each guest machine comprises a "guest" operating system, such as OS 910, and one or more applications, such as application(s) 902, designed to execute on the guest operating system. The VMM 930 presents the guest operating systems with a virtual operating platform and manages the execution of the guest operating systems.

[0124] In some instances, the VMM 930 can allow a guest operating system to run as if it is running on the bare hardware 920 of device 800 directly. In these instances, the same version of the guest operating system configured to execute on the bare hardware 920 directly can also execute on VMM 930 without modification or reconfiguration. In other words, VMM 930 can provide full hardware and CPU virtualization to a guest operating system in some instances.

[0125] In other instances, a guest operating system can be specially designed or configured to execute on VMM 930 for efficiency. In these instances, the guest operating system is "aware" that it executes on a virtual machine monitor. In other words, VMM 930 can provide para-virtualization to a guest operating system in some instances.

[0126] The above-described basic computer hardware and software is presented for purpose of illustrating the basic underlying computer components that can be employed for implementing the example embodiment(s). The example embodiment(s), however, are not necessarily limited to any particular computing environment or computing device configuration. Instead, the example embodiment(s) can be implemented in any type of system architecture or processing environment that one skilled in the art, in light of this disclosure, would understand as capable of supporting the features and functions of the example embodiment(s) presented herein.

7.0 EXTENSIONS AND ALTERNATIVES

[0127] In the foregoing specification, the example embodiment(s) of the present invention have been described with reference to numerous specific details. However, the details can vary from implementation to implementation according to the requirements of the particular implement at hand. The example embodiment(s) are, accordingly, to be regarded in an illustrative rather than a restrictive sense.

CLAIMS

1. A method, comprising:

receiving, from a metadata server, an identification of content item blocks, wherein a content item comprises the content item blocks;

based, at least in part, on the received identification, storing a replication log entry in a replication task log, the replication task log entry comprising the identification of content item blocks and an identifier of a content item block server; and

based, at least in part, on processing the replication task log entry, transmitting, to the content item block server identified in the replication task log entry, an offer to send at least one content item block, of the content item blocks of which the content item is composed, to the content item block server identified in the replication log entry;

receiving, from the content item block server, an acceptance of the offer to send the at least one content item block; and

based, at least in part, on the received acceptance, sending the at least one content item block to the content item block server.

2. The method of Claim 1, further comprising:

based, at least in part, on completing the processing of the replication log entry, updating the replication log entry to indicate the replication log entry is complete.

3. The method of Claim 1, wherein all of the content item blocks of which the content item is composed are offered to the content item block server identified in the replication log entry.

4. The method of Claim 1, wherein the acceptance identifies one or more content item blocks, of the content item blocks of which the content item is composed, to send; wherein the identified one or more content item blocks do not include all of the content item blocks of which the content item is composed; and wherein the method further comprises sending the identified one or more content item blocks to the content item block server.

5. The method of Claim 1, wherein:

the content item block server is a first content item block server; and

the method further comprises, prior to the processing the replication log entry, downloading at least one of the content item blocks, of which the content item is composed, from a second content item block server that is not the first content item block server.

6. The method of Claim 1, further comprising:

determining a current client journal cursor value associated with a content item namespace;

sending the current client journal cursor value and an identifier of the associated content item namespace to the metadata sever; and

wherein the received identification of content item blocks, of which the content item is composed, is based, at least in part, on the sent current client journal cursor value and the content item namespace identifier.

7. The method of Claim 1, further comprising:

determining that a content item namespace is assigned to the content item block server, the content item belonging to the content item namespace;

wherein the offering to send the at least one content item block, of the content item blocks of which the content item is composed, to the content item block server is based, at least in part, on the determining that the content item namespace is assigned to the content item block server.

8. The method of Claim 1, wherein:

the content item block server is a first content item block server;

the replication log entry comprises an identifier of a second content item block server that is not the first content item block server; and

the method further comprises offering to send the at least one content item block, of the content item blocks of which the content item is composed, to the second content item block server identified in the replication log entry, based, at least in part, on the processing the replication log entry.

9. The method of Claim 1, wherein:

the method is performed by a computing system;

the content item block server is a first content item block server;

the content item belongs to a content item namespace;

the method further comprises:

determining a second content item block server that is not the first content item block server to which the content item namespace is assigned;

determining that a peering relationship does not exist with between the computing system and the second content item block server; and

determining not to offer to send to the second content item block server any of the content item blocks, of which the content item is composed, based, at least in part, on the determining that the peering relationship does not exist; and

the replication log entry does not identify the second content item block server.

10. The method of Claim 1, wherein:

the method is performed by a computing system;

the method further comprises identifying, for deletion according to a least recently used (LRU) criterion, one or more content item blocks that are locally stored by the computing system; and

the LRU criterion is based on content item block download time, content item block upload time, or content item block download time and content item block upload time.

11. The method of Claim 1, wherein the receiving, from the metadata server, the identification of content item blocks, of which the content item is composed, is based, at least in part, on an assignment of the content item namespace to the content item block server.

12. A method comprising:

receiving, from a metadata server, an identification of content item blocks wherein a content item comprises the content item blocks;

based, at least in part, on the received identification, storing a replication log entry in a replication log, the replication log entry comprising the identification of content item blocks and an identifier of an off-premises content item block server of a content management service; and

based, at least in part, on processing the replication log entry, downloading at least one content item block, of the content item blocks of which the content item is composed, from the off-premises content item block server identified in the replication log entry.

13. The method of Claim 12, wherein all of the content item blocks, of which the content item is composed, are downloaded from the off-premises content item block server.

14. The method of Claim 12, further comprising:

based, at least in part, on completing the processing of the replication log entry, updating, after downloading all of the content item blocks, of which the content item is composed, the replication log entry to indicate completion.

15. One or more storage media storing instructions which, when executed by one or more processors, cause performance of a method as recited in any one of Claims 1-14.

16. A system, comprising:

one or more processors; and

one or more storage media storing instructions which, when executed by one or more processors, cause performance of a method as recited in any one of Claims 1-14.

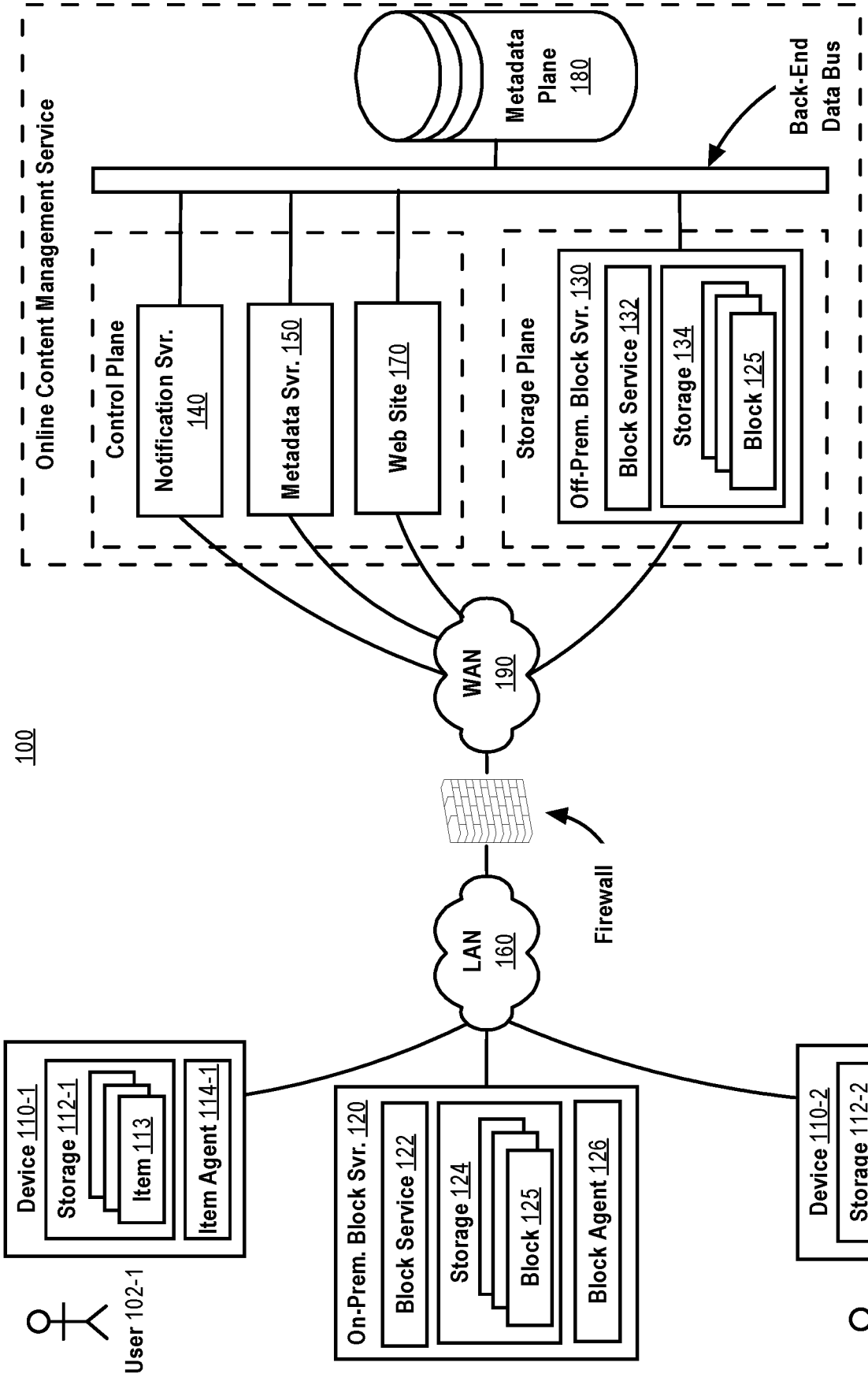
Online Content Management Service
Control Plane
Notification Svr. 140
Metadata Svr. 150
Web Site 170
Storage Plane
Off-Prem. Block Svr. 130
Block Service 132
Storage 134
Block 125
Metadata Plane 180
Back-End Data Bus
WAN 190
Firewall
LAN 160
100
Device 110-1
Storage 112-1
Item 113
Item Agent 114-1
User 102-1
On-Prem. Block Svr. 120
Block Service 122
Storage 124
Block 125
Block Agent 126
Device 110-2
Storage 112-2
Item 113
Item Agent 114-2
User 102-2

Figure 1

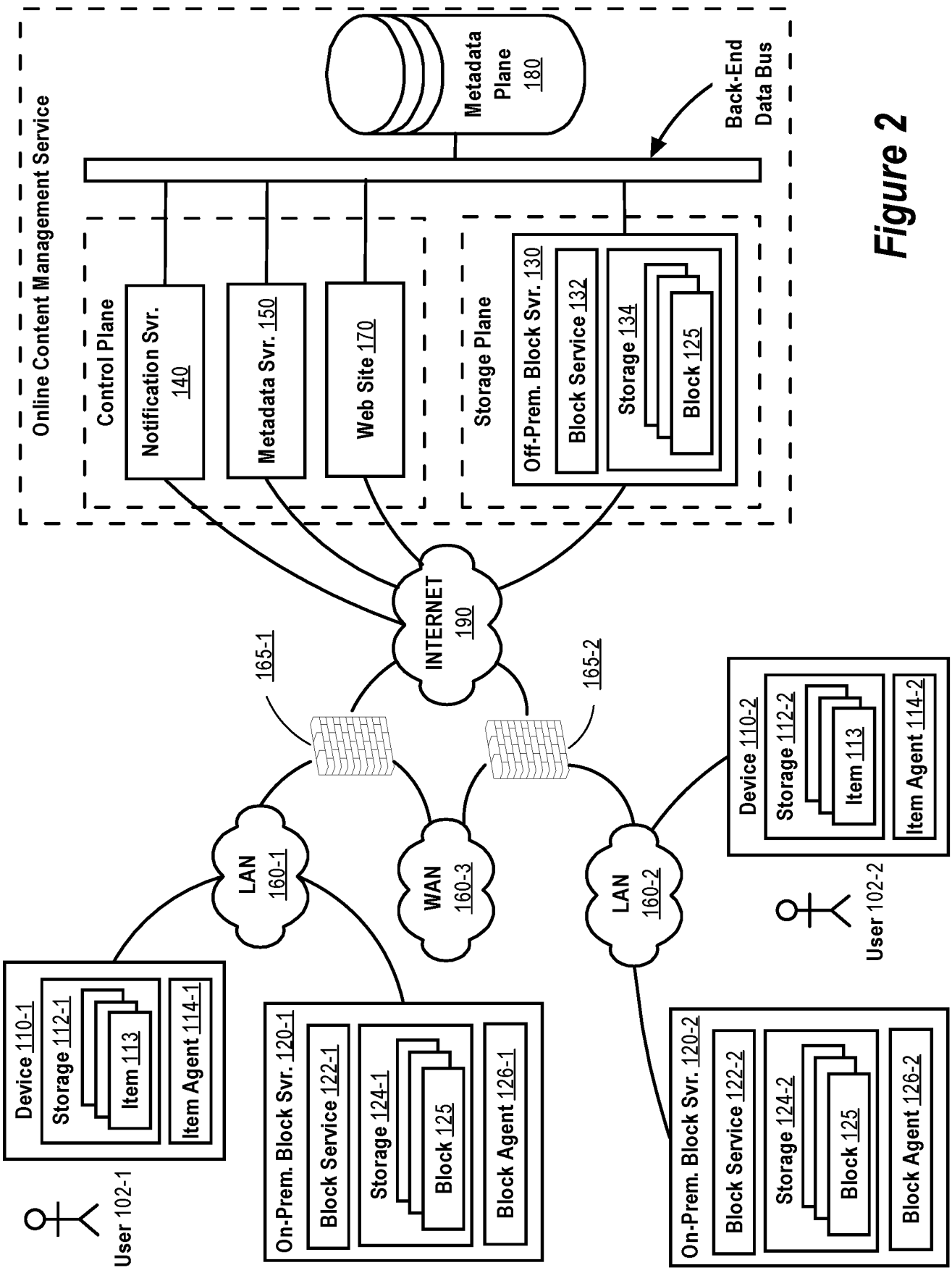
Online Content Management Service
Control Plane
Notification Svr. 140
Metadata Svr. 150
Web Site 170
Storage Plane
Off-Prem. Block Svr. 130
Block Service 132
Storage 134
Block 125
Metadata Plane 180
Back-End Data Bus
INTERNET 190
165-1
165-2
LAN 160-1
WAN 160-3
LAN 160-2
Device 110-1
Storage 112-1
Item 113
Item Agent 114-1
User 102-1
On-Prem. Block Svr. 120-1
Block Service 122-1
Storage 124-1
Block 125
Block Agent 126-1
Device 110-2
Storage 112-2
Item 113
Item Agent 114-2
User 102-2
On-Prem. Block Svr. 120-2
Block Service 122-2
Storage 124-2
Block 125
Block Agent 126-2

Figure 2

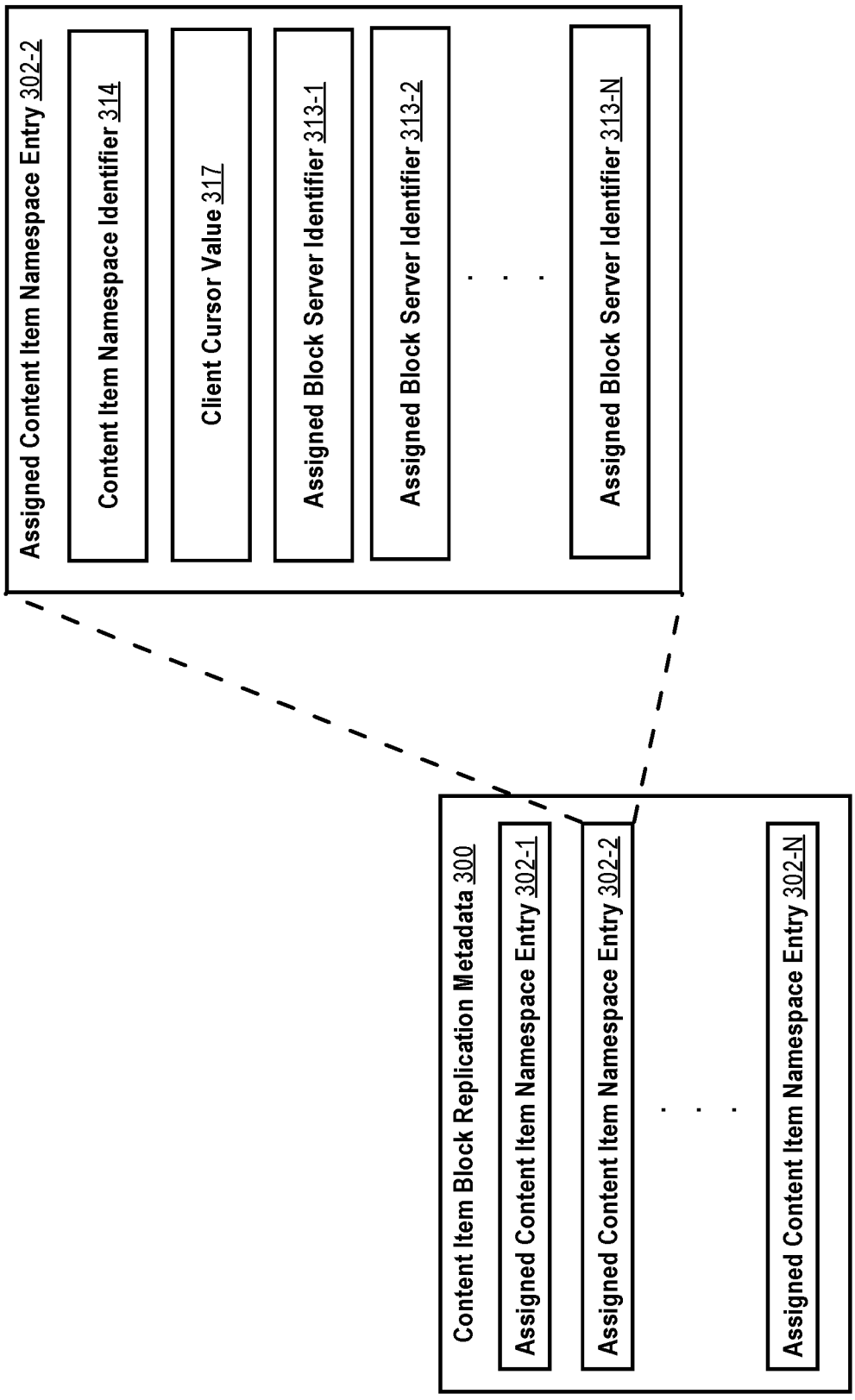
Content Item Block Replication Metadata 300
Assigned Content Item Namespace Entry 302-1
Assigned Content Item Namespace Entry 302-2
Assigned Content Item Namespace Entry 302-N
Assigned Content Item Namespace Entry 302-2
Content Item Namespace Identifier 314
Client Cursor Value 317
Assigned Block Server Identifier 313-1
Assigned Block Server Identifier 313-2
Assigned Block Server Identifier 313-N

Figure 3

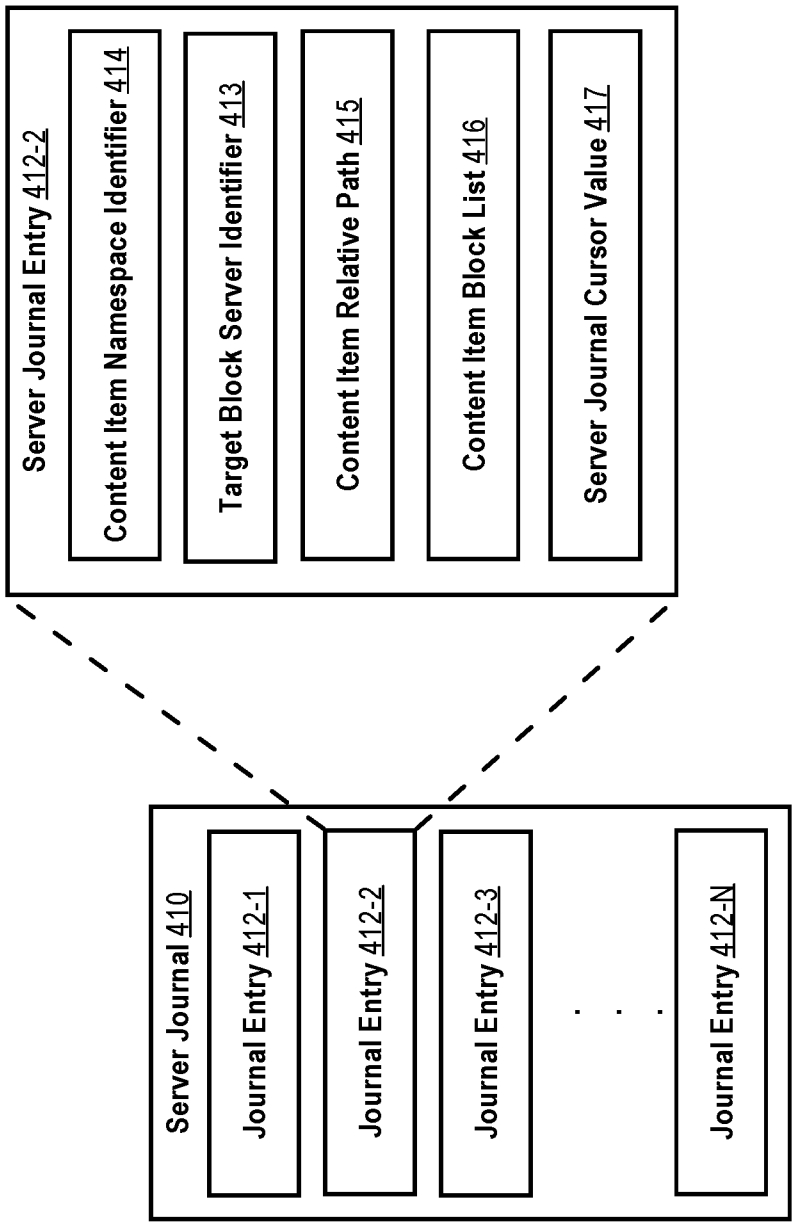
Server Journal Entry 412-2
Content Item Namespace Identifier 414
Target Block Server Identifier 413
Content Item Relative Path 415
Content Item Block List 416
Server Journal Cursor Value 417
Server Journal 410
Journal Entry 412-1
Journal Entry 412-2
Journal Entry 412-3
Journal Entry 412-N

Figure 4

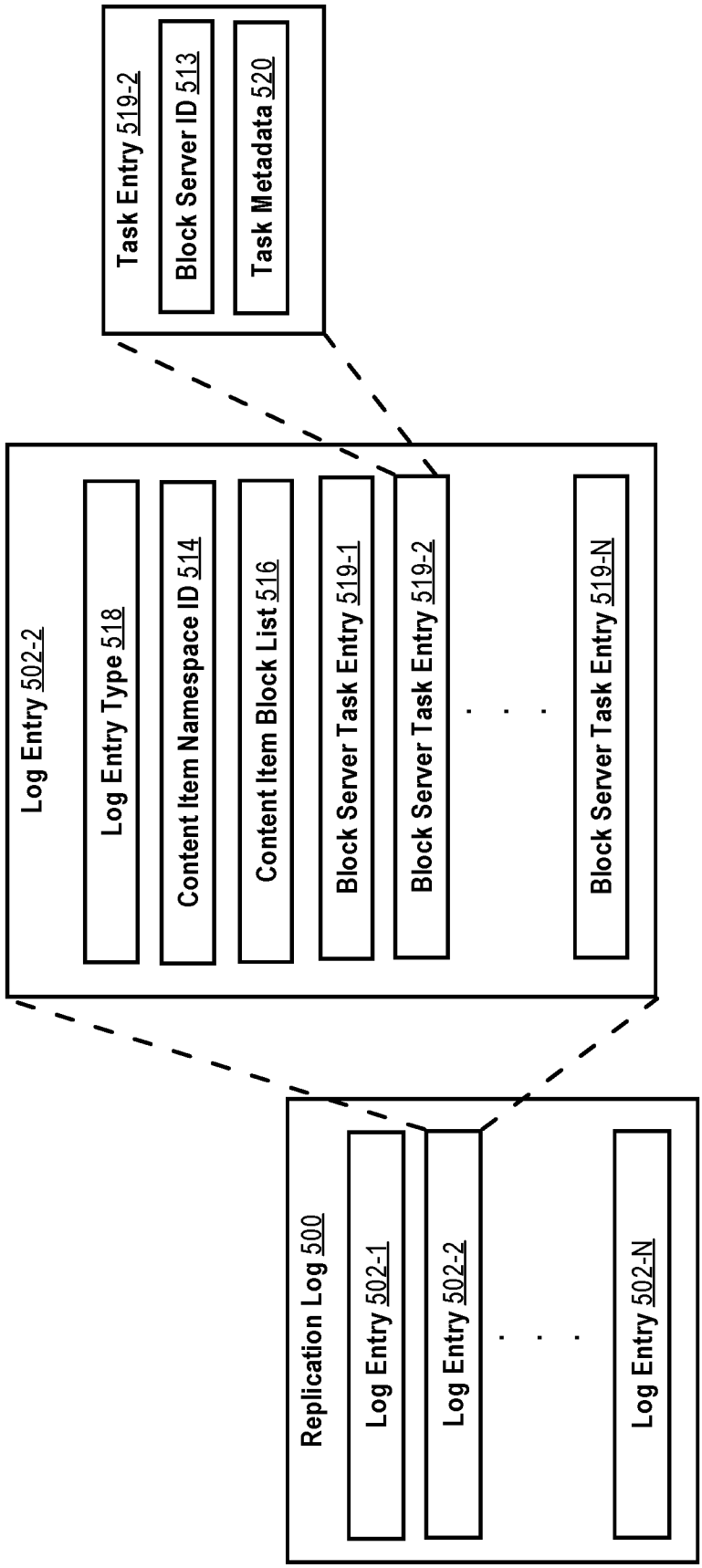
Replication Log 500
Log Entry 502-1
Log Entry 502-2
Log Entry 502-N
Log Entry 502-2
Log Entry Type 518
Content Item Namespace ID 514
Content Item Block List 516
Block Server Task Entry 519-1
Block Server Task Entry 519-2
Block Server Task Entry 519-N
Task Entry 519-2
Block Server ID 513
Task Metadata 520

Figure 5

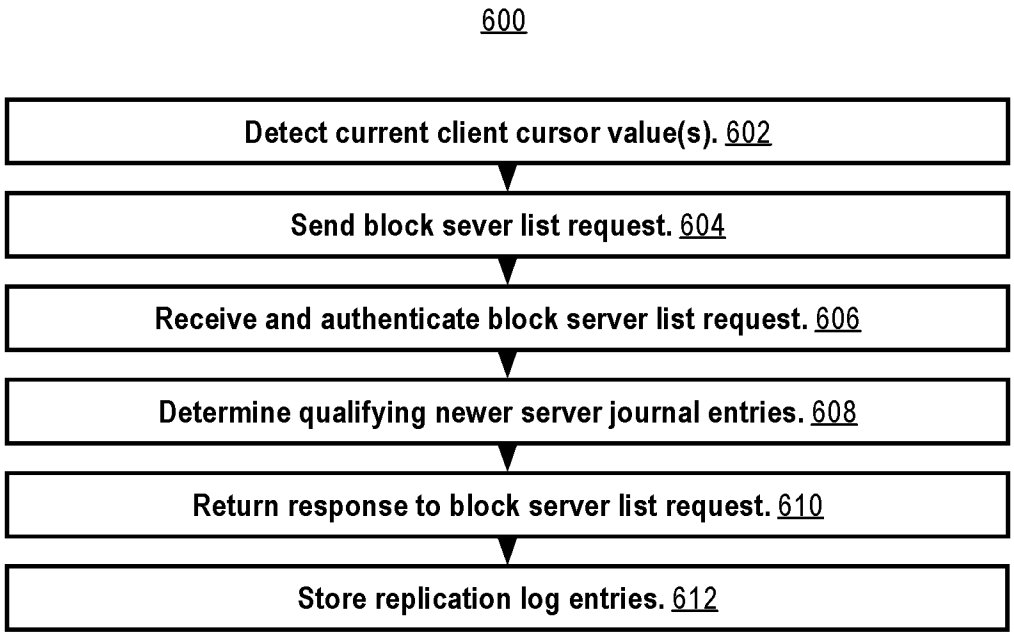
600
Detect current client cursor value(s). 602
Send block sever list request. 604
Receive and authenticate block server list request. 606
Determine qualifying newer server journal entries. 608
Return response to block server list request. 610
Store replication log entries. 612

Figure 6

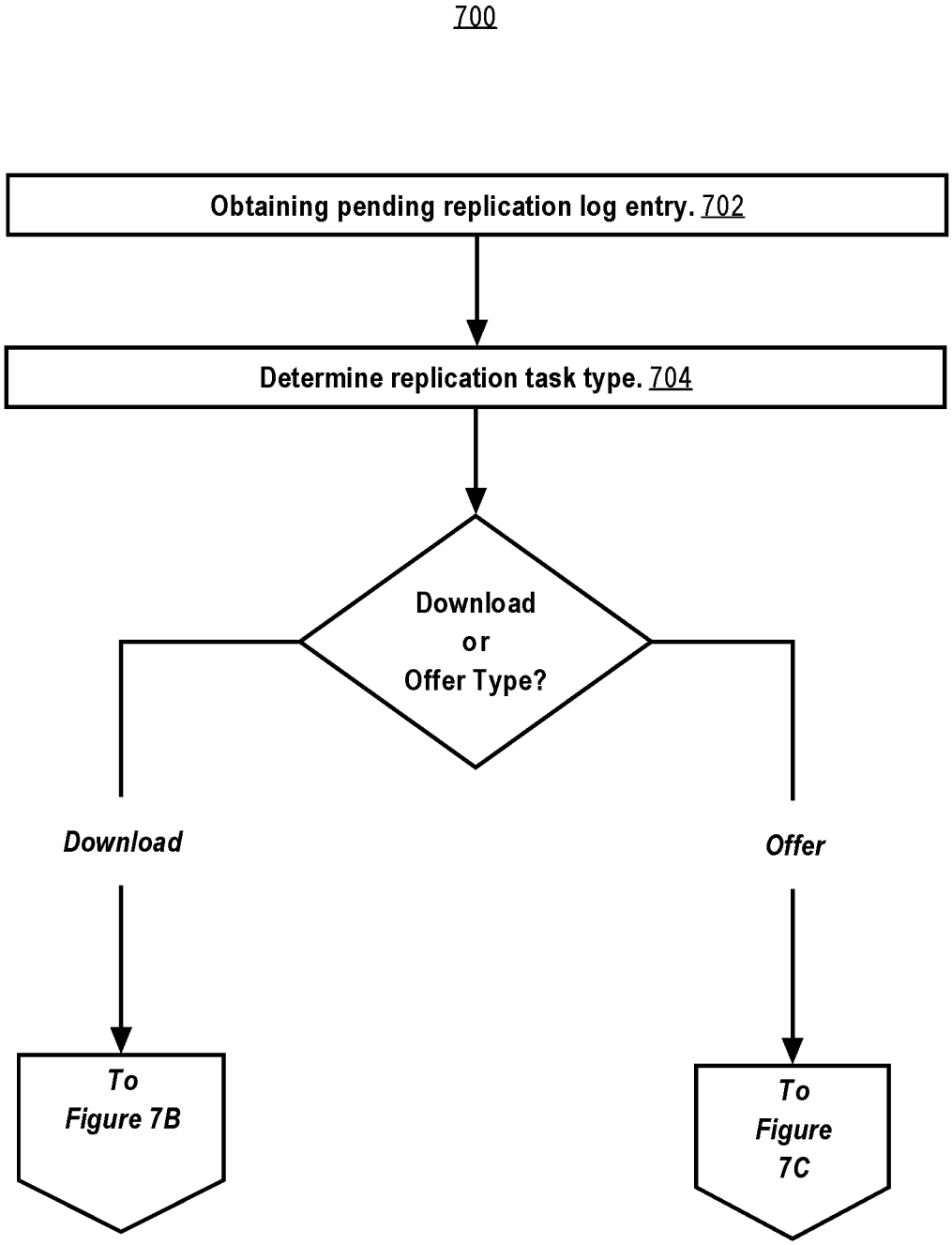
700
Obtaining pending replication log entry. 702
Determine replication task type. 704
Download or Offer Type?
Download
Offer
To Figure 7B
To Figure 7C

Figure 7A

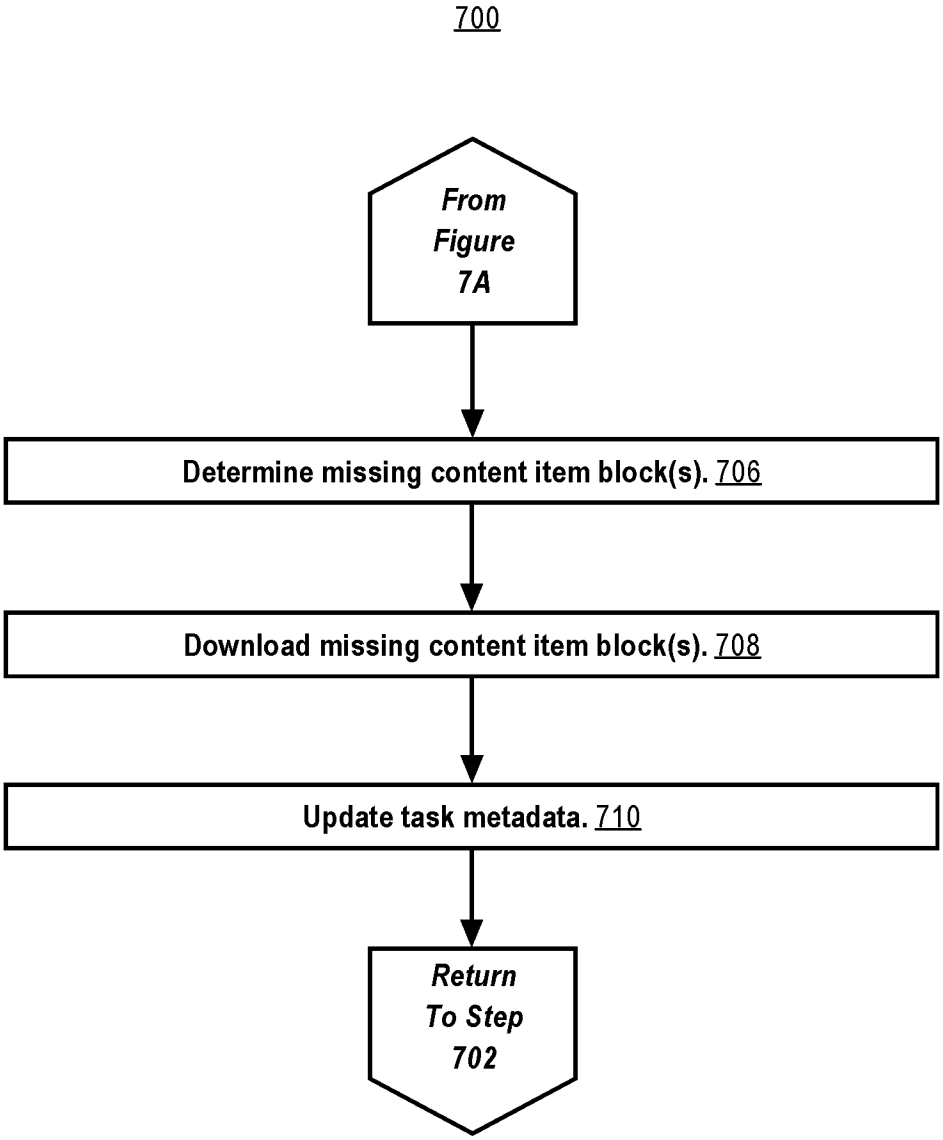
700
From Figure 7A
Determine missing content item block(s). 706
Download missing content item block(s). 708
Update task metadata. 710
Return To Step 702

Figure 7B

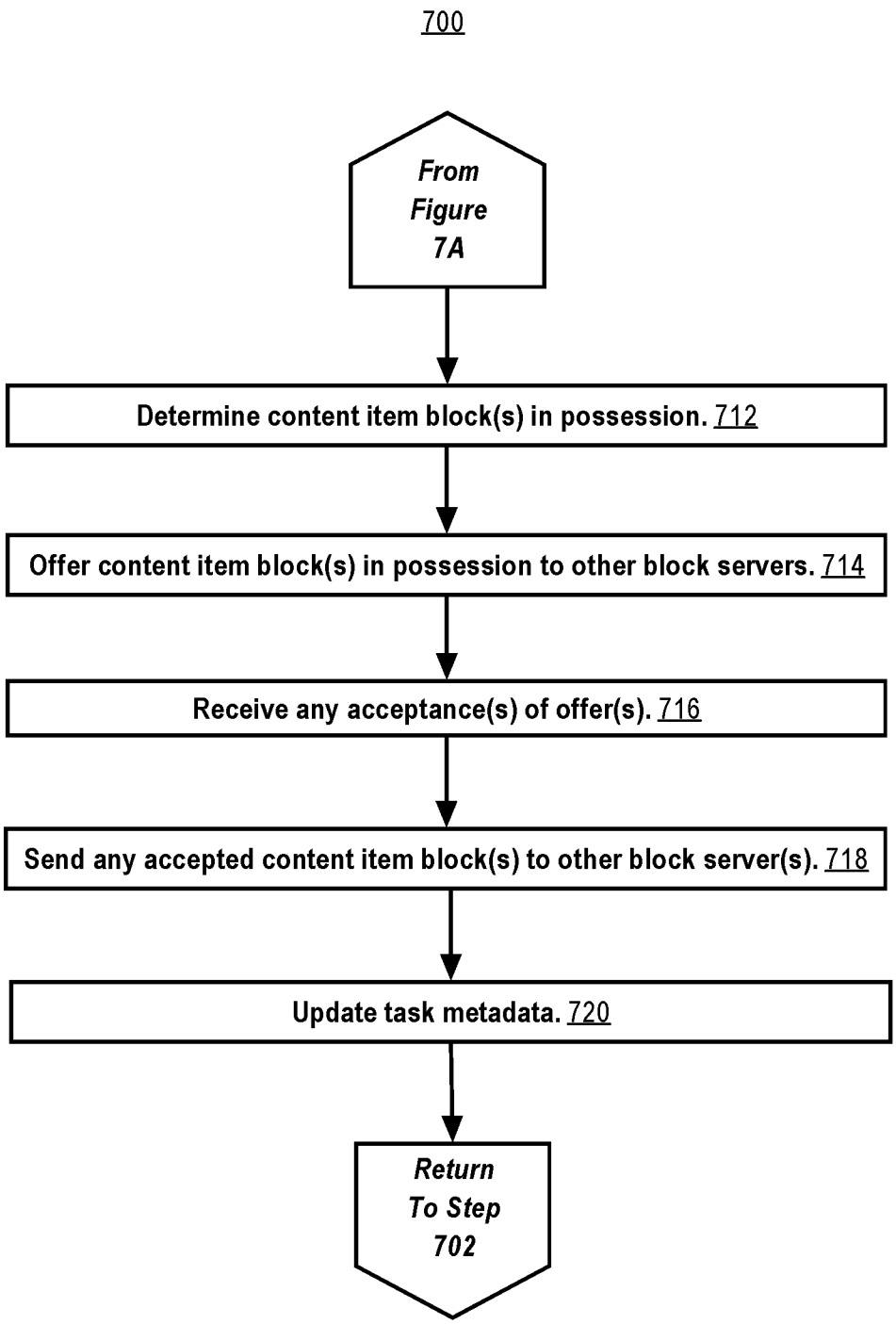
700
From Figure 7A
Determine content item block(s) in possession. 712
Offer content item block(s) in possession to other block servers. 714
Receive any acceptance(s) of offer(s). 716
Send any accepted content item block(s) to other block server(s). 718
Update task metadata. 720
Return To Step 702

Figure 7C

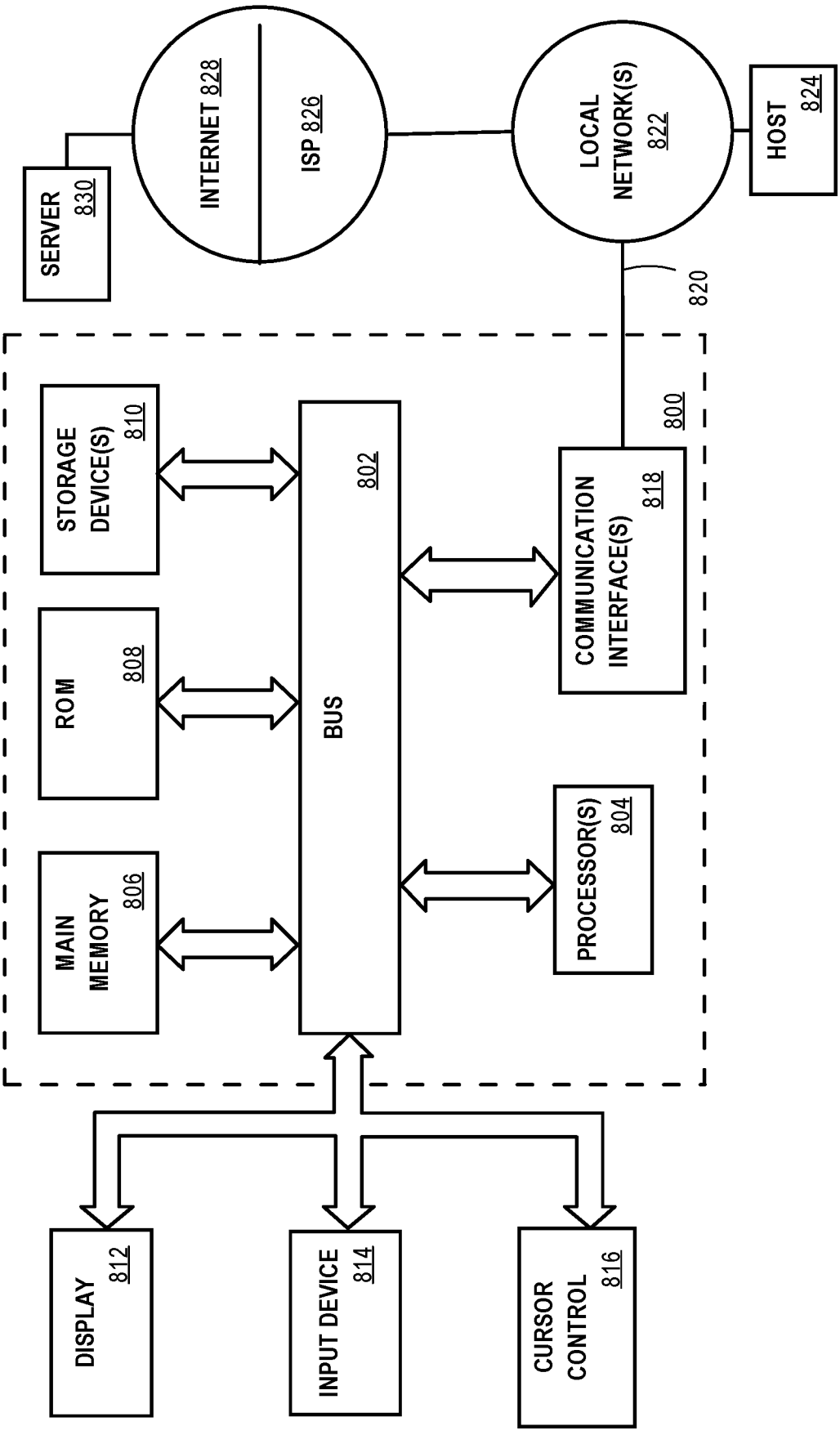
SERVER 830
INTERNET 828
ISP 826
LOCAL NETWORK(S) 822
HOST 824
820
800
STORAGE DEVICE(S) 810
ROM 808
MAIN MEMORY 806
BUS 802
COMMUNICATION INTERFACE(S) 818
PROCESSOR(S) 804
DISPLAY 812
INPUT DEVICE 814
CURSOR CONTROL 816

Figure 8

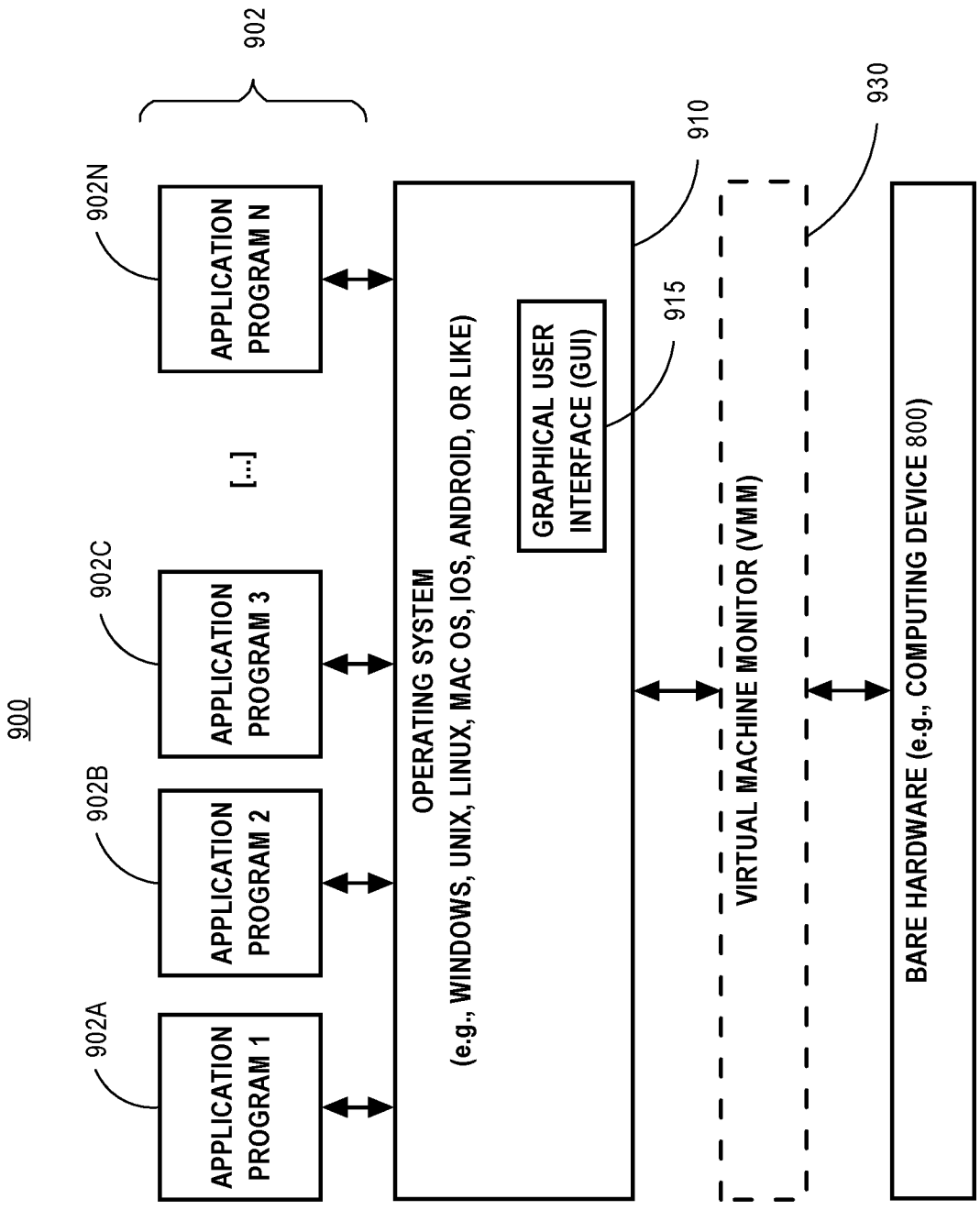
900
902
902A
APPLICATION PROGRAM 1
902B
APPLICATION PROGRAM 2
902C
APPLICATION PROGRAM 3
[...]
902N
APPLICATION PROGRAM N
OPERATING SYSTEM
(e.g., WINDOWS, UNIX, LINUX, MAC OS, IOS, ANDROID, OR LIKE)
910
GRAPHICAL USER INTERFACE (GUI)
915
VIRTUAL MACHINE MONITOR (VMM)
930
BARE HARDWARE (e.g., COMPUTING DEVICE 800)

Figure 9