

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
17 July 2008 (17.07.2008)

PCT

(10) International Publication Number
WO 2008/086077 A1

(51) International Patent Classification:
G06F 11/20 (2006.01)

(21) International Application Number:
PCT/US2008/050086

(22) International Filing Date: 3 January 2008 (03.01.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/883,272 3 January 2007 (03.01.2007) US

(71) Applicant (for all designated States except US):
RAYTHEON COMPANY [US/US]; 870 Winter Street,
Waltham, MA 02451-1449 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **BALLEW, James, D.** [US/US]; 500 Woodhill Court, Grapevine, TX 76051 (US). **DAVIDSON, Shannon, V.** [US/US]; 3484 Jarvis Road, Hillsboro, MO 63050 (US).

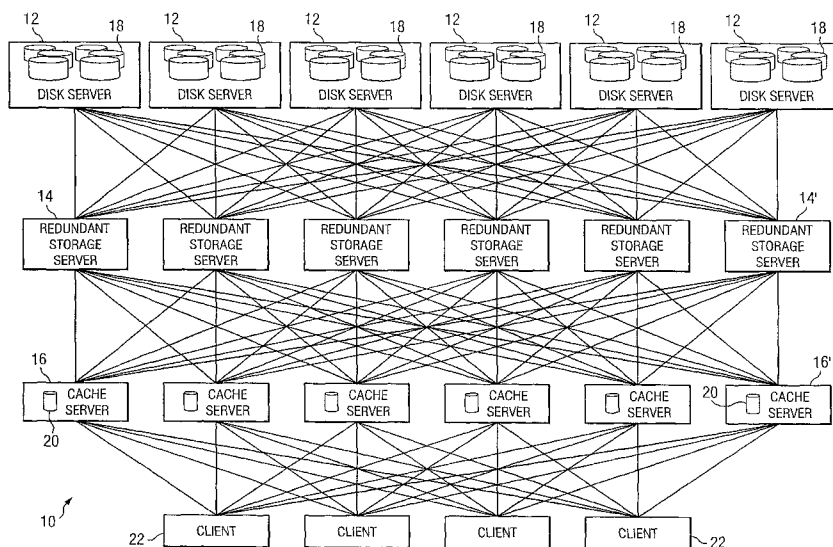
(74) Agents: **MEEK, Kevin, J.** et al.; Baker Botts L.L.P., 2001 Ross Avenue, Suite 600, Dallas, TX 75201 (US).

(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, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, 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, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— with international search report
— before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments

(54) Title: COMPUTER STORAGE SYSTEM



(57) Abstract: According to one embodiment, a computer storage system includes one or more redundant storage servers coupled to one or more cache servers. A redundant storage server is coupled to each disk server. A disk server comprises at least one mass storage disk operable to store data. The data is segmented according to logical blocks, where each logical block has an associated logical block identifier. The redundant storage servers are operable to replicate each logical block of at least two of the disk servers. The cache servers comprise a cache memory and are coupled to each redundant storage server. Each cache server is operable to access the replicated logical blocks according to the associated logical block identifiers, and cache, in the cache memory, the replicated logical block according to the associated logical block identifier.

WO 2008/086077 A1

COMPUTER STORAGE SYSTEM

RELATED APPLICATIONS

This application claims benefit under 35 U.S.C. § 119(e) of U.S. Patent Application Serial No. 60/883,272, entitled "LOW COST COMPUTER STORAGE SYSTEM," which was
5 filed on January 3, 2007.

TECHNICAL FIELD OF THE DISCLOSURE

This disclosure relates to computer storage systems, and more particularly, to a computer storage system and a
10 method of operating the same.

BACKGROUND OF THE DISCLOSURE

Storage area networks have been developed to meet the complex computing requirements of large organizations. A storage area network enables distributed access to data stored in one or more disk servers. The distributed nature of the network provides for storage of relatively large amounts of data and enables the use of redundancy that allows for increased availability.

SUMMARY OF THE DISCLOSURE

According to one embodiment, a computer storage system includes one or more redundant storage servers coupled to one or more cache servers. A redundant storage server is coupled to each disk server. A disk server comprises at least one mass storage disk operable to store data. The data is segmented according to logical blocks, where each logical block has an associated logical block identifier. The redundant storage servers are operable to replicate each logical block of at least two of the disk servers. The cache servers comprise a cache memory and are coupled to each redundant storage server. Each cache server is operable to access the replicated logical blocks according to the associated logical block identifiers, and cache, in the cache memory, the replicated logical block according to the associated logical block identifier.

Certain embodiments may provide numerous technical advantages. A technical advantage of one embodiment may be that redundant disk servers, storage servers, and cache servers provide a relatively fault tolerant

computer storage system having relatively high availability.

Although specific advantages have been enumerated above, various embodiments may include all, some, or none
5 of the enumerated advantages. Additionally, other technical advantages may become readily apparent to one of ordinary skill in the art after review of the following figures and description.

10 BRIEF DESCRIPTION OF THE DRAWINGS

A more complete understanding of embodiments of the disclosure will be apparent from the detailed description taken in conjunction with the accompanying drawings in which:

15 FIGURE 1 is a block diagram showing one embodiment of a computer storage system according to the teachings of the present disclosure;

FIGURE 2 is a diagram showing one embodiment of a cache server of the computer storage system of FIGURE 1;
20 and

FIGURE 3 is a flowchart showing one embodiment of a series of steps that may be taken by the computer storage system of FIGURE 1.

25 DETAILED DESCRIPTION OF EXAMPLE EMBODIMENTS

Storage area networks (SANs) couple mass storage disks to clients through a network. Large entities, such as corporations, may use an array of mass storage disks coupled together through a storage area network to store
30 a relatively large amount of data. Servers may use high performance protocols, such as a fiber channel (FC)

protocol to provide access to the data. Storage systems utilizing these high performance networks, however, may be relatively expensive to implement and maintain.

FIGURE 1 shows one embodiment of a computer storage system 10. Computer storage system 10 includes one or more disk servers 12, one or more redundant storage servers 14, and one or more cache servers 16 that store data that is accessed by one or more clients 22. Data may be stored in one or more mass storage disks 18 configured in each disk server 12. Each redundant storage server 14 may be coupled to each disk server 12. Each cache server 16 may be coupled to each redundant storage server 14. As will be described in greater detail below, computer storage system 10 may provide a storage architecture with a high degree of performance and availability.

Disk servers 12 store data in mass storage disks 18. Mass storage disks 18 may have any suitable capacity, such as greater than or equal to 32 tera-bytes of data. Data in mass storage disks 18 may be accessed at any suitable access rate, such as greater than or equal to 200 Giga-bytes per second.

Data may be segmented in mass storage disks 18 according to logical blocks. Logical blocks generally comprise a portion of useable memory of a mass storage disk 18. The capacity of each logical block generally corresponds with the physical structure of the memory portion of the mass storage disk 18. Logical block identifiers are associated with each logical block. Logical blocks may be accessed according to their associated logical block identifiers.

Redundant storage servers 14 are operable to redundantly store data that is in disk servers 12. Each disk server 12 may have eight 500 Giga-byte mass storage disks 18 with a serial advanced technology attachment (SATA) interface, operating at 7200 revolutions-per-minute (RPM). If a disk server 12 fails, redundant storage server 14 can provide data. In one embodiment, redundant storage servers 14 may comprise a spare redundant storage server 14' that operates in place of a failed redundant storage server 14. A redundant storage server 14 may transmit a heartbeat messages to the other redundant storage servers 14. If a valid response is not received, the redundant storage server 14 transfers operation of the unresponsive redundant storage server 14 to the spare redundant storage server 14'.

In one embodiment, redundant storage servers 14 are implemented according to a Redundant Array of Inexpensive Disk (RAID) protocol; such as a Redundant Array of Inexpensive Disks level 5 (RAID-5) protocol. Any suitable data replication protocol, however, may be used.

Each redundant storage server 14 may be coupled to each disk server 12 using any storage area network protocol. Examples of storage area network protocols include an Internet small computer system interface (iSCSI) protocol, an advanced technology attachment over the Ethernet (AoE) protocol, an Infiniband protocol, a peripheral component interconnect express (PCIe) protocol, or an Ethernet over serial advanced technology attachment (eSATA) protocol. Storage area network protocols provide disk servers 12 access to data that is

stored in logical blocks according to their logical block identifiers.

Cache servers 16 are coupled between clients 22 and redundant storage servers 14. Each cache server 16 may be coupled to each of the redundant storage servers 14. Cache servers 16 access data that is in redundant storage servers 14 and store the data in cache memory 20. Cache memory 20 may be implemented as a level-1 cache. Cache memory 20 may provide up to 96 Giga-bytes of storage and an access rate of up to 400 Giga-bytes per second. Cache servers 16 are described in more detail with reference to FIGURES 2A and 2B.

If a redundant storage server 14 fails, a cache server 16 may provide the data. In one embodiment, cache servers 16 may comprise a spare cache server 16' that operates in place of a failed cache server 16. A cache server 16 may transmit a heartbeat messages to another cache server 16. If a valid response is not received, the cache server 16 transfers operation of the unresponsive cache server 16 to the spare cache server 16'.

Cache servers 16 may be coupled to redundant storage servers 14 and clients 22 using any suitable storage area network protocol. Examples of protocols may include those described with reference to redundant storage system 14. The storage area network protocol may provide clients 22 access to data stored in logical blocks according to their associated logical block identifiers.

Disk servers 12, redundant storage servers 14, and cache servers 16 may have any suitable configuration. In one example, disk servers 12, redundant storage servers

14, and cache servers 16 may comprise commercial-off-the-shelf computing systems having a single socket, dual-core processor. Disk servers 12 and redundant storage servers 14 may be implemented with at least 2 Giga-bytes of system memory, while cache servers 16 may be implemented with at least 12 Giga-bytes of system memory.

Modifications, additions, or omissions may be made to computer storage system 10 without departing from the scope of the invention. The components of computer storage system 10 may be integrated or separated. For example, the operations of cache server 16 may be integrated with redundant storage servers 14. Moreover, the operations of computer storage system 10 may be performed by more, fewer, or other components. For example, disk servers 12 may have any suitable number of mass storage disks 18. Additionally, operations of computer storage system 10 may be performed using any suitable logic comprising software, hardware, and/or other logic. As used in this document, "each" refers to each member of a set or each member of a subset of a set.

FIGURE 2 shows an embodiment of a cache server 16 of FIGURE 1. Cache server 16 includes a processor 24, a cache memory 20, and one or more input/output ports 26. Processor 24 executes instructions stored in cache memory 20. Input/output port 26 couples processor 24 to redundant storage servers 14 and clients 22.

Cache server 16 may comprise any suitable computing system, such as a personal computer, laptop computer, or mainframe computer. In one embodiment, cache server 16 is a blade server that can be placed in a rack among other blade servers. In another embodiment, cache server 16

may comprise a commercial off-the-shelf (COTS) computer having a system memory that may operate as cache memory 20. Cache memory 20 may be any suitable type of memory available for use with commercial off-the-shelf computers, such as dynamic random access memory (DRAM) or static random access memory (SRAM). In another embodiment, cache memory 20 may be implemented as a Level-1 cache.

In one embodiment, cache servers 16 may distribute data over each of the cache servers 16. In this manner, cache memory 20 may be provided by the system memories of a number of commercial off-the-shelf mother boards implemented as cache servers 16.

A portion of data stored in mass storage disks 18 as logical blocks may be replicated in cache memory 20 as logical blocks. Client 22 may access data from cache memory 20 more quickly than from mass storage disks 18. In one embodiment, logical blocks may be formatted according to a Linux block devices that are commonly referred to as logical unit number identifiers (LUNs).

Cache servers 16 may receive requests from clients 22 and either forward the request to redundant storage servers 14 or access the data from a cache memory 20 if available. If the data is retrieved directly from cache memory 20, time latency of a response to the client's request for data may be reduced.

In one embodiment, processor 24 may select a portion of logical blocks for storage in cache memory 20 according to a least-recently-used process. That is, processor 24 may cull logical blocks of data from cache

memory 20 that have been accessed less than other logical blocks.

In a particular embodiment incorporating a number of cache servers 16, logical blocks may be striped over each cache server 16 according to a least significant block address of the logical block identifier. Client 22 may direct requests for data to an appropriate cache server 16 by reading each block address identifier prior to access of the data. In this manner, clients 22 may correctly ascertain the cache server 16 that contains the desired data.

FIGURE 3 shows one embodiment of a series of steps that may be taken by computer storage system 10. In step 100, the process is initiated.

In step 102, cache server 16 receives requests for data from client 22. The requests may include logical block identifiers associated with logical blocks containing the requested data. In one embodiment, cache server 16 may be one of a number of cache servers 16 that are striped according to the least significant block address of the logical block identifiers associated with the logical blocks.

In step 104, cache server 16 accesses the requested logical blocks from mass storage disks 18. In one embodiment, cache server 16 may be coupled to mass storage disks 18 through a storage area network such that the requested logical blocks may be accessed according to their associated logical block identifiers.

In step 106, cache server 16 caches a portion of the requested logical blocks in cache memory 20. In one embodiment, the portion of logical blocks are cached

according to a least-recently-used process. In this manner, logical blocks that are accessed relatively more often may be stored in cache memory 20 for relatively quick access on subsequent access requests from client 5 22. Cache memory 20 stores often used data such that requests for data by clients 22 may be served from cache servers 16, thus alleviating throughput latency of accessing data through disk servers 12.

In step 108, the process ends.

10 Modifications, additions, or omissions may be made to the method without departing from the scope of the disclosure. The method may include more, fewer, or other steps. For example, cache server 16 may access the data from a spare redundant storage server 14 that operates in 15 place of a failed redundant storage server 14.

Although the present disclosure describes several embodiments, a myriad of changes, variations, alterations, transformations, and modifications may be suggested to one skilled in the art, and it is intended 20 that the present disclosure encompass such changes, variations, alterations, transformation, and modifications as they fall within the scope of the appended claims.

What is claimed is:

1. A computer-implemented storage system comprising:
a plurality of redundant storage servers, each
redundant storage server coupled to each of a plurality
of disk servers, each disk server having at least one
5 mass storage disk operable to store data, the data
segmented according to a plurality of logical blocks,
each logical block having an associated logical block
identifier, the plurality of redundant storage servers
operable to:

10 replicate each logical block of at least two of
the plurality of disk servers; and

at least one cache server comprising a cache memory,
the at least one cache server coupled to each redundant
15 storage server, the at least one cache server operable
to:

access the replicated logical blocks according
to the associated plurality of logical block identifiers;
and

20 cache, in the cache memory, a portion of the
replicated logical blocks according to the associated
plurality of logical block identifiers.

2. The computer-implemented storage system of Claim
25 1, wherein the plurality of redundant storage servers
comprise a spare redundant storage server operable to:

operate in place of another redundant storage server
if the other redundant storage server fails.

3. The computer-implemented storage system of Claim 1, wherein the plurality of redundant storage servers comprise a spare redundant storage server, each of the plurality of redundant storage servers operable to:

5 transmit a heartbeat message to another redundant storage server; and

 transfer operation of the other redundant storage server to the spare redundant storage server if a response to the heartbeat message is not received.

10

4. The computer-implemented storage system of Claim 1, wherein the at least one cache server comprises a plurality of cache servers, the plurality of cache servers comprise a spare cache server that is operable to

15 operate in place of another cache server if the other cache server fails.

5. The computer-implemented storage system of Claim 1, wherein the at least one cache server comprises a plurality of cache servers, the plurality of cache servers comprise a spare cache server, each of the plurality of cache servers operable to:

20

 transmit a heartbeat message to another cache server; and

25

 transfer operation of the other cache server to the spare cache server if a response to the heartbeat message is not received.

6. The computer-implemented storage system of Claim 1, wherein the plurality of redundant storage servers are coupled to each of the plurality of storage servers using a storage area network protocol.

5

7. The computer-implemented storage system of Claim 1, wherein the plurality of redundant storage servers are coupled to each of the plurality of storage servers using a storage area network protocol selected from the group consisting of Internet small computer system interface (iSCSI), advanced technology attachment over the Ethernet (AoE), Infiniband, peripheral component interconnect express (PCIe), and Ethernet over serial advanced technology attachment (eSATA).

10

15

8. The computer-implemented storage system of Claim 1, wherein the at least one cache server is coupled to each of the plurality of redundant storage servers using a storage area network protocol.

20

9. The computer-implemented storage system of Claim 1, wherein the at least one cache server is coupled to each of the plurality of redundant storage servers using a storage area network protocol selected from the group consisting of Internet small computer system interface (iSCSI), advanced technology attachment over the Ethernet (AoE), Infiniband, peripheral component interconnect express (PCIe), and Ethernet over serial advanced technology attachment (eSATA).

25

30

10. The computer-implemented storage system of Claim 1, wherein the at least one cache server is further operable to:

5 cache the portion of the replicated logical blocks according to a least-recently-used (LRU) process.

11. The computer-implemented storage system of Claim 1, wherein the plurality of redundant storage servers are configured together according to Redundant Array of
10 Inexpensive Disks level 5 (RAID-5) protocol.

12. A computer-implemented method comprising:

accessing, by a plurality of redundant storage servers, a plurality of disk servers, each redundant storage server coupled to each disk server, each disk server having at least one mass storage disk operable to store data, the data segmented according to a plurality of logical blocks, each logical block having an associated logical block identifier;

replicating each logical block of at least two of the plurality of disk servers;

accessing the replicated logical blocks according to the associated plurality of logical block identifiers, the at least one cache server comprising a cache memory, the at least one cache server coupled to each redundant storage server; and

caching, in the cache memory, a portion of the plurality of replicated logical blocks according to the associated plurality of logical block identifiers.

13. The computer-implemented method of Claim 12, the accessing the replicated logical blocks according to the associated plurality of logical block identifiers further comprising:

accessing the replicated logical block through a spare redundant storage server, the spare redundant storage operating in place of another redundant storage server that has failed.

14. The computer-implemented method of Claim 12, further comprising:

transmitting a heartbeat message to another redundant storage server; and

5 transferring operation of the other redundant storage server to a spare redundant storage server if a response to the heartbeat message is not received.

15. The computer-implemented method of Claim 12, the
10 accessing the replicated logical blocks according to the associated plurality of logical block identifiers further comprising:

accessing the replicated logical block through a spare cache server, the spare cache server operating in
15 place of another cache server that has failed.

16. The computer-implemented method of Claim 12, further comprising:

transmitting a heartbeat message to another cache
20 server; and

transferring operation of the other cache server to a spare cache server if a response to the heartbeat message is not received.

25 17. The computer-implemented method of Claim 12, wherein the plurality of redundant storage servers are coupled to each of the plurality of storage servers using an Ethernet protocol.

18. The computer-implemented method of Claim 12, wherein the plurality of redundant storage servers are coupled to each of the plurality of storage servers using a protocol selected from the group consisting of Internet
5 small computer system interface (iSCSI), advanced technology attachment over the Ethernet (AoE), Infiniband, peripheral component interconnect express (PCIe), and Ethernet over serial advanced technology attachment (eSATA).

10

19. The computer-implemented method of Claim 12, wherein the portion of the replicated logical blocks are cached according to a least-recently-used (LRU) process.

15

20. The computer-implemented storage method of Claim 12, wherein the plurality of redundant storage servers are configured together according to Redundant Array of Inexpensive Disks level 5 (RAID-5) protocol.

21. A computer-implemented storage system comprising:

a plurality of redundant storage servers, each redundant storage server coupled to each of a plurality of disk servers using a storage area network protocol, each disk server having a plurality of mass storage disks operable to store data, the data segmented according to a plurality of logical blocks, each logical block having an associated logical block identifier, the plurality of redundant storage servers comprising a spare redundant storage server, the plurality of redundant storage servers operable to:

replicate each logical block of at least two of the plurality of disk servers; and

at least one cache server comprising a cache memory, the at least one cache server coupled to each redundant storage server using a storage area network protocol, the plurality of cache servers comprising a spare cache server, the at least one cache server operable to:

access the plurality of replicated logical blocks according to the associated plurality of logical block identifiers; and

cache, in the cache memory, a portion of the replicated plurality of logical blocks according to the associated plurality of logical block identifiers.

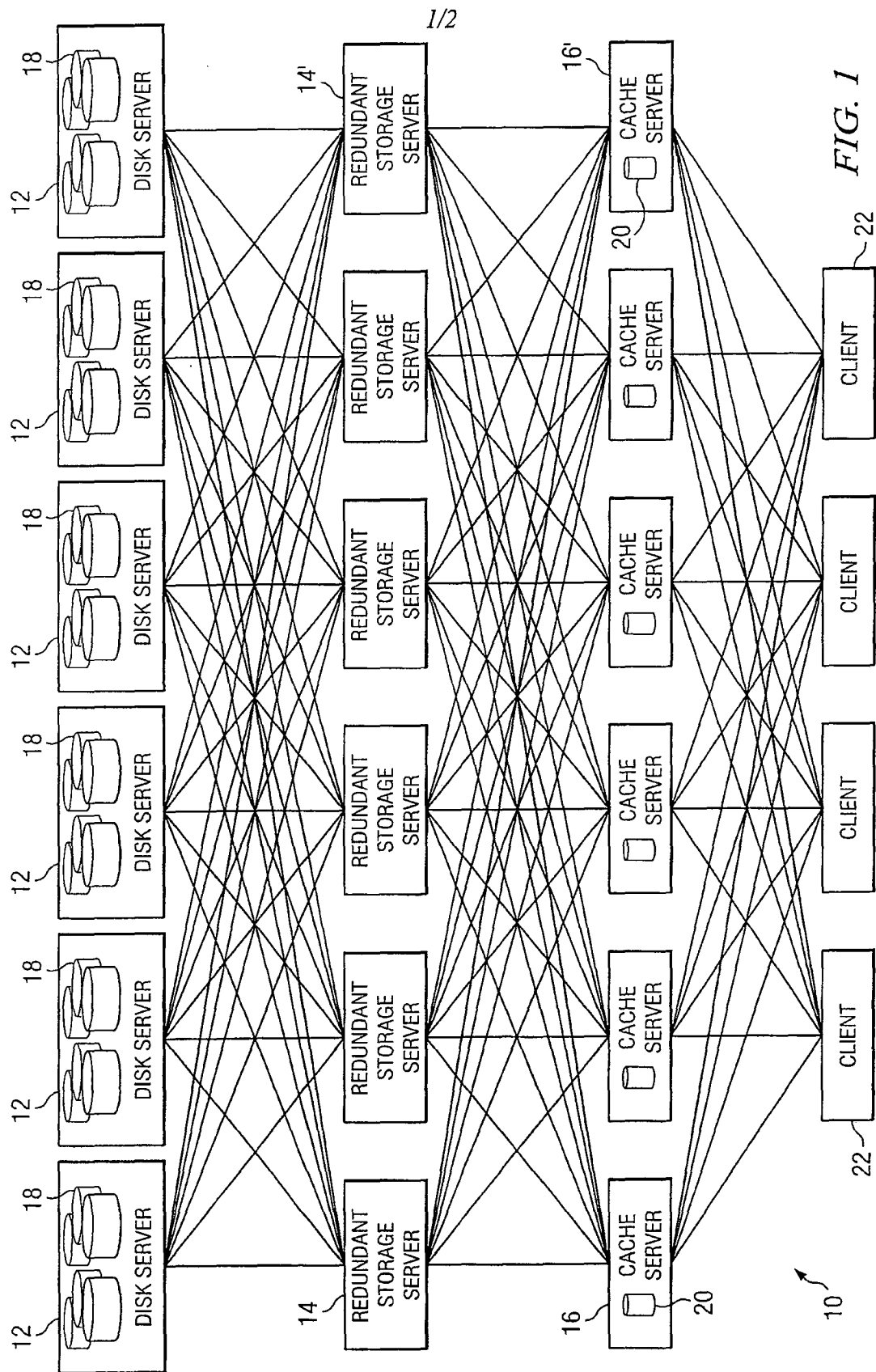


FIG. 1

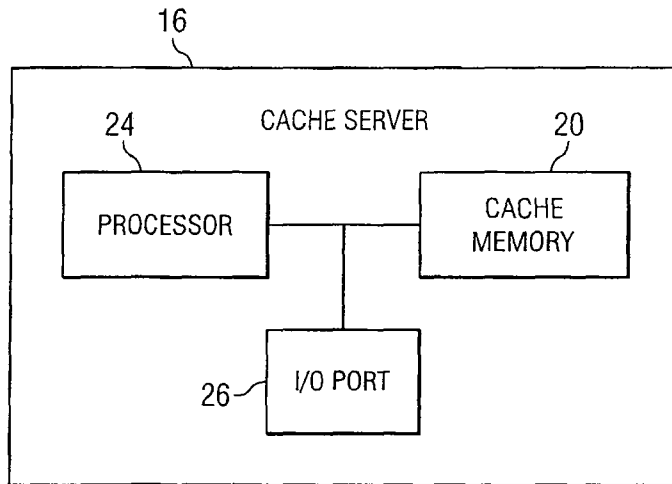


FIG. 2

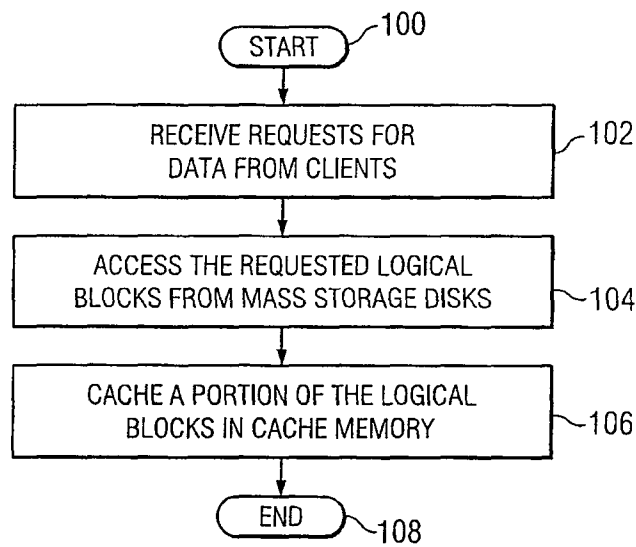


FIG. 3

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2008/050086

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F11/20

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
G06F

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

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

EPO-Internal, WPI Data, IBM-TDB, COMPENDEX, INSPEC

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X Y A	US 6 820 171 B1 (WEBER BRET S [US] ET AL) 16 November 2004 (2004-11-16) column 1, line 26 - line 40 column 2, line 43 - column 3, line 12 column 4, line 55 - column 6, line 29 column 6, line 60 - column 7, line 9 column 7, line 31 - line 51 column 7, line 66 - column 8, line 55 claims 1,11 figures 2,4 ----- -/--	1,6-9,11 2-5 10

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

<p>Special categories of cited documents:</p> <p>*A* document defining the general state of the art which is not considered to be of particular relevance</p> <p>*E* earlier document but published on or after the international filing date</p> <p>*L* document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)</p> <p>*O* document referring to an oral disclosure, use, exhibition or other means</p> <p>*P* document published prior to the international filing date but later than the priority date claimed</p>	<p>*T* later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention</p> <p>*X* document of particular relevance: the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone</p> <p>*Y* document of particular relevance: the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.</p> <p>*Z* document member of the same patent family</p>
---	--

Date of the actual completion of the international search 3 June 2008	Date of mailing of the international search report 10/06/2008
---	---

Name and mailing address of the ISA/ European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl. Fax: (+31-70) 340-3016	Authorized officer Johansson, Ulf
---	---

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2008/050086

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	<p>US 7 127 633 B1 (OLSON MARK DAVID [US] ET AL) 24 October 2006 (2006-10-24) column 3, line 42 - column 4, line 19 column 7, line 9 - line 58 column 9, line 62 - column 10, line 27 column 11, line 1 - line 17 column 12, line 30 - line 49 column 13, line 29 - column 14, line 60 column 20, line 31 - line 67 column 21, line 8 - column 24, line 17 figures 2,5</p> <p style="text-align: center;">-----</p>	2-5
X	<p>US 5 459 857 A (LUDLAM HENRY S [US] ET AL) 17 October 1995 (1995-10-17) column 1, line 60 - column 2, line 19 column 2, line 64 - column 3, line 24 column 3, line 59 - column 4, line 54 column 5, line 20 - column 7, line 30 column 11, line 42 - column 12, line 22 claim 1</p>	1-4,6,8, 10
A	<p>figures 1-3</p> <p style="text-align: center;">-----</p>	5,7,9,11

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2008/050086

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 6820171	B1	16-11-2004	NONE
US 7127633	B1	24-10-2006	NONE
US 5459857	A	17-10-1995	NONE