



(51) International Patent Classification:
G06F 3/06 (2006.01) **H04L 29/08** (2006.01)

(21) International Application Number:
PCT/US2011/043947

(22) International Filing Date:
14 July 2011 (14.07.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
12/837,634 16 July 2010 (16.07.2010) US

(71) Applicant (for all designated States except US): **NETAPP, INC.** [US/US]; 495 East Java Drive, Sunnyvale, California 94089 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **WEBER, Bret** [US/US]; 2521 North Tee Time, Wichita, Kansas 67205 (US). **NOSSOKOFF, Mark** [US/US]; 5400 White Willow Drive, Fort Collins, Colorado 80528 (US). **PEMBLE, Bret** [US/US]; 6185 Davis Creek Road, Silverton, Oregon 97381 (US).

(74) Agents: **KELTON, Thomas** et al.; Haynes and Boone, LLP, IP Section, 2323 Victory Avenue, Suite 700, Dallas, Texas 75219 (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, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

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

Published:

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

(54) Title: ARCHITECTURE FOR IMPROVED CLOUD COMPUTING

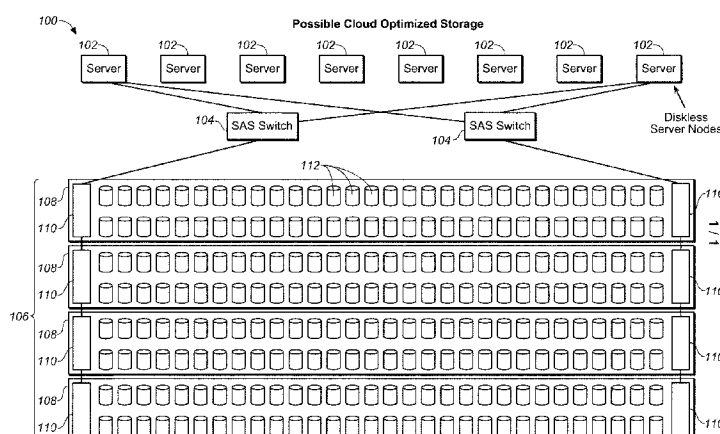


FIG. 1

(57) Abstract: The present invention is directed to an architecture for promoting improved cloud computing. The architecture includes a plurality of diskless server nodes. The architecture further includes a plurality of Serial Attached Small Computer System Interface (SAS) switches, the plurality of SAS switches being connected to the plurality of diskless server nodes. The architecture further includes a storage system, the storage system configured for being communicatively coupled to the plurality of servers via the plurality of SAS switches. Further, the storage system is configured for implementing Controlled Replication Under Scalable Hashing (CRUSH) redundancy. Still further, the architecture is configured for dynamically mapping data stores of the storage system to the diskless server nodes.



ARCHITECTURE FOR IMPROVED CLOUD COMPUTING

FIELD OF THE INVENTION

The present invention relates to the field of storage resource and data management and particularly to an architecture for promoting improved cloud computing.

5 BACKGROUND OF THE INVENTION

Currently available cloud architectures have deficiencies that do not allow them to quickly adapt to different usage deployment models. Compute-based clusters need compute cycles with minimal storage. Storage-based clusters need few compute cycles, but need large amounts of storage. Further, currently available cloud specific nodes are limited in
10 what they can configure.

Therefore, it may be desirable to provide a cloud computing architecture which addresses the above-referenced shortcomings of currently available solutions.

SUMMARY OF THE INVENTION

Accordingly, an embodiment of the present invention is directed to an
15 architecture, including: a plurality of servers; a plurality of Serial Attached Small Computer System Interface (SAS) switches, the plurality of SAS switches being connected to the plurality of servers; and a storage system, the storage system configured for being communicatively coupled to the plurality of servers via the plurality of SAS switches, wherein the architecture is configured for dynamically mapping data stores of
20 the storage system to the servers.

A further embodiment of the present invention is directed to an architecture, including:

a plurality of diskless server nodes; a plurality of Serial Attached Small Computer System Interface (SAS) switches, the plurality of SAS switches being connected to the plurality of diskless server nodes; and a storage system, the storage system configured for being communicatively coupled to the plurality of servers via the plurality of SAS switches, wherein the architecture is configured for dynamically mapping data stores of the storage system to the server nodes.

A still further embodiment of the present invention is directed to an architecture, including: a plurality of diskless server nodes; a plurality of Serial Attached Small Computer System Interface (SAS) switches, the plurality of SAS switches being connected to the plurality of diskless server nodes; and a storage system, the storage system configured for being communicatively coupled to the plurality of servers via the plurality of SAS switches, wherein the storage system is configured for implementing Controlled Replication Under Scalable Hashing (CRUSH) redundancy, wherein the architecture is configured for dynamically mapping data stores of the storage system to the server nodes.

It is to be understood that both the foregoing general description and the following detailed description are exemplary and explanatory only and are not necessarily restrictive of the invention as claimed. The accompanying drawings, which are incorporated in and constitute a part of the specification, illustrate embodiments of the invention and together with the general description, serve to explain the principles of the invention.

BRIEF DESCRIPTION OF THE DRAWINGS

The numerous advantages of the present invention may be better understood by those skilled in the art by reference to the accompanying figures in which:

FIG. 1 is a block diagram illustration of an architecture for promoting improved cloud computing in accordance with an exemplary embodiment of the present invention.

DETAILED DESCRIPTION OF THE INVENTION

Reference will now be made in detail to the presently preferred embodiments of the invention, examples of which are illustrated in the accompanying drawings.

5 Currently available cloud architectures have deficiencies that do not allow them to quickly adapt to different usage deployment models. Compute-based clusters need compute cycles with minimal storage. Storage-based clusters need few compute cycles, but need large amounts of storage. Further, currently available cloud specific nodes are limited in what they can configure. Still further, currently available cloud architectures do not
10 utilize traditional Redundant Array of Inexpensive Disks (RAID) capability since redundancy is inherent in the cloud middleware. The architecture of the present invention disclosed herein: a.) allows for complete dynamic configurability; b.) is compatible with existing cloud middleware components; and c.) implements new methods of Controlled Replication Under Scalable Hashing (CRUSH) redundancy to
15 provide more efficient data redundancy while still allowing higher level mechanisms for more extreme failure mechanisms.

Referring to FIG. 1, an architecture 100 for promoting improved cloud computing (ex. — a cloud computing architecture 100) is shown. In an exemplary embodiment of the present
20 invention, the architecture 100 may include a plurality of servers 102 (exs. — server nodes, processor cards, processors, central processing units (CPUs)). For example, the architecture 100 may include eight servers 102 (ex. — eight processor cards 102). In further exemplary embodiments of the present invention, the servers 102 (ex. — processor cards 102) may include limited or no storage (ex. - may not include drives). For instance, the
25 servers 102 (ex. — server nodes 102) may be diskless server nodes (DSN) 102, such that the servers 102 do not include a boot drive(s). Thus, the servers 102 (ex. — processors 102) of the architecture 100 of the present invention are not tied to (ex. — do not include) storage, even for boot.

30 In current exemplary embodiments of the present invention, the architecture 100 may

include one or more switches 104. For example, the switches 104 may be Serial Attached Small Computer System Interface (SAS) switches 104. The switches 104 may be configured for being connected to the servers 102.

5 In exemplary embodiments of the present invention, the architecture 100 may include a storage system 106. In further embodiments of the present invention, the storage system 106 may be configured for being connected to the plurality of servers 102 via the SAS switches 104, said SAS switches 104 configured for facilitating data communications between servers 102 and the storage system 106. In still further
10 embodiments of the present invention, the storage system 106 may include a plurality of storage subsystems 108, each of the storage subsystems 108 configured for being connected (ex. — communicatively coupled) to each other. In further embodiments of the present invention, each storage subsystem 108 may include one or more storage controllers 110. In still further embodiments of the present invention, each storage
15 subsystem 108 may further include a plurality of disk drives 112, said disk drives 112 being connected to the storage controllers 110. For example, the storage system 106 may include six hundred disk drives 112. In exemplary embodiments of the present invention, the storage subsystems 108 may be communicatively coupled to each other via the storage controllers 110. In further embodiments of the present invention, the storage controllers 110 of the
20 storage system 106 may be communicatively coupled to the servers 102 via the SAS switches 104.

In current exemplary embodiments of the present invention, the storage system 106 is configured for implementing Controlled Replication Under Scalable Hashing (CRUSH)
25 redundancy (ex. — is configured for utilizing large CRUSH data configuration) to provide more efficient data redundancy (ex. — flexible CRUSH mappings) while still allowing higher level mechanisms for more extreme failure mechanisms. Controlled Replication Under Scalable Hashing (CRUSH) is a mechanism for mapping data to storage objects which was developed by the University of California at Santa Cruz. For example, CRUSH techniques
30 are disclosed in: *CRUSH: Controlled, Scalable, Decentralized Placement of*

Replicated Data., Weil et al., Proceedings of SC '06, November 2006, which is herein incorporated by reference in its entirety. CRUSH allows redundancy methods to operate independently of data placement algorithms. For example, a CRUSH system may have as its redundancy mechanism a Redundant Array of Inexpensive Disks (RAID) mechanism/a RAID stripe, such as a RAID 5 4+1 stripe. Each stripe of information on this redundancy group/redundancy mechanism may be mapped by CRUSH to a set/subset of 5 drives within a set of drives of the CRUSH system. Each subsequent stripe of data may be mapped to another set/subset of 5 drives within the set of drives of the CRUSH system.

10 In exemplary embodiments of the present invention, the servers 102 (ex. — processor cards 102) may be dynamically mapped to application and storage requirements. In further embodiments of the present invention, the architecture 100 allows for dynamic virtualized storage. In still further embodiments of the present invention, the architecture 100 allows for flexible mappings between the servers 102 (ex. — CPUs 102) and the disk drives 112 of the storage system 106. In further embodiments of the present invention, the architecture 100, by providing for flexible CRUSH mappings (as mentioned above), allows for: high performance on all volumes of the storage system 106; implementation of RAID redundancy mechanisms, including RAID 6; and fast rebuilds (which may promote a reduction in upper level data copies as well as promoting a decrease in network traffic (such as in a drive failure environment)). In still further embodiments of the present invention, the architecture 100 allows for dynamic configuration of performance node(s) versus storage node(s).

In current exemplary embodiments of the present invention, the servers 102 (ex. — processor cards 102) of the architecture 100 may not include drives (as mentioned above), thus, the architecture 100 may allow for the use of operating system (OS) snapshots from a single volume and/or may allow for the implementation or use of flash swap space. In still further embodiments of the present invention, the architecture 100 may allow for replication at a Distributed File System (DFS) layer, there allowing said architecture to be compatible with current cloud computing infrastructure. In further

embodiments of the present invention, the architecture 100 may allow for extremely quick rebuilds after failures occur. In still further embodiments of the present invention, the architecture 100 may promote the elimination of thrashing at a DFS layer except in the case of catastrophic errors (such as server failures, multiple drive failures), thus
5 increasing effective user bandwidth. In further embodiments of the present invention, the architecture 100 allows for the elimination of traditional Storage Area Network (SAN) infrastructure.

In exemplary embodiments of the present invention, the architecture 100 may allow for improved control for provisioning of resources (ex. — provisioning of processor and
10 storage resources). For example, the architecture 100 of the present invention may allow for allocation of amounts of storage power and processor power for an application. In further embodiments of the present invention, the architecture 100 may allow for expansion capability (ex. — scale-out expansion capability) for promoting improved bandwidth and capacity. In still further embodiments of the present invention, the
15 architecture 100 allows full customer replaceability. In further embodiments of the present invention, the architecture 100 is compatible with currently available cloud software, which may run on the servers 102 (ex. — server nodes 102) without change. In still further embodiments of the present invention, the architecture 100 may be configured (ex. — sized) for implementation within a server cabinet (ex. — a 44U server
20 cabinet). In further embodiments of the present invention, the architecture 100 may be configured (ex. — sized) such that it abstracts well to a container (ex. — a shipping container). The dynamic mapping capability provided by the architecture 100 allows for such abstraction capabilities. For instance, the servers 102 and storage system 106 may be sized such that at least two thousand servers 102 and their associated storage system
25 106 may fit into a standard cloud shipping container.

In current exemplary embodiments of the present invention, the architecture 100 removes any dependencies on processor and storage nodes, thereby allowing for complete flexibility in terms of dynamically configuring any type of cloud computing node. In

- further embodiments of the present invention, the architecture 100 allows for very fast recovery from disk failures and allows any components of the architecture 100 to be replaced with a customer replaceable unit (CRU), all while retaining existing cloud middleware. In still further embodiments of the present invention, the storage system
- 5 106 of the architecture 100 is SAS-switched and utilizes large CRUSH data configuration that allows for fast rebuilds of drive failures. In further embodiments of the present invention, the architecture 100 utilizes a virtualized mapping structure which allows for data stores (of the storage system 106) to be dynamically custom-mapped to the appropriate server 102 (ex — processor complex 102) for a task that is being allocated.
- 10 This also includes boot capability of the node (ex. — the cloud computing node), which may be a writable snapshot of an operating system (OS) boot node. In still further embodiments of the present invention, the architecture 100 is configured for allowing full customer replaceability of the architecture's components and promotes improved performance over existing architectures for cloud computing.
- 15 It is believed that the present invention and many of its attendant advantages will be understood by the foregoing description. It is also believed that it will be apparent that various changes may be made in the form, construction and arrangement of the components thereof without departing from the scope and spirit of the invention or without sacrificing all of its material advantages. The form herein before described being merely an explanatory
- 20 embodiment thereof, it is the intention of the following claims to encompass and include such changes.

CLAIMS

What is claimed is:

1. An architecture, comprising:
a plurality of servers;
5 a plurality of Serial Attached Small Computer System Interface (SAS) switches,
the plurality of SAS switches being connected to the plurality of servers; and
a storage system, the storage system configured for being communicatively
coupled to the plurality of servers via the plurality of SAS switches,
wherein the architecture is configured for dynamically mapping data stores of the
10 storage system to the servers.
2. An architecture as claimed in claim 1, wherein the storage system includes a
plurality of storage subsystems, each storage subsystem included in the plurality of
storage subsystems including at least one storage controller.
3. An architecture as claimed in claim 2, wherein each storage subsystem included
15 in the plurality of storage subsystems includes a plurality of disk drives, the plurality of
disk drives being connected to the at least one storage controller.
4. An architecture as claimed in claim 1, wherein the plurality of servers are diskless
server nodes (DSN).
5. An architecture as claimed in claim 1, wherein the storage system is configured
20 for implementing Controlled Replication Under Scalable Hashing (CRUSH) redundancy.

6. An architecture as claimed in claim 1, wherein the architecture is a cloud computing architecture.

7. An architecture as claimed in claim 1, wherein the storage system is configured
5 for implementing Redundant Array of Inexpensive Disks (RAID) redundancy.

8. An architecture, comprising:

a plurality of diskless server nodes;

a plurality of Serial Attached Small Computer System Interface (SAS) switches, the
10 plurality of SAS switches being connected to the plurality of diskless server nodes; and

a storage system, the storage system configured for being communicatively coupled to the plurality of servers via the plurality of SAS switches,

wherein the architecture is configured for dynamically mapping data stores of the storage system to the server nodes.

15 9. An architecture as claimed in claim 8, wherein the storage system includes a plurality of storage subsystems, each storage subsystem included in the plurality of storage subsystems including at least one storage controller.

10. An architecture as claimed in claim 9, wherein each storage subsystem included in the plurality of storage subsystems includes a plurality of disk drives, the plurality of disk
20 drives being connected to the at least one storage controller.

11. An architecture as claimed in claim 8, wherein the storage system is configured for implementing Controlled Replication Under Scalable Hashing (CRUSH) redundancy.

12. An architecture as claimed in claim 8, wherein the architecture is a cloud computing architecture.

13. An architecture as claimed in claim 8, wherein the storage system is configured for implementing Redundant Array of Inexpensive Disks (RAID) redundancy.

14. An architecture, comprising:

5 a plurality of diskless server nodes;

a plurality of Serial Attached Small Computer System Interface (SAS) switches, the plurality of SAS switches being connected to the plurality of diskless server nodes; and

10 a storage system, the storage system configured for being communicatively coupled to the plurality of servers via the plurality of SAS switches, wherein the storage system is configured for implementing Controlled Replication Under Scalable Hashing (CRUSH) redundancy,

wherein the architecture is configured for dynamically mapping data stores of the storage system to the server nodes.

15 15. An architecture as claimed in claim 14, wherein the storage system includes a plurality of storage subsystems, each storage subsystem included in the plurality of storage subsystems including at least one storage controller.

16. An architecture as claimed in claim 15, wherein each storage subsystem included in the plurality of storage subsystems includes a plurality of disk drives, the plurality of
20 disk drives being connected to the at least one storage controller.

17. An architecture as claimed in claim 16, wherein the plurality of storage subsystems are communicatively coupled to each other.

18. An architecture as claimed in claim 14, wherein the architecture is a cloud computing architecture.

19. An architecture as claimed in claim 14, wherein the storage system is configured for implementing Redundant Array of Inexpensive Disks (RAID) redundancy.
20. An architecture as claimed in claim 19, wherein the storage system is configured for implementing RAID 6.

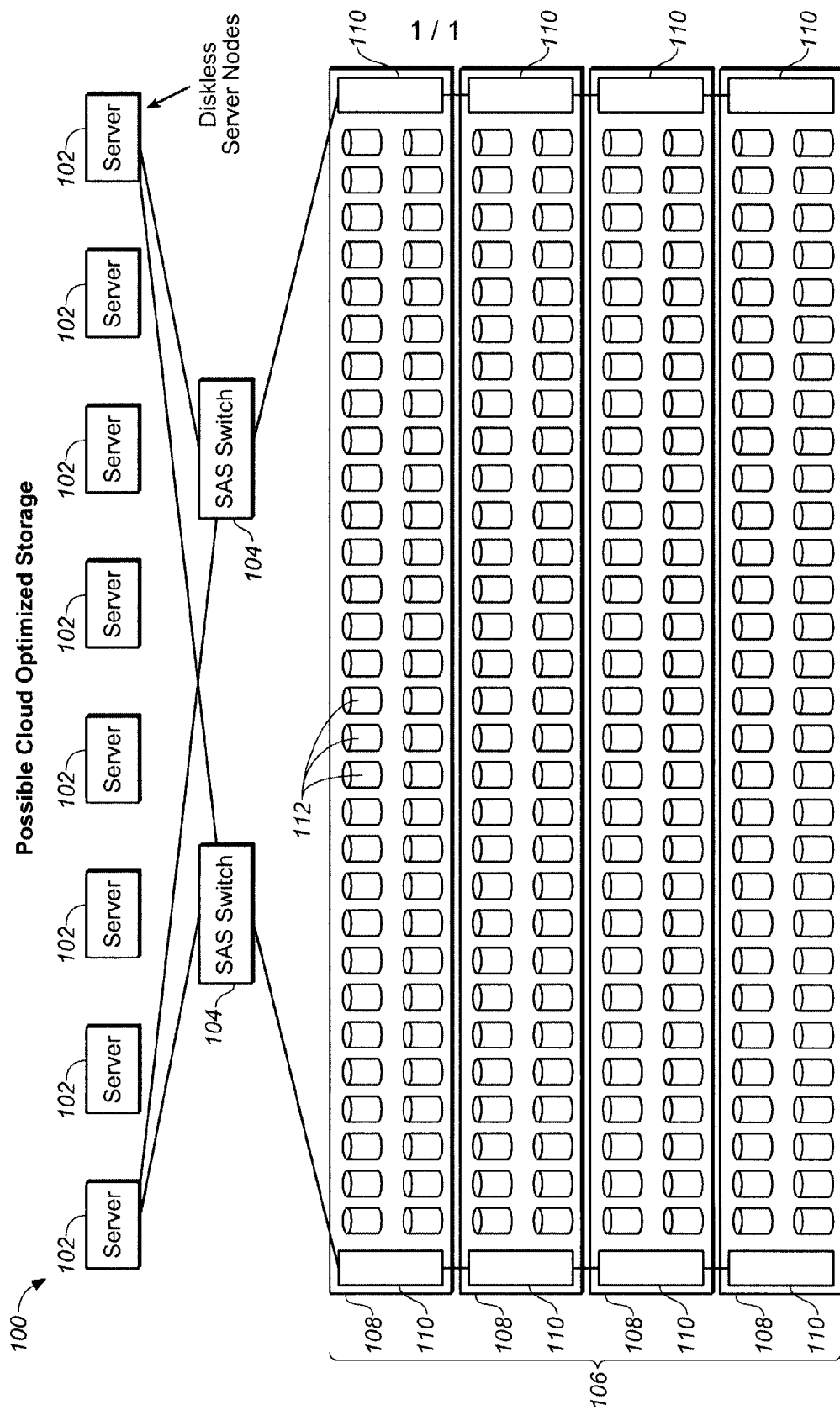


FIG. 1

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2011/043947

A. CLASSIFICATION OF SUBJECT MATTER

INV. G06F3/06 H04L29/08
ADD.

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

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06F H04L

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	BRINKMANN A ET AL: "Cost-Effectiveness of Storage Grids and Storage Clusters", PARALLEL, DISTRIBUTED AND NETWORK-BASED PROCESSING, 2007. PDP '07 . 15TH EUROMICRO INTERNATIONAL CONFERENCE ON, IEEE, PI, 1 February 2007 (2007-02-01), pages 517-525, XP031064203, DOI: 10.1109/PDP.2007.33 ISBN: 978-0-7695-2784-0	1-4, 6-10, 12, 13
Y	the whole document	5, 11, 14-20
Y	----- WEIL ET AL.: "CRUSH: Controlled, Scalable, Decentralized Placement of Replicated Data.", PROCEEDINGS OF SC '06, November 2006 (2006-11), XP040053733,	5, 11, 14-20
A	the whole document ----- -/-	6, 12

☒ Further documents are listed in the continuation of Box C.

☒ See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search

26 August 2011

Date of mailing of the international search report

02/09/2011

Name and mailing address of the ISA/

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

Authorized officer

Limacher, Rolf

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2011/043947

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2007/162592 A1 (MARKS KEVIN [US] ET AL) 12 July 2007 (2007-07-12) paragraph [0001] paragraph [0017] - paragraph [0019]; figure 1 paragraph [0025] -----	1-4, 6-10,12, 13
X	US 2008/126696 A1 (HOLLAND WILLIAM GAVIN [US] ET AL) 29 May 2008 (2008-05-29) paragraph [0002] paragraph [0005] paragraph [0013] paragraph [0019] paragraph [0029]; figure 4 paragraph [0056] -----	1-4, 6-10,12, 13

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2011/043947

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2007162592	A1	12-07-2007	NONE

US 2008126696	A1	29-05-2008	NONE
