

(19) World Intellectual Property Organization
International Bureau



(43) International Publication Date
7 September 2007 (07.09.2007)

PCT

(10) International Publication Number
WO 2007/100521 A2

(51) International Patent Classification: **Not classified**

(21) International Application Number:
PCT/US2007/004142

(22) International Filing Date:
14 February 2007 (14.02.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/365,474 28 February 2006 (28.02.2006) US

(71) Applicant (for all designated States except US): **GEAR SIX, INC.** [US/US]; 1360 Willow Road, Suite 201, Menlo Park, CA 94025 (US).

(72) Inventors; and

(75) Inventors/Applicants (for US only): **PATTERSON, Martin** [GB/US]; 332 Arbor Boulevard, Menlo Park, CA 94025 (US). **OBERDORFER, Matthias** [DE/US]; 1360 Willow Road, Menlo Park, CA 94025 (US).

(74) Agents: **DANIELSON, Mark, J.** et al.; Pillsbury Winthrop Shaw Pittman LLP, P.O. Box 10500, McLean, VA 22102 (US).

(81) Designated States (unless otherwise indicated, for every kind of national protection available): AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, 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, LV, LY, MA, MD, 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.

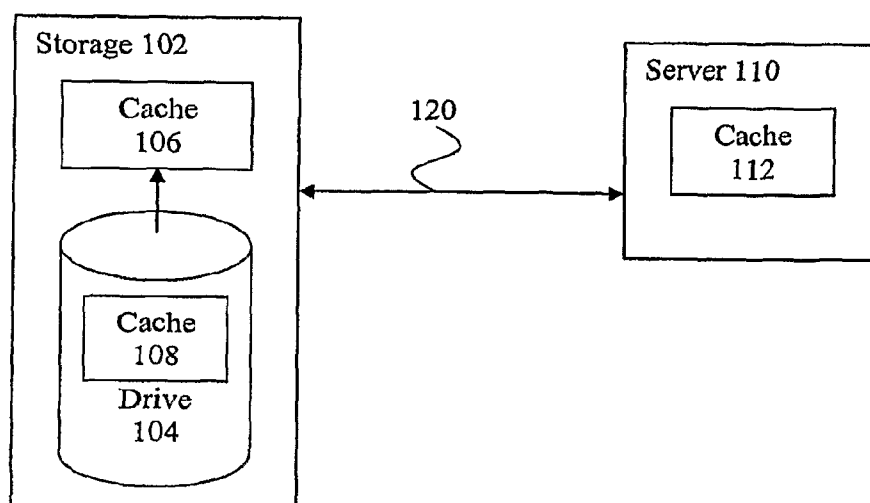
(84) Designated States (unless otherwise indicated, for every kind of regional protection available): ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, NL, 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:

— without international search report and to be republished upon receipt of that report

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

(54) Title: METHOD AND APPARATUS FOR PROVIDING HIGH-PERFORMANCE AND HIGHLY-SCALABLE STORAGE ACCELERATION



(57) Abstract: A method and apparatus of providing high performance and highly scalable storage acceleration includes a cluster node-spanning RAM disk (CRD) interposed in the data path between a storage server and a computer server. The CRD addresses performance problems with applications that need to access large amounts of data and are negatively impacted by the latency of classic disk-based storage systems. It solves this problem by placing the data the application needs into a large (with respect to the server's main memory) RAM-based cache where it can be accessed with extremely low latency, hence improving the performance of the application significantly. The CRD is implemented using a novel architecture which has very significant cost and performance advantages over existing or alternative solutions.

WO 2007/100521 A2

METHOD AND APPARATUS FOR PROVIDING HIGH-PERFORMANCE AND HIGHLY-SCALABLE STORAGE ACCELERATION

FIELD OF THE INVENTION

The present invention relates to data storage, and more particularly to an architecture and approach for a high performance, highly scalable storage accelerator for computer networks.

BACKGROUND OF THE INVENTION

In computing architectures that use externally attached storage such as Network Attached Storage (NAS) or Storage Area Networks (SANs), there is a growing mismatch between the increasing speed of compute servers and the ability of these storage systems to deliver data in a timely fashion. This inability for storage systems to keep up with fast compute servers causes applications to stall and overall throughput of the system to plateau or even regress under significant load.

When looking more closely at the root causes of this scalability problem, one common factor is latency of fetching data from a disk drive, in particular the rotation and seek time. While drives can deliver large contiguous amounts of data with an initial latency of 1-5ms in seek time (moving the drive heads to the correct location on disk) frequent access to non-contiguous data can be of the order of ~40ms per access. For datasets that involve a lot of randomly accessed data (such as relational databases), the drive seek time becomes a major bottleneck in delivering data in a timely fashion.

Traditional attempts to solve this problem include adding a hierarchy of RAM-based data caches in the data path. This conventional approach is illustrated in FIG. 1. As shown in FIG. 1, when a compute server 110 attempts to access data from storage system 102 via a network 120, there are typically at least three different caches in the overall data path. A hard drive data cache 108 provides about 8Mbytes of cache, a storage system cache 106 provides between about 128Mbytes and 16Gbytes, and a compute server data cache 112 provides between about 100M and 2Gbytes (typical lightly loaded system).

While such caches are generally beneficial, certain drawbacks remain. For example, the performance problems mentioned above still occur when the active data set is being accessed randomly or is too large to fit into the caches normally present.

There have been a number of companies that have created caching products which try to attack this problem through custom hardware solutions. Examples of this include

RAMSAN from Texas Memory Systems (<http://www.superssd.com/default.asp>) and e and n-series products from Solid Data (<http://www.soliddata.com/>). These products are inadequate because they rely on solid-state disk technology which tends to be both expensive and limited in maximum storage size.

SUMMARY OF THE INVENTION

The present invention is related to providing a high performance, highly scalable and cost-effective storage accelerator. According to one aspect, an additional extremely large and scaleable RAM-based cache is introduced into the storage hierarchy between the Storage Array/NAS Filer and the compute servers. While external caching devices are not new and several established products exist, the architecture and approach of the present invention are unique. In one example, the system architecture is based on a parallel computing cluster design which yields an extraordinarily large scaleable cache at a very attractive price point.

In furtherance of these and other aspects, an apparatus that accelerates an access between a storage server and a client over a network according to the invention includes a plurality of computing elements each having an available portion of system memory, a memory pool being comprised of the combined available system memory of the computing elements, and programs respectively executing on the computing elements that cause the access to be intercepted and determine whether data corresponding to the access should be provided from the memory pool rather than the storage server.

In additional furtherance of these and other aspects, a storage system according to the invention comprises a scaleable RAM-based cache system separate from, and in a network path between, a storage server and a compute server.

In additional furtherance of these and other aspects, a method of accelerating storage access, comprises providing a scaleable RAM-based cache system separate from, and in a network path between, a storage server and a compute server.

BRIEF DESCRIPTION OF THE DRAWINGS

These and other aspects and features of the present invention will become apparent to those ordinarily skilled in the art upon review of the following description of specific embodiments of the invention in conjunction with the accompanying figures, wherein:

FIG. 1 is a block diagram of a conventional storage hierarchy;

FIG. 2 is a block diagram illustrating an example storage hierarchy in accordance with the invention;

FIGs. 3A to 3C illustrate a high-level read flow;

FIGs. 4A to 4C illustrate a high-level write flow;

FIG. 5 is a block diagram illustrating an example implementation of a cluster RAM disk according to the invention; and

FIGs. 6A to 6D illustrate an example operation of providing storage acceleration by distributing data across multiple computing elements in a cluster RAM disk according to the invention.

DETAILED DESCRIPTION OF THE PREFERRED EMBODIMENTS

The present invention will now be described in detail with reference to the drawings, which are provided as illustrative examples of the invention so as to enable those skilled in the art to practice the invention. Notably, the figures and examples below are not meant to limit the scope of the present invention to a single embodiment, but other embodiments are possible by way of interchange of some or all of the described or illustrated elements. Moreover, where certain elements of the present invention can be partially or fully implemented using known components, only those portions of such known components that are necessary for an understanding of the present invention will be described, and detailed descriptions of other portions of such known components will be omitted so as not to obscure the invention. In the present specification, an embodiment showing a singular component should not be considered limiting; rather, the invention is intended to encompass other embodiments including a plurality of the same component, and vice-versa, unless explicitly stated otherwise herein. Moreover, applicants do not intend for any term in the specification or claims to be ascribed an uncommon or special meaning unless explicitly set forth as such. Further, the present invention encompasses present and future known equivalents to the known components referred to herein by way of illustration.

In general, the present invention greatly improves storage scalability by introducing an additional extremely large and scaleable RAM-based cache into the storage hierarchy between the Storage Array/NAS Filer and the compute servers.

A block diagram illustrating an example architecture in accordance with some aspects of the invention is shown in FIG. 2. As shown in FIG. 2, a cluster RAM disk (CRD) 230 is provided in the data path between one or more storage servers 202-1 to 202-n and one or

more compute servers 210-1 to 210-s. The data path is formed by the portion of a network 220 that links the stores and servers. It should be noted, however, that although network elements and protocols can be used to implement the data path, certain of servers 202 and/or servers 210 need not be remotely located from either each other or CRD 230.

Storage servers 202 are in one example NAS filers such as the FAS 900 series from Network Appliance. In another example storage servers 202 are comprised of SAN products such as the Symmetrix DMX series from EMC Corporation. It should be noted that although a plurality of servers are shown, there may only be one. It should be further noted that servers 202 may comprise combinations of different types of servers (e.g. some of both NAS and SAN).

Compute servers 210 are typically high-performance servers running OLTP, batch processing, and other intensive operations under Apple OS X, AIX, Solaris, Linux, and Windows environments, for example. One example of compute servers 210 that can be used include the DL145 from Hewlett Packard.

Network 220 is an Ethernet in a NAS example, or a Fibre Channel in a SAN example. Hardware from Fibre Channel vendors including Cisco, Emulex, Brocade, McData, QLogic, LSI Logic, and Vixel can be used in a Fibre Channel example. Other types of wired and wireless networks and connections, as well as combinations of disparate types of networks, are possible.

CRD 230, as will be described in more detail below, preferably contains a large pool of RAM or other type of semiconductor memory, and contains functionality for recognizing and fulfilling requests for reading and writing data between stores 202 and servers 210.

According to one aspect of the invention illustrated in FIG. 2, a single CRD 230 can be interposed in the data path of many storage servers 202, and so its cache can be effectively be shared between these servers. Conventional designs that rely on cache inside the file server effectively "strand" it. For example, one file server may be completely saturated (with all its cache resources in use) whereas a neighboring file server may be idle with plenty of cache resources. As there is no way to share cache between file servers, performance suffers. Moreover, file server cache memory can be very expensive in large quantities, whereas the same amount of other types of RAM are generally cheaper, and are used in CRD 230 in accordance with one aspect of the invention.

FIGs. 3A to 3C illustrate the advantages of the invention in connection with a high level read flow. FIG. 3A illustrates a conventional flow where compute servers 210 access

data from storage servers 202 using an NFS protocol, servers 210 singly or together implement one or more NFS client processes 302 and servers 202 include one or more NFS server processes 304. NFS is a well-known protocol in the art of network storage. Other protocols that can be used in accordance with the invention include CIFS and Fibre Channel. As shown in FIG. 3A, clients 302 use one or more of the RPC calls found in NFS (e.g. READ) to access data in servers 202 via NFS server process 304. These calls use a direct TCP or UDP connection between NFS client processes 302 and server process 304, as well as the existing caches discussed in connection with FIG. 1.

FIG. 3B illustrates a read flow when CRD 230 according to the invention is provided in the data path between storage servers 202 and compute servers 210. As shown in FIG. 3B, a read request made by NFS client process 302 is intercepted by CRD 230, and the corresponding response data by NFS server process 304 is cached. Accordingly, as shown in FIG. 3C, when a subsequent read request for the same data is made by one of NFS client processes 302, the cached data from CRD 230 can be returned much more quickly than would be possible without CRD 230.

FIGs. 4A to 4C illustrate the advantages of the invention in connection with a high level write flow. FIG. 4A illustrates a conventional flow where compute servers 210 write data to storage servers 202 using an NFS protocol, wherein clients 302 use one or more of the RPC calls found in NFS (e.g. WRITE) to write data in storage servers 202 via NFS server process 304, as well as a direct TCP or UDP connection between NFS client processes 302 and server process 304 and the existing caches discussed in connection with FIG. 1.

FIG. 4B illustrates a write flow when CRD 230 according to the invention is provided in the data path between storage servers 202 and compute servers 210. As shown in FIG. 4B, a write request made by NFS client process 302 is intercepted by CRD 230, and the corresponding data provided by NFS client process 302 is cached. Accordingly, as shown in FIG. 4C, when a subsequent read request for the same data is made by one of NFS client processes 302, the cached data from CRD 230 can be returned much more quickly than would be possible without CRD 230.

Further details of how a CRD 230 can be implemented to intercept and process read and write flows in various example embodiments will be provided below.

One example implementation of CRD 230 is further illustrated in FIG. 5. As shown in FIG. 5, in this example, CRD 230 is implemented as a cluster of computing elements or blades 502-1 to 502-n. The elements are interconnected using a standard interconnect 516,

for example Gigabit Ethernet, which is used to exchange management and control information between blades 502. As shown in FIG. 5, interconnect 516 can further connect the cluster of blades 502-1 to 502-n to additional clusters or blades (not shown), which may or may not be further used to implement CRD 230.

In one example, elements 502 are comprised of 64-bit blade servers such as ScaleOut series blades from Rackable Systems, each having 16 GBytes of system memory (e.g. RAM) 504. The memories of the individual blades are combined together in a manner to be described in more detail below to form a single very large and scalable memory pool 512. According to an aspect of the invention, therefore, the system offers an exceptionally large cache at an attractive price compared to conventional approaches which require expensive custom hardware designs which make the cost of a large cache (>320GBytes at the time of writing) prohibitive. It should be noted that elements 502 or clusters of elements need not be implemented using exactly the same type of blade server or other computing element, as long as they are capable of supporting an application 506 as described below.

As further shown in FIG. 5, each blade 502 includes a custom application 506 that executes on each respective blade 502 and communicates with other applications 506 via interconnect 516 to thereby preferably implement a proxy 508 and a global directory 510. In one example implementation, the application 506 executes in an embedded Linux or VxWorks environment.

In one example, proxy 508 implements well-known virtual IP addresses and proxy server techniques to intercept and, if possible, fulfill data requests from clients to servers. However, according to an aspect of the invention as will be described in more detail below, the physical blade 502 that handles any given request (or any portion thereof) is determined by a connection handoff mechanism. According to another aspect, proxy 508 includes support for one or more industry standard storage protocols (such as NFS, CIFS, Fibre Channel) and is implemented as a "bump in the wire" tee. Proxy 508 also handles communications from servers back to clients.

Global directory 510 maps the cached files to the physical blade(s) 502 on which they reside. According to an aspect of the invention, a single file can be distributed across multiple blades 502. To be able to locate file content system wide by means of a Global Directory system (i.e. how one blade can determine which other blade has the content it needs to fulfill a particular request), application 506 preferably includes an efficient update

mechanism where all blades see a consistent view of the Global Directory, and those skilled in the art will understand various ways to implement such an update mechanism.

Memory pool 512 is comprised of the aggregate of free system memory in RAM 504 of each of blades 502 in one example implementation. According to an aspect of the invention, this allows the total size of the memory pool 512 to be scalable depending on the number of blades 502 that are added or removed from CRD 230, either physically or through any number of software or hardware configuration methods known to those skilled in the art. In this regard, applications 506 preferably communicate with each other to determine the number of blades 502 that are available at any given moment, thus allowing the memory pool to be freely scaled.

As each blade 502 operates as an independent system, one or more of applications 506 further implement algorithms to decide on how best to distribute file content which is too big to fit in a single blade's memory across multiple blades. In one example, the algorithms consider: Placement of file data onto cluster blades based on a static round-robin algorithm, and/or Placement of data onto the next "least loaded blade" with a computed metric of the load factor of each blade. This computed metric can be a combination of one or more of: Total amount of memory in use; Input/Output rate of the blade; Cache activity – how often Cache Reclaim operations are taking place; and other considerations.

In one example, one of blades 502 is designated a default blade for initially intercepting and, if further processing is required, opening a socket in its corresponding application 506 to handle the request. A TCP connection handoff mechanism is then used to allow cloned sockets to be opened by applications 506 in other blades 502 to service a client request for data which it has in its own physical memory.

For example, application 506 in all blades 502 share a single virtual IP address for use of the CRD as a proxy server, and clients 302 are configured to send data requests destined for one of servers 304 to this IP address. The proxy 508 distributed across all blades 502 monitors the requested connection between the specific client 302 and server 304 associated with this connection. When one of blades 502 starts communicating with a client 302 using the virtual IP address, and it is determined that another blade 502 should handle communications (e.g. when the other blade 502 contains data in its physical memory corresponding to a data request from the client), that blade takes over the network connection (e.g. by transparently migrating the connected TCP endpoint from one blade 502 to another blade 502 without interaction on behalf of the client). This allows the other blade to directly

deliver its data into the network stream. It should be noted that, in this example, applications 506 may communicate among themselves to determine the default blade at any given point in time.

According to one aspect, CRD 230 also provides for parallel data delivery. This means that many blades can operate in parallel delivering data to clients. An example might be when a single large file is distributed across all the blades in the system, each blade in turn can deliver its portion of the file to clients requesting data. This compares with a single file server where such requests are typically serialized and hence delivered one-by-one.

In some example implementations, application 506 can include a high speed lossless compress/decompress function which doubles or triples the effective size of the memory pool 512 versus the combined size of physically present RAM 504 on the blades. Many compress/decompress functions that are known in the art can be used, such as Lempel-Ziv or Run Length Encoding. Application 506 in other example implementations can further include a high speed checksum function which acts as an integrity check for data going into the memory pool 512 and coming out of it. Application 506 can still further include efficient and scalable file and block level lookup functions.

Further aspects of a CRD as described above and in accordance with the principles of the invention will become apparent from an example read operation illustrated in FIGs. 6A to 6D. In this example, a CRD according to the invention includes four elements 502 designated as blades 0 to 3 and client/server proxy 508 (implemented as a NFS proxy in this example) and global directory 510 allow pool 512 to be used as a large and high-speed cache for data stored by servers 304 and requested by clients 302. In the example of FIGs. 6A to 6D, storage is provided by a NAS server that can be accessed by a client using NFS over a TCP/UDP Ethernet transport. Those skilled in the art will understand alternative applications and environments after being taught by this example.

As shown in FIG. 6A, a NAS client (executing in a compute server 210 for example) attempts to access File F in a NAS server (executing in a store 202 for example) over a network (e.g. network 220). The request is intercepted by the NFS Proxy which has a virtual IP address that can be shared among blades 0 to 3, and is configured to either process or pass through requests to the real IP address of the NAS server. A socket for handling the request is opened by application 506 in one of the blades 502 of the CRD, designated blade 0, which in one example is configured as a default blade for all requests. If there is no entry for file F

in the Global Directory, the request will be forwarded to the NAS server without further processing by the NFS Proxy.

As further shown in FIG. 6B, as the NAS client begins an NFS read operation for file F, data from NAS server is cached in a portion of the CRD Memory Pool that is physically located in blade 0 in addition to being returned to the NAS client via the socket. An entry for file F is created in the Global Directory.

As more file F data is requested by the NAS client in FIG. 6C, the CRD determines that the Memory Pool should be extended to include physical memory in blade 1. Accordingly, the socket is cloned to node 1 and file F data provided by the NAS server is cached in node 1. The connection between the CRD and the client is handed off to blade 1 and the Global Directory is further updated to reflect the range of file F data cached in blade 1.

FIG. 6D further illustrates the progression of the Memory Pool extension to physical memory in blades 2 and 3 as more and more file F data is requested by the NAS client and returned by the NAS server. FIG. 6D further illustrates how the Global Directory is correspondingly updated.

Although the present invention has been particularly described with reference to the preferred embodiments thereof, it should be readily apparent to those of ordinary skill in the art that changes and modifications in the form and details may be made without departing from the spirit and scope of the invention. It is intended that the appended claims encompass such changes and modifications.

WHAT IS CLAIMED IS:

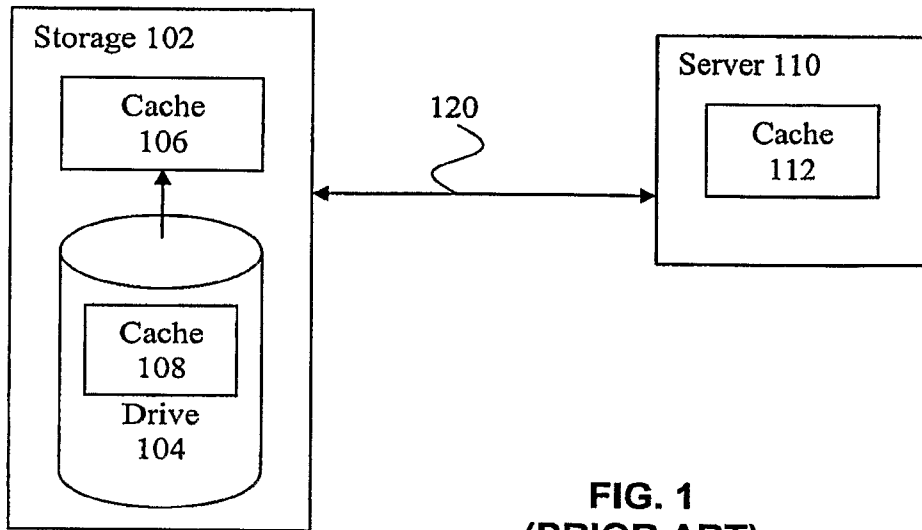
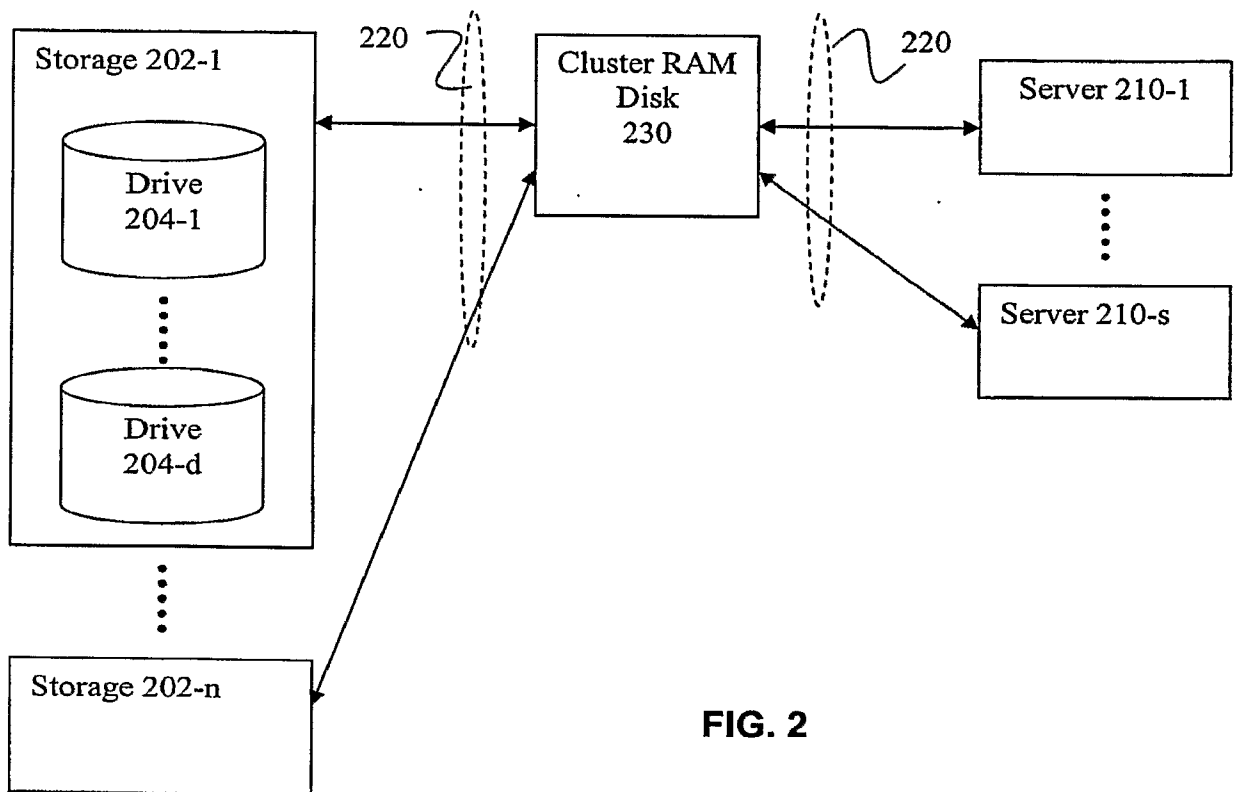
1. An apparatus that accelerates an access between a storage server and a client over a network, the apparatus comprising:
a plurality of computing elements each having an available portion of system memory, a memory pool being comprised of the combined available system memory of the computing elements; and
programs respectively executing on the computing elements that cause the access to be intercepted and determine whether data corresponding to the access should be provided from the memory pool rather than the storage server.
2. An apparatus according to claim 1, wherein the programs implement a proxy that causes communications corresponding to the access from the client addressed to the storage server to be intercepted and handled by a certain one of the computing elements.
3. An apparatus according to claim 2, wherein the proxy includes a connection handoff mechanism that causes the communications to be transparently switched to another one of the computing elements .
4. An apparatus according to claim 3, wherein the connection handoff mechanism includes a mechanism for transparently migrating a TCP endpoint corresponding to the communications from the certain computing element to the another computing element.
5. An apparatus according to claim 1, wherein the program implements a global directory indicating where content is located in the system memories of the computing elements.
6. An apparatus according to claim 5, wherein content corresponding to one data file can be spread across multiple ones of the computing elements.
7. An apparatus according to claim 1, wherein the computing elements comprise blade servers.

8. An apparatus according to claim 1, further comprising an interconnect linking the computing elements and through which the programs communicate.
9. An apparatus according to claim 8, wherein the interconnect comprises Gigabit Ethernet.
10. An apparatus according to claim 1, wherein a size of the memory pool is scalable in accordance with a variable number of the computing elements being included in the apparatus.
11. An apparatus according to claim 1, wherein a size of the memory pool is at least 300 GB.
12. An apparatus according to claim 11, wherein the system memory is comprised of RAM.
13. An apparatus according to claim 1, wherein the system memory is comprised of RAM.
14. An apparatus according to claim 1, wherein the storage access is in accordance with NFS protocol.
15. An apparatus according to claim 1, wherein a second access between the storage server and a second client occurs simultaneously with the access, and wherein the programs cause a first one of the computing elements to handle the access and a second one of the computing elements to handle the second access in parallel.
16. An apparatus according to claim 1, wherein the programs cause other accesses to a plurality of other storage servers in addition to the storage server to be intercepted.
17. An apparatus according to claim 1, wherein the network is Ethernet.
18. An apparatus according to claim 1, wherein the network is Fibre Channel.

19. A storage system, comprising:
a scaleable RAM-based cache system separate from, and in a network path between, a storage server and a compute server.
20. A storage system according to claim 19, wherein the cache system includes a proxy for intercepting storage accesses between the storage server and the compute server.
21. A storage system according to claim 19, wherein the cache system comprises a configurable plurality of computing elements.
22. A storage system according to claim 21, wherein the computing elements comprise blade servers.
23. A storage system according to claim 21, wherein the computing elements are interconnected by a local area network.
24. A storage system according to claim 21, wherein the local area network comprises Gigabit Ethernet.
25. A storage system according to claim 21, wherein each of the computing elements includes an available amount of system memory, a memory pool for the cache system being comprised of the combined available system memory of all of the computing elements.
26. A method of accelerating storage access, comprising:
providing a scaleable RAM-based cache system separate from, and in a network path between, a storage server and a compute server.
27. A method according to claim 26, wherein the providing step includes providing a proxy for intercepting storage accesses between the storage server and the compute server.
28. A method according to claim 26, wherein the providing step includes:
implementing the cache system with a configurable plurality of computing elements.

29. A method according to claim 28, wherein the computing elements comprise blade servers.
30. A method according to claim 28, further comprising:
interconnecting the computing elements by a local area network.
31. A method according to claim 30, wherein the local area network comprises Gigabit Ethernet.
32. A method according to claim 28, wherein each of the computing elements includes an available amount of system memory, the method further comprising:
forming a memory pool for the cache system from the combined available system memory of all of the computing elements.

1/5

**FIG. 1**
(PRIOR ART)**FIG. 2**

2/5

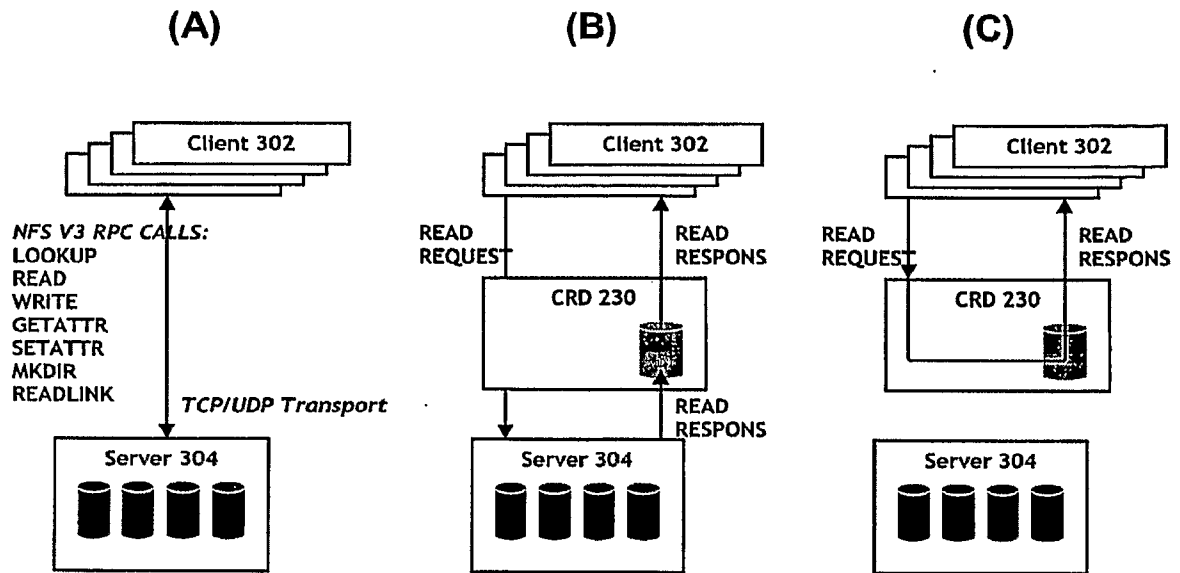


FIG. 3

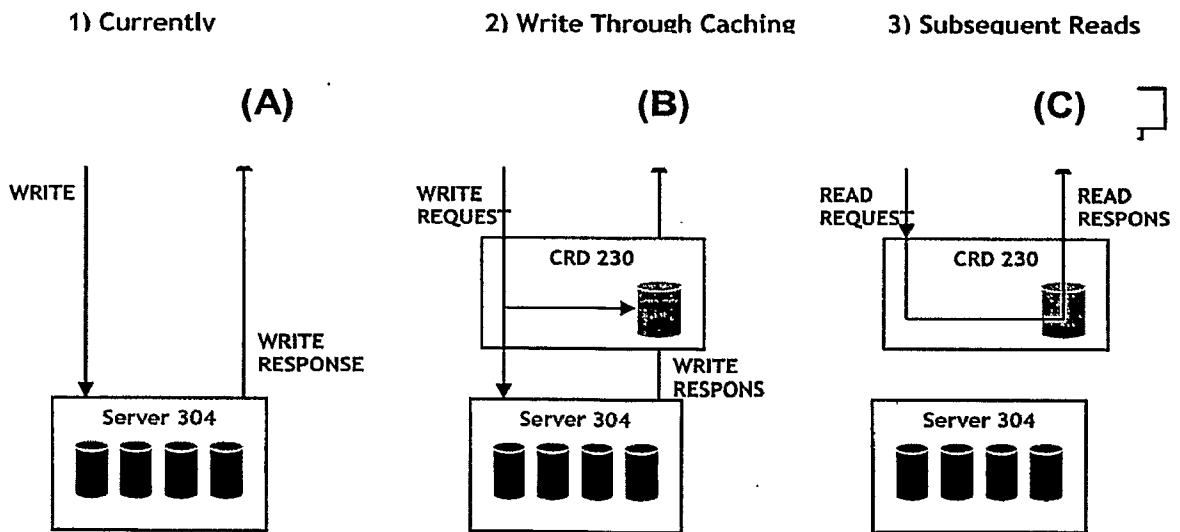


FIG. 4

3/5

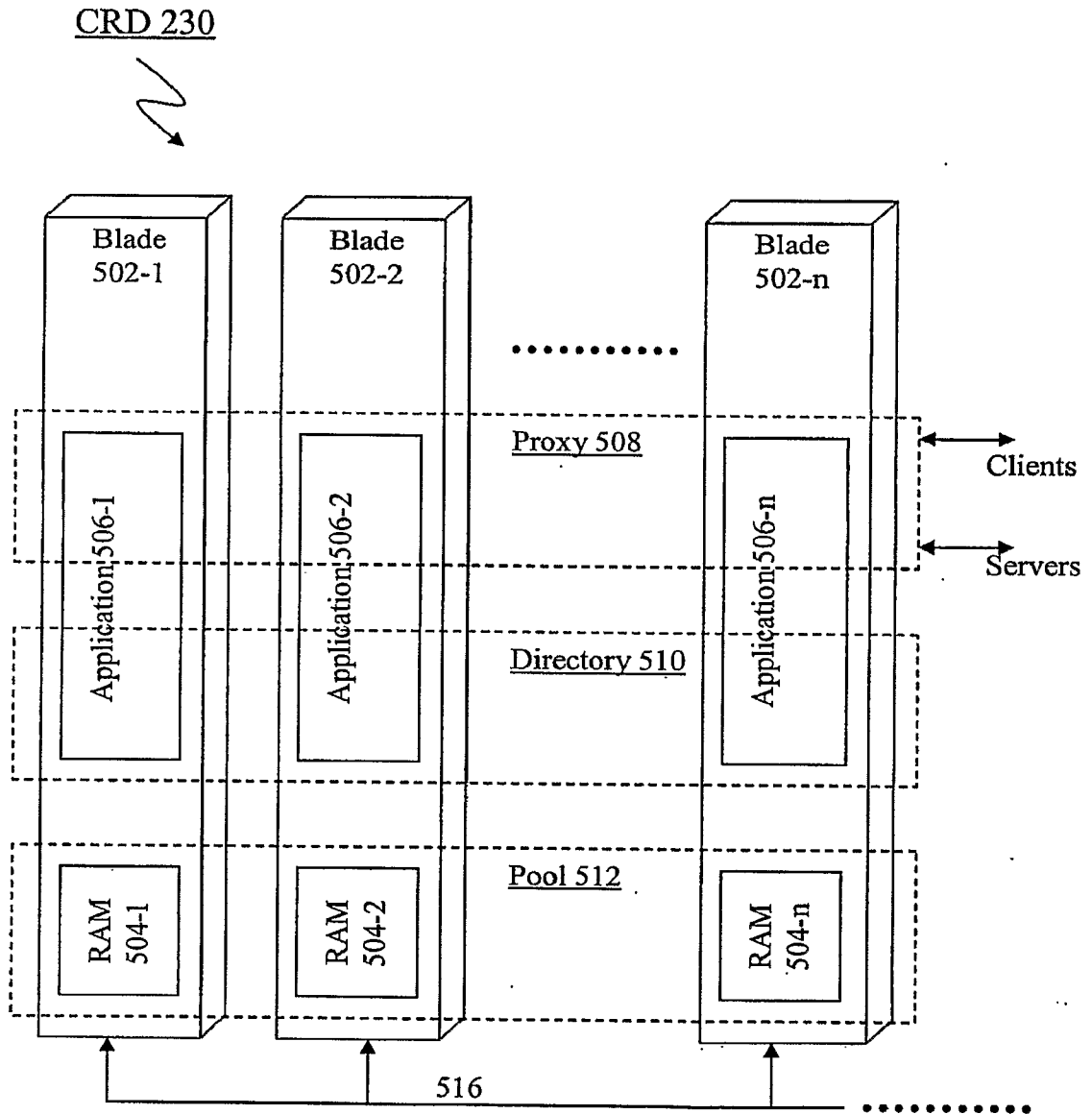


FIG. 5

4/5

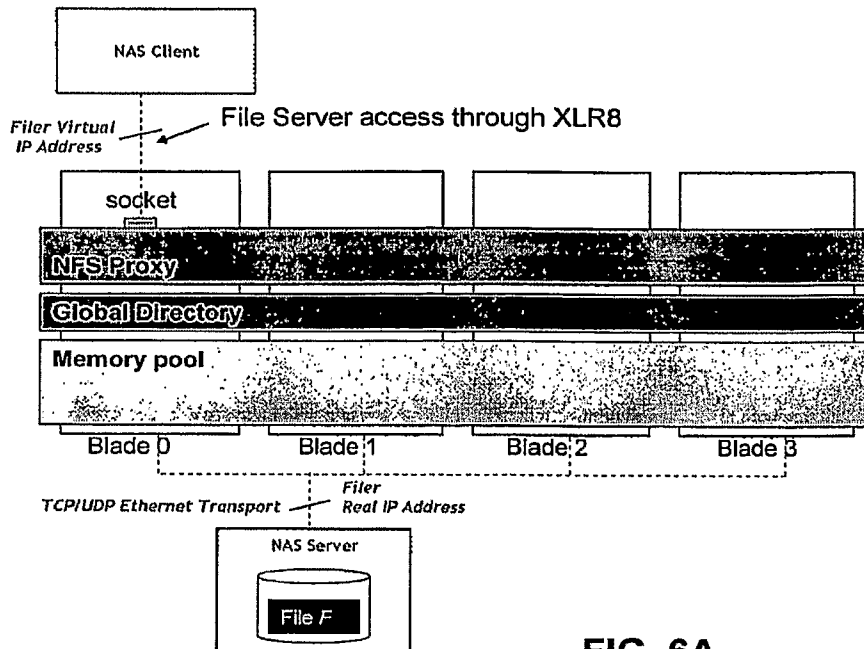


FIG. 6A

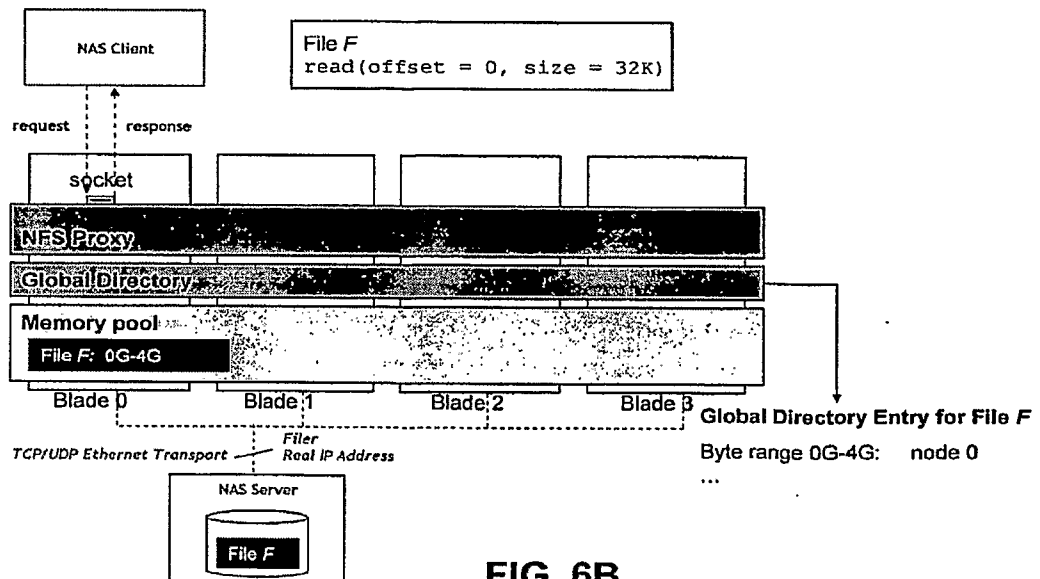


FIG. 6B

5/5

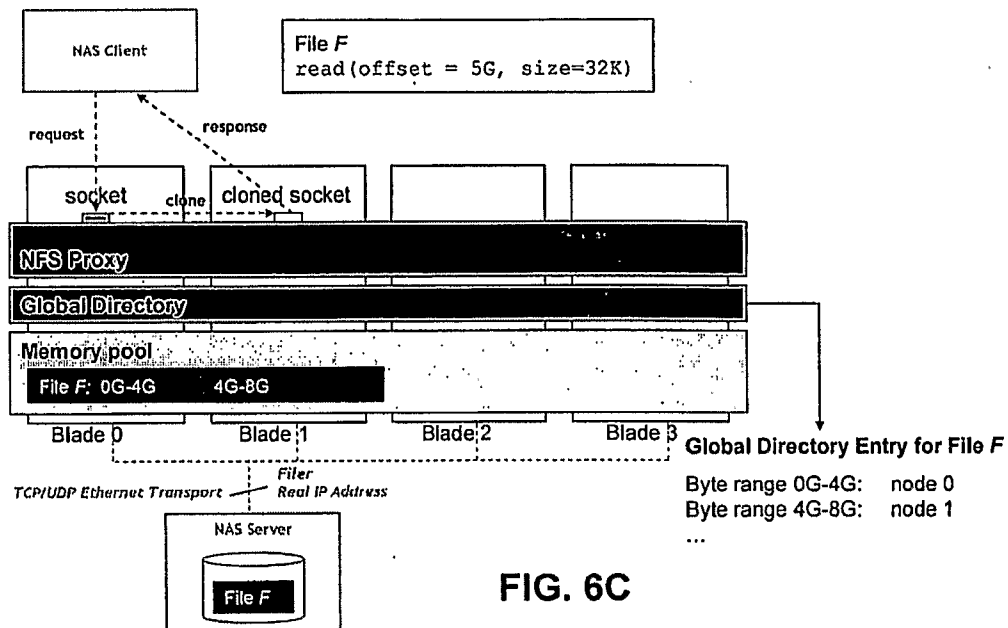


FIG. 6C

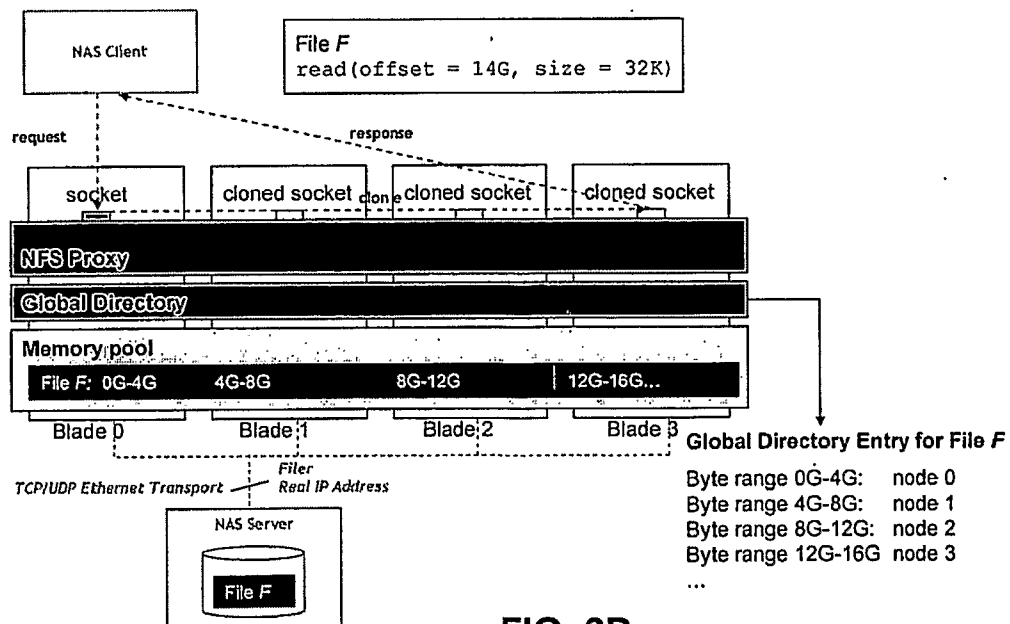


FIG. 6D