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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 September 2007 (27.09.2007)

PCT

(10) International Publication Number
WO 2007/108840 A1

(51) International Patent Classification:
G06F 11/14 (2006.01)

(21) International Application Number:
PCT/US2006/045552

(22) International Filing Date:
28 November 2006 (28.11.2006)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
11/378,722 17 March 2006 (17.03.2006) US

(71) Applicant *(for all designated States except US)*: **EMC CORPORATION** [US/US]; 176 South Street, Hopkinton, MA 01748 (US).

(72) Inventor; and
(75) Inventor/Applicant *(for US only)*: **LONG, Matthew** [US/US]; 538 Mendon Street, Uxbridge, MA 01569 (US).

(74) Agent: **DUQUETTE, Jeffrey, J.**; Highpoint Center, 2 Connector Road, Suite 200, Westborough, MA 01581 (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.

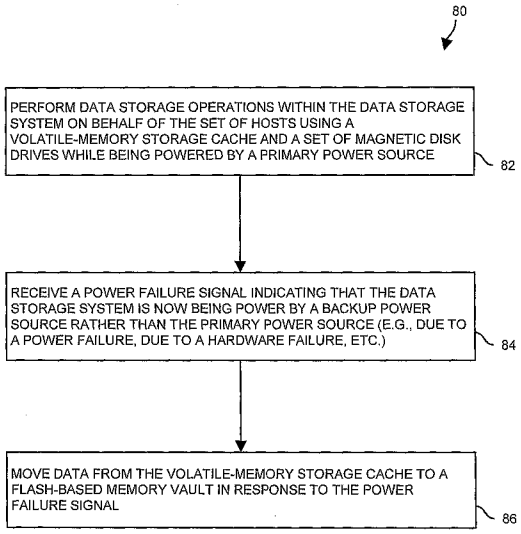
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT,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

Declarations under Rule 4.17:
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*
— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

Published:
— *with international search 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:** TECHNIQUES FOR MANAGING DATA WITHIN A DATA STORAGE SYSTEM UTILIZING A FLASH-BASED MEMORY VAULT



(57) **Abstract:** A technique for managing data within a data storage system involves performing data storage operations on behalf of a set of hosts (i.e., one or more hosts) using a volatile-memory storage cache and a set of magnetic disk drives while the data storage system is being powered by a primary power source (e.g., a main power feed). The technique further involves receiving a power failure signal (e.g., from a sensor) indicating that the data storage system is now being powered by a backup power source rather than by the primary power source (e.g., due to a loss of the main power feed, due to a failure of a power converter, etc.), and moving data from the volatile-memory storage cache of the data storage system to a flash-based memory vault of the data storage system in response to the power failure signal.

WO 2007/108840 A1

TECHNIQUES FOR MANAGING DATA WITHIN A DATA STORAGE SYSTEM UTILIZING A FLASH-BASED MEMORY VAULT

BACKGROUND

One conventional data storage system includes a storage processor, an array of magnetic disk drives and a backup power supply. The storage processor carries out a variety of data storage operations on behalf of an external host device (or simply host). In particular, the storage processor temporarily caches host data within its storage cache and, at certain times, de-stages that cached data onto the array of magnetic disk drives. If the data storage system is set up so that it acknowledges write requests from the host once the data reaches the storage cache rather than once the data reaches the array of magnetic disk drives, the host will enjoy shorter transaction latency.

Some data storage systems employ backup power supplies (e.g., uninterruptible power supplies) to prevent the loss of data from the storage caches in the event of power failures. For example, suppose that such a data storage system fails to receive power from a main power feed (e.g., power from the street) during operation. In such a situation, a set of backup power supplies provides reserve power to the storage processor and to the array of magnetic disk drives for a short period of time (e.g., 30 seconds). During this time, the storage processor writes the data from its storage cache onto a dedicated section of the magnetic disk drives called a "vault" so that any data which has not yet been properly de-staged is not lost. Once power from the main power feed returns, the storage processor loads the data from the magnetic disk drive vault back into the storage cache. At this point, the data storage system is capable of continuing normal operation.

It should be understood that some data storage systems include two storage processors for high availability (e.g., fault tolerant redundancy, higher throughput, etc.). Furthermore, some data storage systems position arrays of magnetic disk drives within enclosures which are separated from other enclosures holding the storage processors. These data storage systems typically rely on an external backup power supply for each storage processor and the array of magnetic disk drives that contain

the vault. Typically the backup power supplies for the storage processors and the magnetic disk drives communicate with the various components of the data storage system through external cables in order to properly coordinate their operations.

SUMMARY

Unfortunately, there are deficiencies to the above-described conventional data storage systems which store data from storage caches to magnetic disk drive vaults during power failures. For example, magnetic disk drives typically consume a significant amount of power even during a short time duration (e.g., 30 seconds) since power is required for disk drive motors to spin, for fans to provide cooling, for actuators to move magnetic heads, and so on. Accordingly, the backup power supplies for arrays of magnetic disk drives are often large, costly and complex.

Additionally, the backup power supplies are external to the storage processor and disk array enclosures and as such require power and control cabling between the backup power supplies and the various enclosures. These external backup power supplies and the associated cabling impose relatively-high serviceability demands as well as increase the number of components which are susceptible to failure.

In contrast to the above-described conventional approaches to storing data from storage caches into magnetic disk drive vaults during power failures, an improved technique involves moving data within a data storage system from a storage cache into a flash-based memory vault (e.g., a module containing flash memory with no mechanical moving parts) in response to a power failure signal. Such operation alleviates the need to provide backup power to magnetic disk drives. Rather, data can be moved from the storage cache to the flash-based memory vault using a relatively-small backup power source (e.g., a battery that only powers a storage processor). Without the need for backup power to the magnetic disk drives, there is no burden of having to provide large, costly and complex backup power supplies and the associated external cabling for magnetic disk drives. That is, the magnetic disk drives can simply turn off as soon as primary power is lost. With the storage processor still running from a backup power source (e.g., a relatively small battery), the storage processor is capable of moving the contents of the storage cache

to the flash-based memory vault thus preserving data integrity of the data storage system so that no data is ever lost.

One embodiment is directed to a technique for managing data within a data storage system. The technique involves performing data storage operations on behalf of a set of hosts (i.e., one or more hosts) using a volatile-memory storage cache and a set of magnetic disk drives while the data storage system is being powered by a primary power source (e.g., a main power feed). The technique further involves receiving a power failure signal (e.g., from a sensor, from a backup power source, etc.) indicating that the data storage system is now being powered by a backup power source rather than by the primary power source (e.g., due to a loss of the main power feed, due to a failure of a power converter, etc.), and moving data from the volatile-memory storage cache of the data storage system to a flash-based memory vault of the data storage system in response to the power failure signal.

BRIEF DESCRIPTION OF THE DRAWINGS

The foregoing and other objects, features and advantages of the invention will be apparent from the following description of particular embodiments of the invention, as illustrated in the accompanying drawings in which like reference characters refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead being placed upon illustrating the principles of the invention.

Fig. 1 is a block diagram of a data storage system which utilizes a flash-based memory vault.

Fig. 2 is a diagram of the data storage system of Fig. 1 in a multiple storage processor context.

Fig. 3 is a flowchart of a procedure performed by the data storage system of Figs. 1 and 2.

Fig. 4 is a block diagram illustrating a particular use of the flash-based memory vault of Fig. 2.

Fig. 5 is a block diagram of a first technique for restoring contents of the flash-based memory vault of Fig. 2 to a storage cache.

Fig. 6 is a block diagram of a second technique for restoring contents of the flash-based memory vault of Fig. 2 to a storage cache.

Fig. 7 is a block diagram of a third technique for restoring contents of the flash-based memory vault of Fig. 2 to a storage cache.

DETAILED DESCRIPTION

An improved technique involves moving data within a data storage system from a storage cache into a flash-based memory vault in response to a power failure signal. Such operation alleviates the need to provide backup power to magnetic disk drives. Rather, data can be moved from the storage cache to the flash-based memory vault using a relatively-small backup power source, e.g., a battery that only powers a storage processor. Without the need for backup power to the magnetic disk drives, there is no burden of having to provide large, costly and complex backup power supplies and the associated external cabling for magnetic disk drives. That is, the magnetic disk drives can simply turn off as soon as the primary power source is lost. With the storage processor still running from a backup power source (e.g., a dedicated battery), the storage processor is capable of moving the contents of the storage cache to the flash-based memory vault thus preserving data integrity of the data storage system so that no data is ever lost.

Fig. 1 shows a data storage system 20 which is configured to manage data behalf of a set of hosts 22(1), 22(2), ... (collectively, hosts 22). In particular, the data storage system 20 exchanges communications signals 24 with at least one host 22 to perform a variety of data storage operations (e.g., read, write, read-modify-write, etc.).

As shown in Fig. 1, the data storage system 20 includes a primary power source 26, a secondary power source 28, storage processing circuitry 30 and a set of magnetic disk drives 32 (i.e., one or more magnetic disk drives 32). The primary power source 26 (e.g., a set of power supplies which connects to an external main power feed) is configured to provide primary power 34 to the storage processing circuitry 30 under normal conditions. The secondary power source 28 (e.g., a set of

batteries) is configured to provide backup power 36 to the storage processing circuitry 30 in the event of a loss of primary power 34.

As further shown in Fig. 1, the storage processing circuitry 30 is configured to receive a power failure signal 38 which indicates whether the storage processing circuitry 30 is running off of primary power 34 or backup power 36. In some arrangements, the power failure signal 38 is a power supply signal from the primary power source 26 or from the secondary power source 28. In other arrangements, the power failure signal 38 is a separate signal, e.g., from a sensor connected to the main power feed.

The storage processing circuitry 30 includes a controller 40, a volatile-memory storage cache 42 (a data storage cache between 100 MB to 1 GB), a flash-based memory vault 44, a clock generator circuit 46, and isolation circuitry 48. While the controller 40 is being powered by the primary power source 28, the controller 40 performs data storage operations on behalf of the set of hosts 22 using the volatile-memory storage cache 42 and the set of magnetic disk drives 32. For example, when a host 22 sends the controller 40 a request to write data, the controller 40 stores the data in volatile memory 42 and then, in parallel to scheduling the data to be written to the magnetic disk drives 32, conveys the completion of the write data request to the host 22. As a result, the write request completes to the host 22 as soon as the data is written to the volatile-memory storage cache 42 which takes less time than writing the magnetic disk drives 32.

Now, suppose that the controller 40 receives the power failure signal 38 indicating that the controller 40 is now being powered by the secondary power source 28 rather than by the primary power source 26. In this situation, primary power 34 from the primary power source 26 is no longer available but backup power 36 from the secondary power source 28 is available at least temporarily. Accordingly, the controller 40 remains operational and moves data from the volatile-memory storage cache 42 to the flash-based memory vault 44 in response to the power failure signal 38. The amount of power necessary to move the data from the volatile-memory storage cache 42 to the flash-based memory vault 44 is significantly less than that which would be required to write that data out to a vault on the set of

magnetic disk drives 32 since flash-based memory (which has no motors or actuators to operate) requires relatively little power to store data.

When the primary power source 26 becomes available again, the storage processing circuitry 30 receives primary power 34 and no longer receives the power failure signal 38. In some arrangements, the omission of the power failure signal 38 (or the de-asserted state of the power failure signal 38) is essentially a power normal signal indicating that the storage processing circuitry 30 is running off of primary power 34. At this point, the controller 40 restores the contents of volatile-memory storage cache 42. In particular, the controller 40 moves the data from the flash-based memory vault 44 back into the volatile-memory storage cache 42 thus enabling the storage processing circuitry 30 to resume data storage operations where it left off, e.g., the storage processing circuitry 30 is now capable of properly de-staging the data in the volatile-memory storage cache 42 to the set of magnetic disk drives 32 as well as performing new data storage operations on behalf of the set of hosts 22 in a normal manner.

It should be understood that, in contrast to conventional data storage systems which store data from storage caches into magnetic disk drive storage vaults in response to power failures, there is no need to run the set of magnetic disk drives 32 of the data storage system 20. Rather, the set of magnetic disk drives 32 is allowed to deactivate in response to loss of primary power 34 from the primary power source 26 since the controller 40 transfers data from the volatile-memory storage cache 42 to the flash-based memory vault 44 for safe keeping. Thus, data within the volatile-memory storage cache 42, which has not yet been de-staged, is not lost.

It should be further understood that other components within the storage processing circuitry 30 enable enhanced operation in the event of a power failure. For example, the clock generator circuit 46 and the isolation circuitry 48 are configured to perform certain duties during a loss of primary power 34 from the primary power source 26.

In connection with the clock generator circuit 46, the clock generator 46 is configured to provide a relatively-fast clock signal (or multiple clock signals) to the processing circuitry of the controller 40 during normal operation when the controller

40 is performing data storage operations on behalf of the set of hosts 22. In some arrangements, a microprocessor of the controller 40 runs within a range of 50 to 100 Watts when operating at this normal operating clock speed.

However, if there is a loss of primary power 34, the clock generator 46 is configured to provide a significantly slower clock signal to the processing circuitry of the controller 40 while the controller 40 moves data from the volatile-memory storage cache 42 to the flash-based memory vault 44. In some arrangements, the microprocessor of the controller 40 runs at less than 30 Watts (e.g., substantially within a range of 15 to 20 Watts) when operating at this reduced clock speed. As a result, less power is consumed thus enabling the use of a smaller-sized backup power source 28 (e.g., a relatively small battery).

In connection with the isolation circuitry 48, it should be understood that various components of the data storage system 20 form a processing core 50. In some arrangements, the controller 40, the volatile-memory storage cache 42 and the flash-based memory vault 44 (perhaps among other components) form this core 50. During normal operation, primary power 34 from the primary power source 26 reaches all of the components of the data storage system 20 (e.g., the set of magnetic disk drives 32). However, during a loss of the primary power 34 and a switch to backup power 36 from the secondary power source 28, the isolation circuitry 48 is configured to electrically isolate the processing core 50 from the other areas of the data storage system 20 (e.g., the set of magnetic disk drives 32) so that only the processing core 50 receives the backup power 36. Accordingly, the backup power 36 is not wasted by unnecessarily powering the non-vital areas of the data storage system 20 and only reaches the vital areas thus enabling the controller 40 to dump the contents of the volatile-memory storage cache 42 into the flash-based memory vault 44. Such electrical isolation conserves backup power by removing interference, i.e., power consumption by circuits of the data storage system 20 which are non-essential during the loss of primary power such as the set of magnetic disk drives 32. Further details will now be provided with reference to Fig. 2.

Fig. 2 is a diagram of the data storage system 20 in the context of a dual storage processor configuration 60. Here, the data storage system 20 includes a first

storage processor 62(A), a second storage processor 62(B) and a high-speed bus 64 which interconnects the first and second storage processors 62(A), 62(B) (collectively, storage processors 62). The storage processor 62(A) includes, among other things, an enclosure 66(A) which contains a controller 40(A), a volatile-memory storage cache 42(A), and a flash-based memory vault 44(A). Within the enclosure 66(A) also resides a battery 68(A) which forms a portion of the secondary power source 28 (also see Fig. 1).

Similarly, the storage processor 62(B) includes, among other things, an enclosure 66(B) which contains a controller 40(B), a volatile-memory storage cache 42(B), and a flash-based memory vault 44(B). Within the enclosure 66(B) also resides a battery 68(B) which forms another portion of the secondary power source 28 (again, also see Fig. 1).

Each storage processor 62 sends communications 70 to the other storage processor 62 through the bus 64. In particular, each storage processor 62 is capable of providing status to the other storage processor 62 through the bus 64 (e.g., an indication of whether it is running in a normal operating mode or whether it has switched from the normal operating mode to a data vaulting mode). Additionally, the storage processors 62 exchange data through the bus 64 thus enabling the storage processors 62 to mirror the contents of the volatile-memory storage caches 42(A), 42(B). Accordingly, the volatile-memory storage caches 42(A), 42(B) can be viewed as forming the volatile-memory storage cache 42 of Fig. 1, and the flash-based memory vaults 44(A), 44(B) can be viewed as forming the flash-based memory vault 44 of Fig. 1. Further details will now be provided with reference to Fig. 3.

Fig. 3 is a flowchart of a procedure 80 for managing data within the data storage system 20 during a power failure event. In step 82, the controller 40 performs data storage operations on behalf of the set of hosts 22 using the volatile-memory storage cache 42 and the set of magnetic disk drives 44 while the data storage system 20 is being powered by the primary power source 26 (also see Fig. 1).

In step 84, the controller 40 receives the power failure signal 38 indicating that the data storage system 20 is now being powered by the backup power source 28

rather than by the primary power source 26. Accordingly, a power failure event has occurred. For example, the data storage system 20 may lose access to a main power feed (e.g., power from the street). As another example, the primary power source 26 may suffer a hardware failure.

In step 86, the controller 40 moves data from the volatile-memory storage cache 42 to the flash-based memory vault 44 in response to the power failure signal 38. In view of certain electrical behaviors of flash-memories, a significant amount of data is capable of being written to flash memory in a relatively short period of time (e.g., a data storage rate of 12MB/second).

It should be understood that, once the data is written to flash memory, the data is capable of residing on the flash memory indefinitely. As will be explained in further detail momentarily, this feature provides flexibility when restoring data storage system operations. Furthermore, in contrast to conventional data storage systems which require external UPS's and external cabling, the backup power supplies for the data storage system 20 can be relatively small (e.g., see the batteries 68 in Fig. 2) and there is no external cabling necessary thus making the above-described technique an attractive, simple and low cost mechanism for managing data during a power failure event.

It should be further understood that, in the context of a dual storage processor configuration 60 (also see Fig. 2), the controller 40 of each storage processor 62 preferably moves the contents of the volatile-memory storage caches 42 of that storage processor 62 to the flash-based memory vault 44 of that storage processor 62. That is, the storage processor 62(A) is configured to transfer the contents of the volatile-memory storage cache 42(A) to the flash-based memory vault 44(A) and, concurrently the storage processor 62(B) is configured to transfer the contents of the volatile-memory storage cache 42(B) to the flash-based memory vault 44(B). This contemporaneous operation is superior to the operation of conventional data storage systems where only one of a pair of storage processors writes the contents of its storage cache out to the magnetic disk drive vault on an array of magnetic disk drives. Further details will now be provided with reference to Fig. 4.

Fig. 4 illustrates a recovery procedure which is easily accomplished through

use of the flash-based memory vault 44. In particular, in some arrangements, the flash-based memory vault 44 is configured as a removable module that conveniently connects to and disconnects from other portions of the data storage system 20 through module connectors, e.g., in a manner similar to attaching and detaching a common memory stick to a general purpose computer through a USB port, in a manner similar to connecting a daughter card to and disconnecting the daughter card from a motherboard, and so on.

Moreover, in the situation of a dual storage processor configuration 60 such as that shown in Fig. 2, it should be understood that either flash-based memory vault 44(A), 44(B) contains the entire storage cache contents since the volatile-memory storage caches 42(A), 42(B) mirror each other. As such, only one flash-based memory vault 44(A), 44(B) is necessary to restore the storage cache state of the data storage system 20.

Accordingly, in the event of a hardware failure after safely storing the contents of the volatile-memory storage cache 42 into the flash-based memory vault 44, the flash-based memory vault 44 is then capable of being disconnected from the data storage system 20 and connected to new storage processing hardware (e.g., a new data storage system 20'), as generally shown by the arrow 90 in Fig. 4. The contents of the flash-based memory vault 44 are then capable of being restored onto each volatile-memory storage cache 42(A), 42(B) of the new hardware (e.g., mirrored through the bus 64 of the new data storage system 20', also see Fig. 2) thus enabling the new storage processing hardware to continue to perform data storage operations on behalf of the set of hosts 22. Under this situation, there is no loss of data. Further details will now be provided with reference to Fig. 5.

Fig. 5 is a block diagram of a first technique for restoring contents of the flash-based memory vault 44 to the volatile-memory storage cache 42 in the context of a dual storage processor configuration 60 (also see Fig. 2). Recall that the flash-based memory vaults 44(A), 44(B) (Fig. 2) form the flash-based memory vault 44 (Fig. 1), and that each flash-based memory vault 44(A), 44(B) is configured to contain the same up-to-date information since the volatile-memory storage caches 42(A), 42(B) mirror each other. Whether both volatile-memory storage caches

42(A), 42(B) contain the same current data can be confirmed by a check of time information on the flash-based memory vaults 44(A), 44(B) (e.g., by comparing the output of generation counters, by comparing timestamps, etc.).

If it turns out that one flash-based memory vault 44 contains more recent information, the contents of both volatile-memory storage caches 42(A), 42(B) can be restored from that flash-based memory vault 44. Otherwise, it does not matter which flash-based memory vault 44 provides the data during data restoration.

As shown in Fig. 5, in accordance with the first technique, the restoration transfer occurs in a two step process. In particular, one of the flash-based memory vaults 44(A), 44(B) (e.g., the flash-based memory vault 44(A)) provides the data to its respective volatile-memory storage cache 42(A), 42(B) directly (e.g., the volatile-memory storage cache 42(A)). Then, the data is copied through the bus 64 to the other volatile-memory storage cache 42 (e.g., the volatile-memory storage cache 42(A)). At completion, the data is mirrored by both volatile-memory storage caches 42(A), 42(B). Figs. 6 and 7 show alternative restoration techniques which are available if the flash-based memory vaults 44(A), 44(B) contain the same information.

Fig. 6 is a block diagram of a second technique for restoring contents of the flash-based memory vault 44 to the volatile-memory storage cache 42 in the context of a dual storage processor configuration 60 (also see Fig. 2). Here, the controller 40(A) (also see Fig. 2) restores the contents of flash-based memory vault 44(A) into the volatile-memory storage cache 42(A). Simultaneously, the controller 40(B) (also see Fig. 2) restores the contents of flash-based memory vault 44(B) into the volatile-memory storage cache 42(B).

It should be understood that the restoration technique illustrated in Fig. 6 provides an additional level of thoroughness. In particular, once the contents of the volatile-memory storage caches 42(A), 42(B) are restored, the controllers 40 can perform further tasks to guarantee accuracy of the data, e.g., a comparison of the contents of the volatile-memory storage caches 42(A), 42(B).

Fig. 7 is a block diagram of a third technique for restoring contents of the flash-based memory vault 44 to the volatile-memory storage cache 42 in the context

of a dual storage processor configuration 60 (also see Fig. 2). Under this third technique, the controller 40(A) (also see Fig. 2) restores half of the contents of flash-based memory vault 44(A) into the volatile-memory storage cache 42(A) (e.g., an upper half of the address space). Simultaneously, the controller 40(B) (also see Fig. 2) restores an opposite half of the contents of flash-based memory vault 44(B) into the volatile-memory storage cache 42(B) (e.g., a lower half of the address space). Next, the controllers 40 exchange their restored contents with each other through the bus 64 (also see Fig. 2) so that each volatile-memory storage cache 42 is completely restored (e.g., the upper half is copied from the volatile-memory storage cache 42(A) to the volatile-memory storage cache 42(B), and the lower half is copied from the volatile-memory storage cache 42(B) to the volatile-memory storage cache 42(A)).

It should be understood that the restoration technique illustrated in Fig. 7 provides a faster restoration time since the data transfer rate between the volatile-memory storage caches 42(A), 42(B) (e.g., the bandwidth of the bus 64, also see Fig. 2) is faster than the data transfer rate between the flash-based memory vaults 44 and the volatile-memory storage caches 42. Accordingly, this third technique is well-suited for situations in which recovery time must be kept to a minimum.

As mentioned above, an improved technique involves moving data within a data storage system 20 from a storage cache 42 into a flash-based memory vault 44 in response to a power failure signal 38. Such operation alleviates the need to provide backup power to magnetic disk drives 32. Rather, data can be moved from the storage cache 42 to the flash-based memory vault 44 using a relatively-small backup power source 28, e.g., a battery that only powers storage processing circuitry 30. Without the need for backup power to the magnetic disk drives 32, there is no burden of having to provide large, costly and complex backup power supplies and the associated external cabling for magnetic disk drives. That is, the magnetic disk drives 32 can simply turn off as soon as the primary power source 26 is lost. With the storage processing circuitry 30 still running from a backup power source (e.g., a dedicated battery), the storage processing circuitry 30 is capable of moving the contents of the storage cache 42 to the flash-based memory vault 44 thus preserving

data integrity of the data storage system 20 so that no data is ever lost.

While this invention has been particularly shown and described with references to preferred embodiments thereof, it will be understood by those skilled in the art that various changes in form and details may be made therein without departing from the spirit and scope of the invention as defined by the appended claims.

CLAIMS

What is claimed is:

1. A data storage system, comprising:

a volatile-memory storage cache;

a flash-based memory vault; and

a controller coupled to the volatile-memory storage cache and the flash-based memory vault, the controller being configured to:

while the controller is being powered by a primary power source, performing data storage operations on behalf of a set of hosts using the volatile-memory storage cache and a set of magnetic disk drives,

receiving a power failure signal indicating that the controller is now being powered by a backup power source rather than by the primary power source, and

moving data from the volatile-memory storage cache to a flash-based memory vault in response to the power failure signal.

2. A data storage system as in claim 1 wherein the controller includes processing circuitry which is configured to consume power (i) at a first power consumption rate when running at a first clock speed and (ii) at a second power consumption rate when running at a second clock speed, the second power consumption rate being lower than the first power consumption rate, and the second clock speed being slower than the first clock speed;

wherein the controller, when performing the data storage operations, is configured to run the processing circuitry at the first clock speed; and

wherein the controller, when moving the data from the volatile-memory storage cache to the flash-based memory vault, is configured to run

the processing circuitry at the second clock speed to conserve power from the backup power source while the processing circuitry moves the data from the volatile-memory storage cache to the flash-based memory vault.

3. A data storage system as in claim 2 wherein the controller, when running the processing circuitry at the first clock speed, is configured to operate a microprocessor of the processing circuitry within a first power range which is between 50 Watts and 100 Watts; and

wherein running the processing circuitry at the second clock speed includes operating the microprocessor of the processing circuitry within a second power range which is less than 30 Watts.

4. A data storage system as in claim 1 wherein (i) the volatile-memory storage cache, (ii) the flash-based memory vault and (iii) a controller configured to perform data storage operations reside within a processing core of the data storage system; and wherein the data storage system further comprises:

isolation circuitry configured to electrically isolate the processing core from other portions of the data storage system in response to the power failure signal to enable the controller to move the data from the volatile-memory storage cache to the flash-based memory vault without interference from the other portions of the data storage system.

5. A data storage system as in claim 1 wherein the flash-based memory vault is configured as a removable module to enable a user to

disconnect the flash-based memory vault from the data storage system,

connect the flash-based memory vault to a new data storage system, and

restore the data from the flash-based memory vault to a new volatile-memory storage cache of the new data storage system to continue to perform data storage operations on behalf of the set of hosts.

6. A data storage system as in claim 1 wherein the controller is further configured to:

receive a power normal signal; and

restore the data from the flash-based memory vault to the volatile-memory storage cache in response to the power normal signal.

7. A data storage system as in claim 1 wherein the data storage system includes:

a first storage processor having a first storage cache and a first memory vault, and

a second storage processor having a second storage cache and a second memory vault, the first and second storage caches being configured to mirror data, the first and second storage caches forming the volatile-memory storage cache of the data storage system, the first and second memory vaults forming the flash-based memory vault; and

wherein the controller, when moving the data from the volatile-memory storage cache to the flash-based memory vault, is configured to:

store contents of the first storage cache to the first memory vault, and

store contents of the second storage cache to the second memory vault.

8. A data storage system as in claim 7 wherein the controller is further configured to:

receive a power normal signal; and

restore the first and second storage caches from the first and second memory vaults which form the flash-based memory vault in response to the power normal signal, the first and second storage caches mirroring each other once the first and second storage caches have been restored.

9. A data storage system as in claim 8 wherein the controller, when restoring the first and second storage caches, is configured to:

transfer the contents of the first storage cache stored in the first memory vault back to the first storage cache, and

copy the contents transferred back to the first storage cache from the first storage cache to the second storage cache.

10. A data storage system as in claim 8 wherein the controller, when restoring the first and second storage caches, is configured to:

transfer a top half of the contents of the first storage cache stored in the first memory vault back to the first storage cache,

transfer a bottom half of the contents of the second storage cache stored in the second memory vault back to the second storage cache,

copy the bottom half of the contents transferred back to the second storage cache from the second storage cache to the first storage cache, and

copy the top half of the contents transferred back to the first storage cache from the first storage cache to the second storage cache.

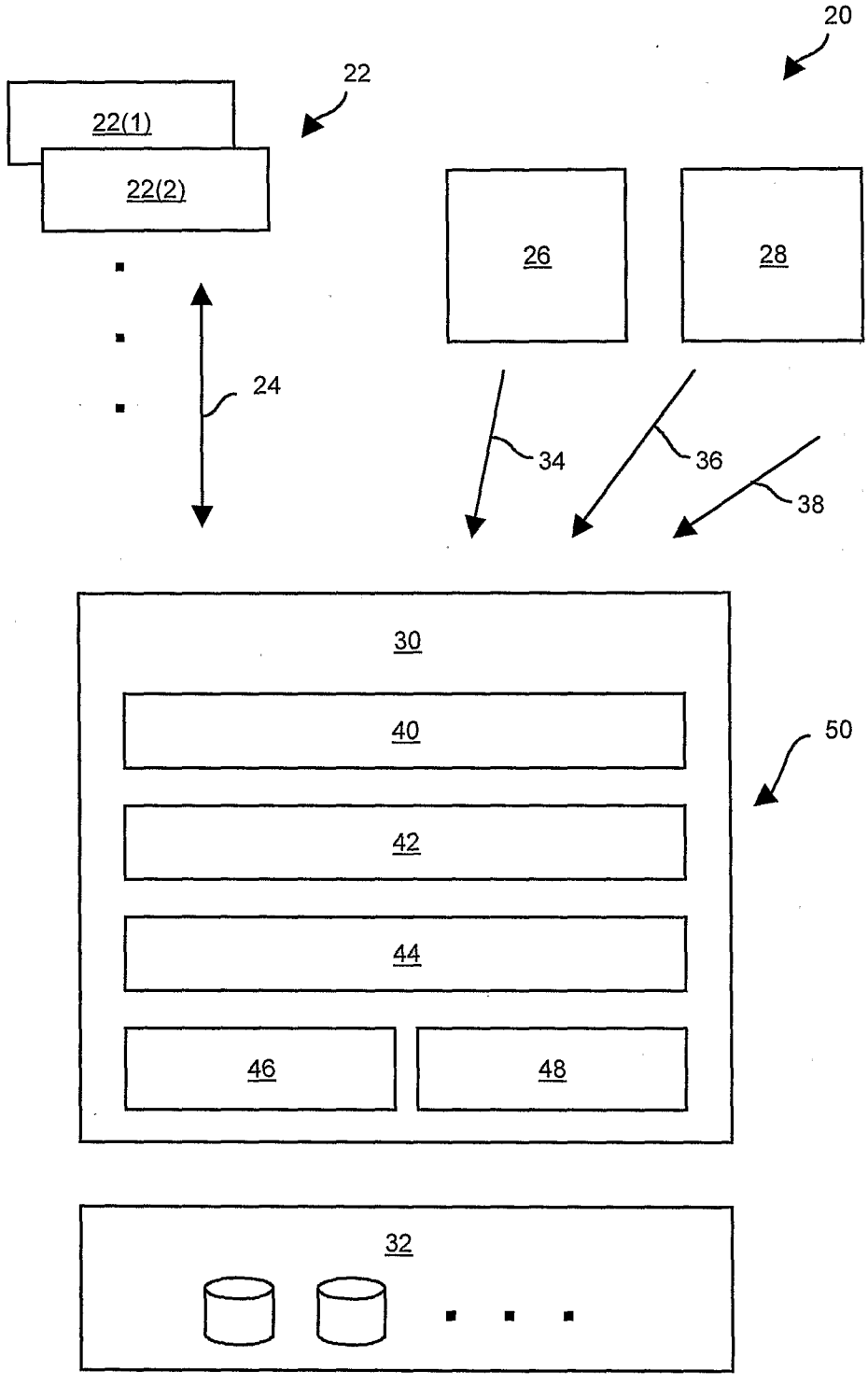


FIG. 1

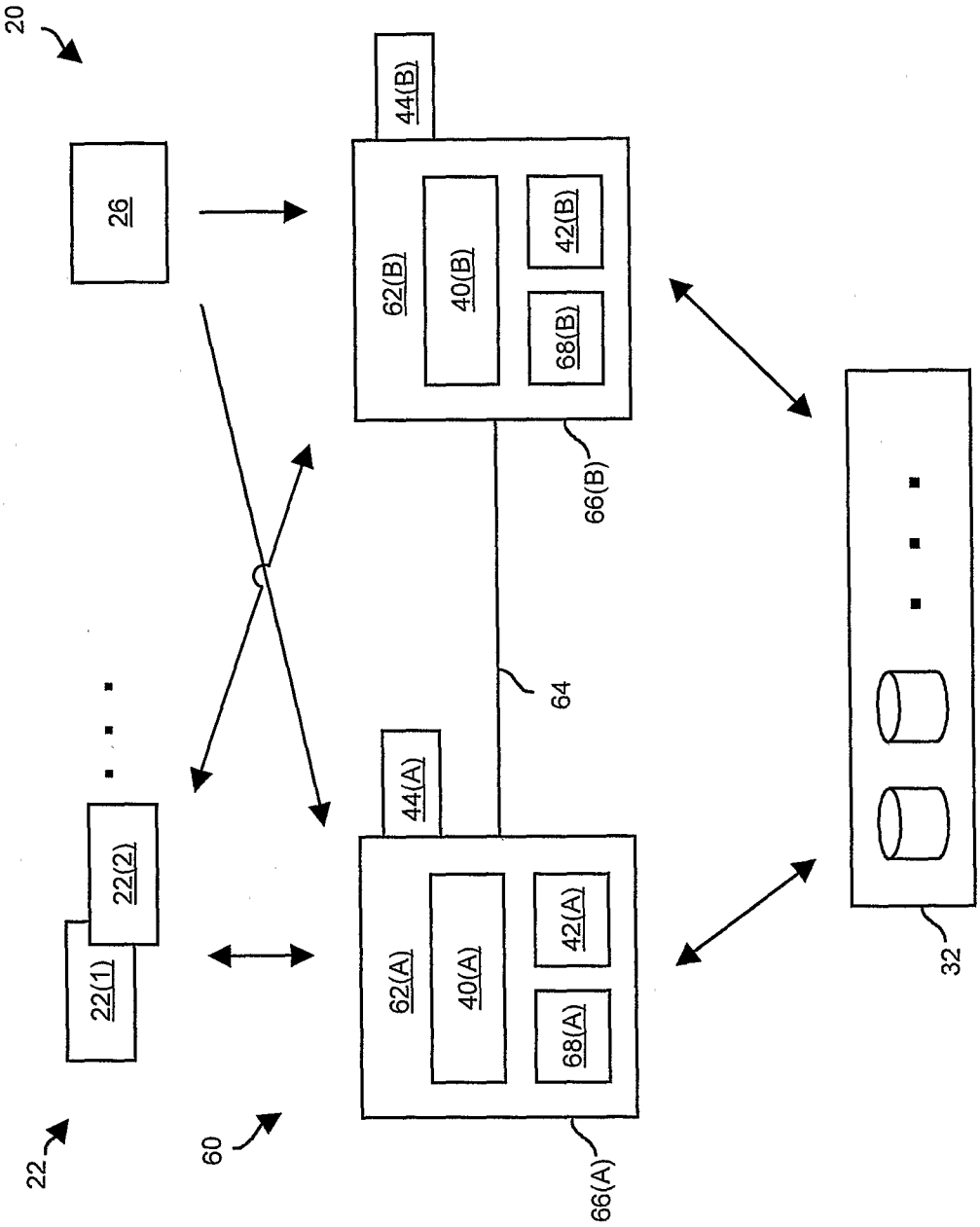

20
26
44(B)
62(B)
40(B)
42(B)
68(B)
66(B)
64
44(A)
62(A)
40(A)
42(A)
68(A)
66(A)
22(1)
22(2)
22
60
32

FIG. 2

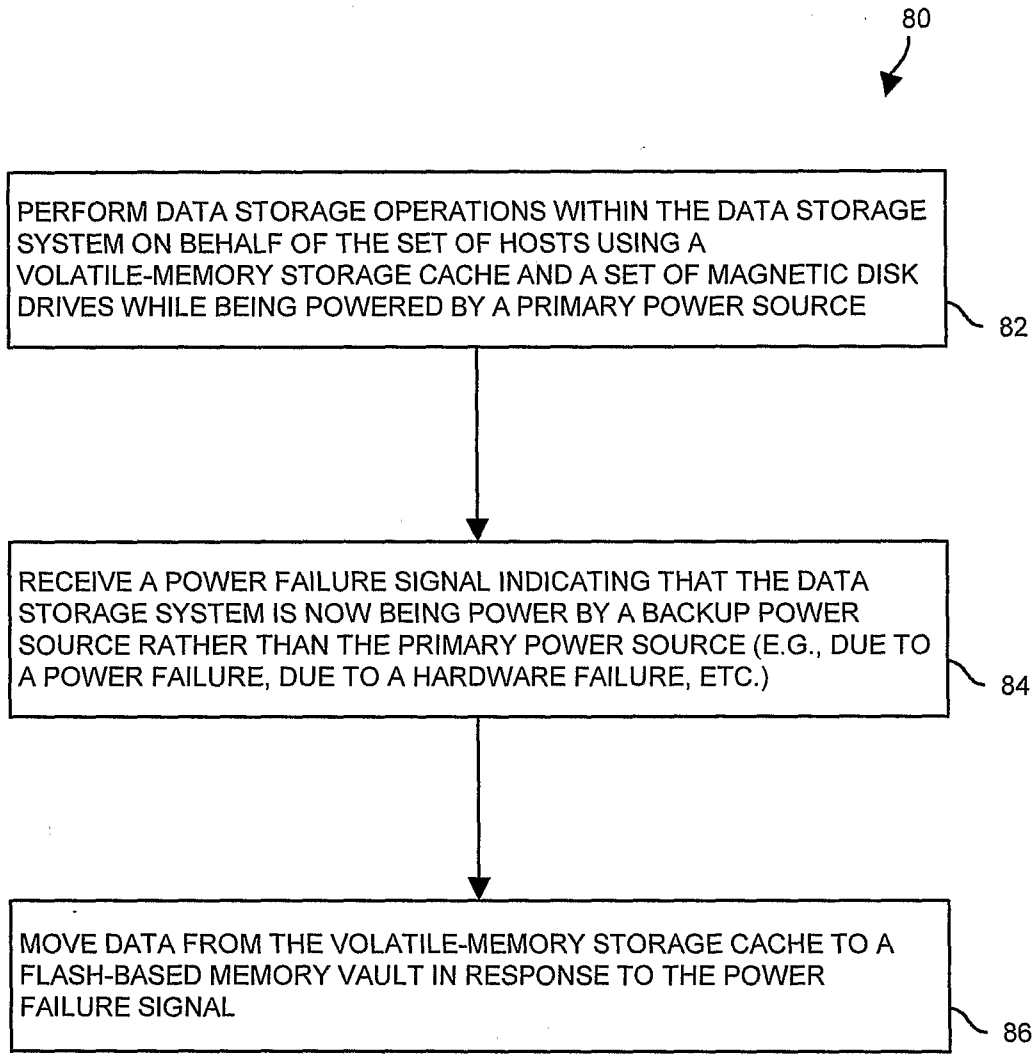

80
PERFORM DATA STORAGE OPERATIONS WITHIN THE DATA STORAGE SYSTEM ON BEHALF OF THE SET OF HOSTS USING A VOLATILE-MEMORY STORAGE CACHE AND A SET OF MAGNETIC DISK DRIVES WHILE BEING POWERED BY A PRIMARY POWER SOURCE
82
RECEIVE A POWER FAILURE SIGNAL INDICATING THAT THE DATA STORAGE SYSTEM IS NOW BEING POWER BY A BACKUP POWER SOURCE RATHER THAN THE PRIMARY POWER SOURCE (E.G., DUE TO A POWER FAILURE, DUE TO A HARDWARE FAILURE, ETC.)
84
MOVE DATA FROM THE VOLATILE-MEMORY STORAGE CACHE TO A FLASH-BASED MEMORY VAULT IN RESPONSE TO THE POWER FAILURE SIGNAL
86

FIG. 3

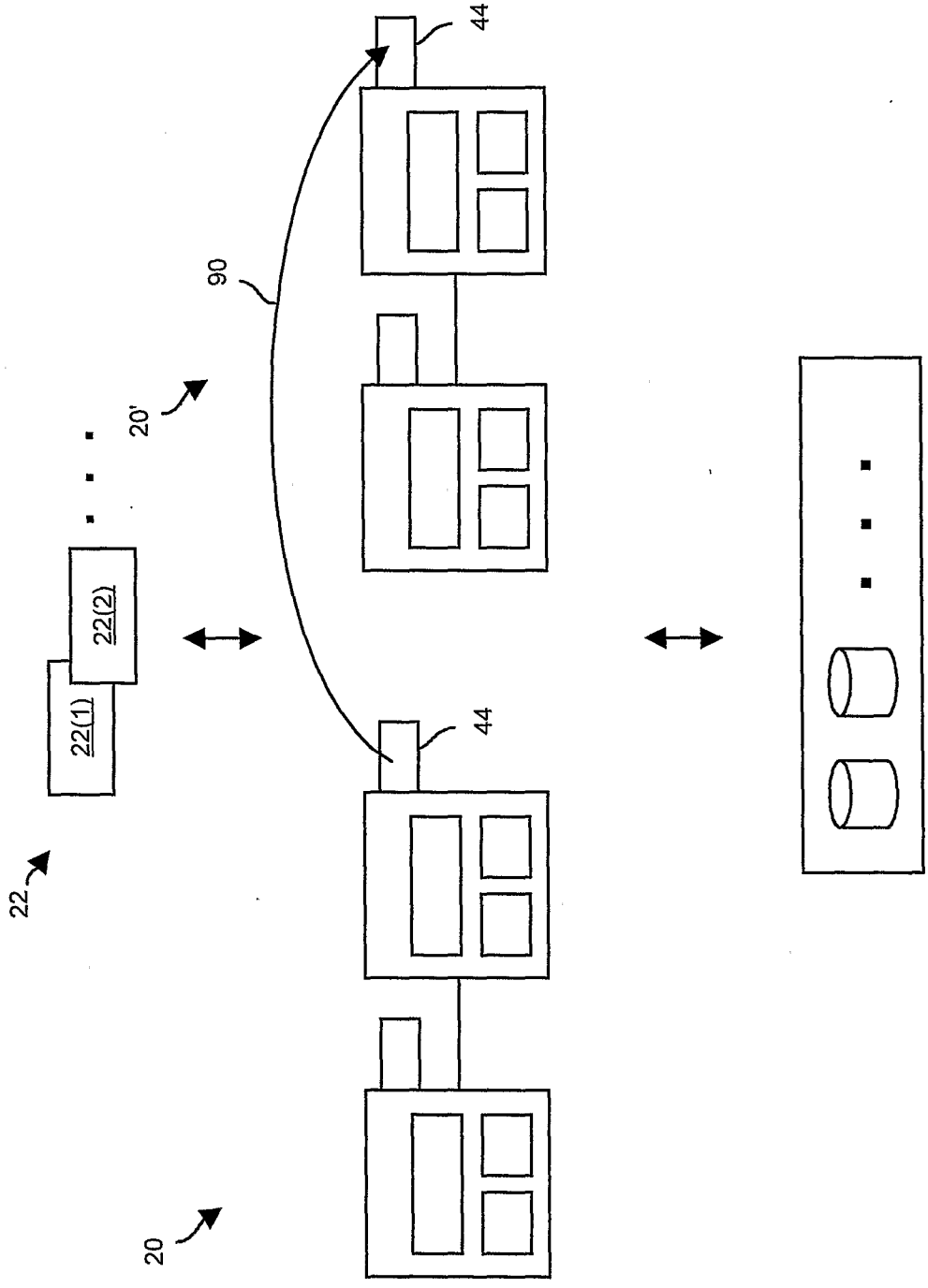


FIG. 4

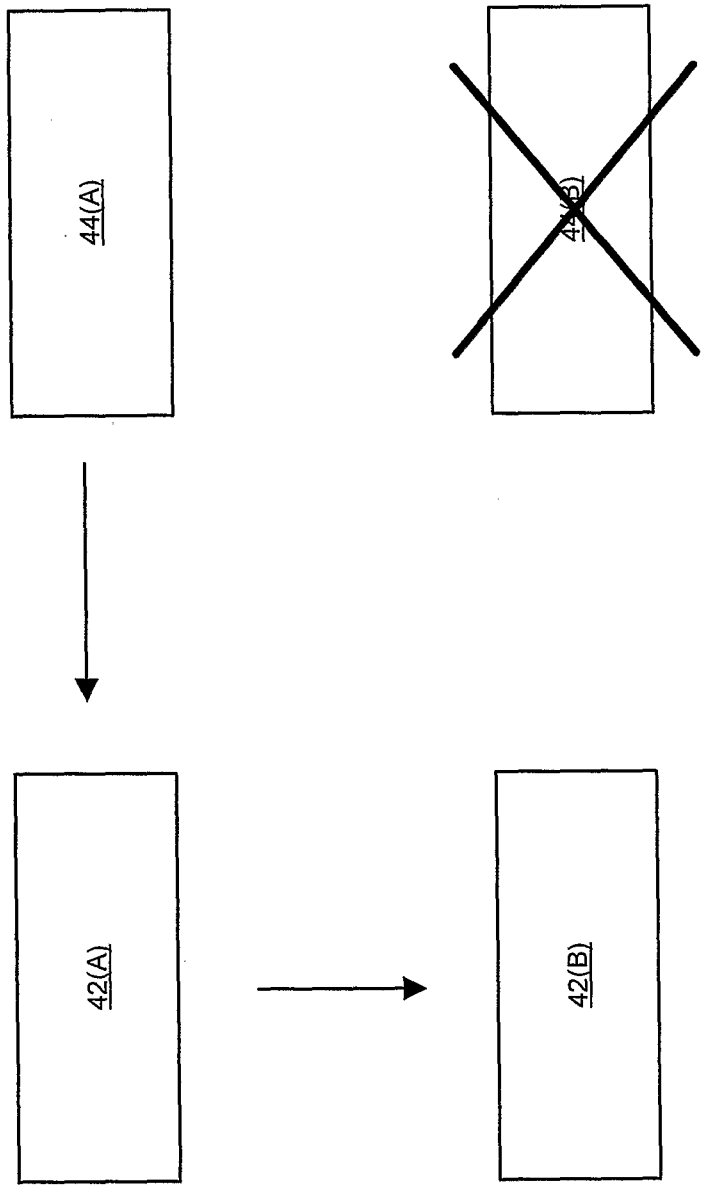


FIG. 5

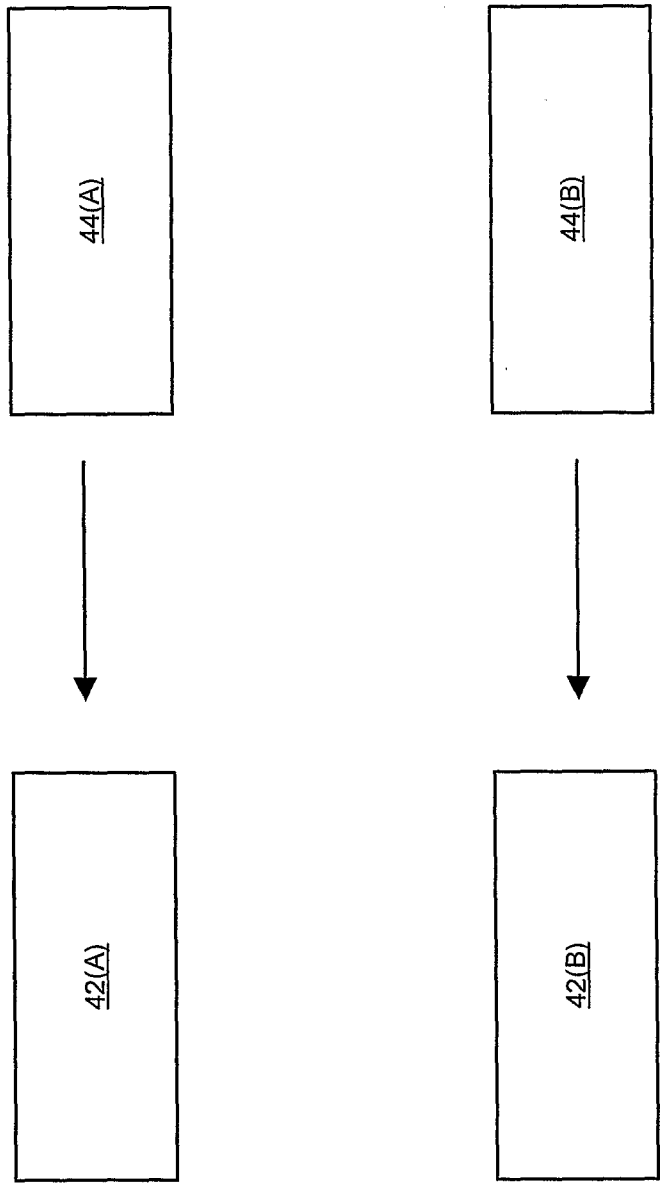


FIG. 6

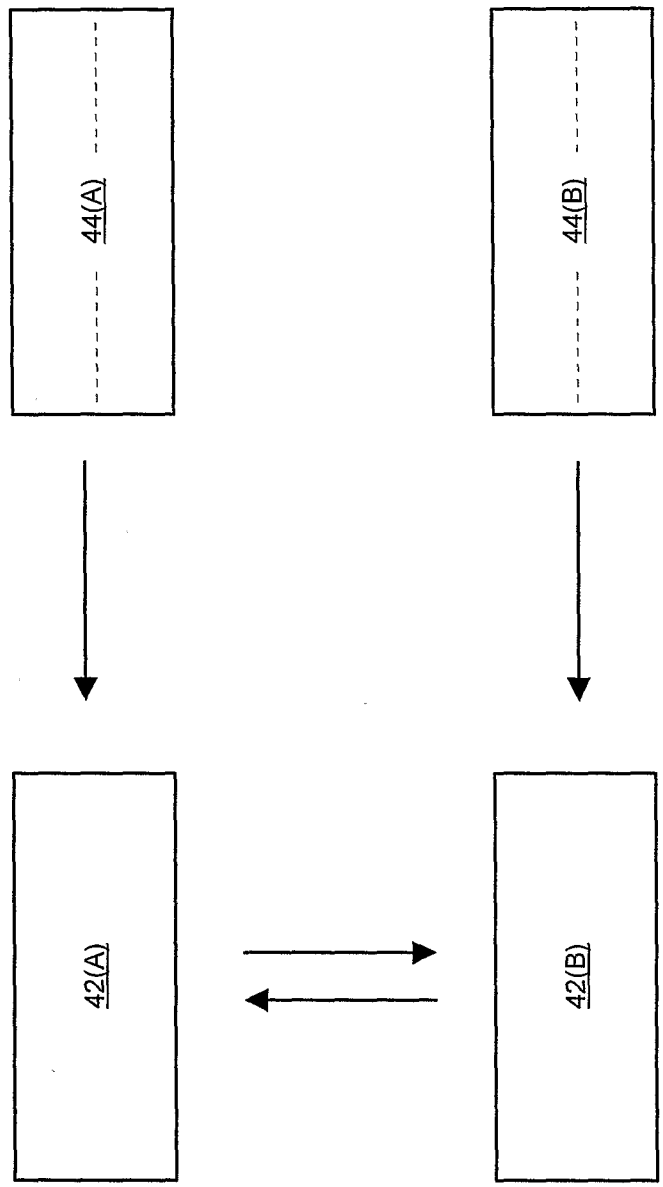


FIG. 7

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2006/045552

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06F11/14

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 5 799 200 A (BRANT WILLIAM A [US] ET AL) 25 August 1998 (1998-08-25)	1,4-10
Y	abstract figures 1,2 column 2, lines 14-24 column 4, lines 16-59 column 5, lines 13-22,52-64 column 6, lines 4-14 column 9, lines 52-54	2,3
Y	US 2002/152417 A1 (NGUYEN DON [US] ET AL) 17 October 2002 (2002-10-17) abstract paragraph [0012]	2,3

-/--

[X] Further documents are listed in the continuation of Box C.

[X] 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	Date of mailing of the international search report
26 April 2007	08/05/2007

Name and mailing address of the ISA/	Authorized officer
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	Weber, Vincent

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2006/045552

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 2003/126494 A1 (STRASSER AMNON A [IL]) 3 July 2003 (2003-07-03) abstract paragraphs [0004], [0010], [0022], [0033]	1,4-10
X	US 5 677 890 A (LIONG THOMAS SINGKIAT [US] ET AL) 14 October 1997 (1997-10-14) abstract column 3, lines 10-33	1,5,6
X	US 2004/103238 A1 (AVRAHAM MEIR [IL] ET AL) 27 May 2004 (2004-05-27) abstract paragraph [0008]	1,5

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2006/045552

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 5799200 A	25-08-1998	NONE	
US 2002152417 A1	17-10-2002	NONE	
US 2003126494 A1	03-07-2003	AU 2002367054 A1	30-07-2003
		EP 1470483 A1	27-10-2004
		WO 03060716 A1	24-07-2003
US 5677890 A	14-10-1997	AU 736507 B2	26-07-2001
		AU 2319797 A	22-09-1997
		BR 9702115 A	15-06-1999
		CA 2220391 A1	12-09-1997
		DE 69722100 D1	26-06-2003
		DE 69722100 T2	11-03-2004
		EP 0826218 A1	04-03-1998
		JP 11505659 T	21-05-1999
		WO 9733285 A1	12-09-1997
US 2004103238 A1	27-05-2004	KR 20040047584 A	05-06-2004