(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 538 129**

(51) Int. CI.:

G06F 17/30 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **21.10.2009 E 09740675 (5)**

(97) Fecha y número de publicación de la concesión europea: **11.03.2015 EP 2342663**

(54) Título: **Almacenamiento de datos distribuido**

(30) Prioridad:

24.10.2008 SE 0802277

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
17.06.2015

(73) Titular/es:

COMPUVERDE AB (100.0%)
Östra Vittusgatan 36
371 33 Karlskrona, SE

(72) Inventor/es:

MELANDER, CHRISTIAN;
BERNBO, STEFAN;
PETERSSON, GUSTAV y
PERSSON, ROGER

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Almacenamiento de datos distribuido

Campo técnico

La presente exposición se refiere a métodos para escribir y mantener datos en un sistema de almacenamiento de datos que comprende una pluralidad de nodos de almacenamiento de datos, utilizándose los métodos en un servidor y en un nodo de almacenamiento del sistema de almacenamiento de datos. La exposición se refiere además a nodos de almacenamiento o servidores con capacidad de llevar a cabo dichos métodos.

Antecedentes

Se da a conocer un método de este tipo, por ejemplo, en el documento US. 2005/0246393. A1. Este método se da a conocer para un sistema que usa una pluralidad de centros de almacenamiento en ubicaciones geográficamente dispares. Se incluyen administradores de almacenamiento distribuido de objetos con el fin de mantener información referente a datos almacenados.

Uno de los problemas asociados a un sistema de este tipo es cómo lograr una escritura así como un mantenimiento de datos, que sean sencillos aunque robustos y fiables.

El documento US-7266556-B1 describe un sistema de almacenamiento en red con un sistema de archivos virtual. El sistema incluye una serie de administradores de almacenamiento distribuido y una serie de nodos de almacenamiento inteligentes. Se almacenan duplicados de archivos en dos nodos de almacenamiento, y en caso de fallo de un nodo de almacenamiento, se dirige una solicitud del archivo al otro nodo de almacenamiento en el que se guarda el archivo.

HONG TANG ET AL: “An Efficient Data Localization Protocol for Self-organizing Storage Clusters”, 2003 ACM/IEEE CONFERENCE PHOENIX, AZ, USA 15 a 21 de NOV. de 2003, describe un agrupamiento para almacenamiento a gran escala. Se usa una combinación de diferentes esquemas para el posicionamiento de datos, que incluye posicionar bloques pequeños de datos sobre la base de un *hashing* y bloques más grandes con filtros Bloom.

Sumario de la Invención

El objetivo de la invención se alcanza con las características de las reivindicaciones independientes. Otras realizaciones constituyen la materia objeto de las reivindicaciones dependientes.

Por lo tanto, un objetivo de la presente exposición es materializar una escritura o mantenimiento robustos de datos en un sistema de almacenamiento distribuido, sin utilizar servidores de mantenimiento centralizados, los cuales pueden constituir en sí mismos uno de los puntos débiles de un sistema. Este objetivo se alcanza por medio de un método del tipo mencionado en el inicio, que se materializa en un nodo de almacenamiento y comprende: monitorizar el estado de otros nodos de almacenamiento en el sistema así como operaciones de escritura llevadas a cabo en el sistema de almacenamiento de datos, detectar, basándose en la monitorización, condiciones en el sistema de almacenamiento de datos, que implican la necesidad de duplicación de datos entre los nodos en el sistema de almacenamiento de datos, e iniciar un proceso de duplicación en caso de que se detecte dicha condición. El proceso de duplicación incluye enviar un mensaje de multidifusión a una pluralidad de nodos de almacenamiento, de manera que el mensaje consulta sobre cuáles de esos nodos de almacenamiento almacenan datos específicos.

Con un método del tipo mencionado, cada nodo de almacenamiento puede participar de manera activa en el mantenimiento de datos del sistema completo. En caso de que un nodo de almacenamiento falle, sus datos pueden ser recuperados por otros nodos del sistema, pudiéndose considerar por lo tanto que dicho sistema es auto-reparable.

La monitorización puede incluir la escucha de señales de latidos (*heart-beat*) provenientes de otros nodos de almacenamiento del sistema. Una condición que implica una necesidad de duplicación es entonces un nodo de almacenamiento que está funcionando deficientemente.

Los datos incluyen archivos y una condición que implica la necesidad de duplicaciones puede ser entonces una de entre una eliminación de archivos o una inconsistencia de archivos.

Se puede mantener una lista de duplicación, que incluye archivos que requieren duplicación, y la misma puede incluir prioridades.

El proceso de duplicación puede incluir: enviar una solicitud en forma de mensaje de multidifusión a una pluralidad de nodos de almacenamiento consultando sobre cuáles de dichos nodos de almacenamiento almacenan datos específicos, recibir respuestas de aquellos nodos de almacenamiento que contienen dichos datos específicos, determinar si dichos datos específicos están almacenados en un número suficiente de nodos de almacenamiento, y, en caso negativo, seleccionar por lo menos un nodo de almacenamiento adicional y transmitir dichos datos

específicos a ese nodo de almacenamiento. Además, los datos específicos de nodos de almacenamiento que contienen versiones obsoletas de los mismos se pueden actualizar.

Adicionalmente, el proceso de duplicación puede comenzar con el nodo de almacenamiento intentado obtener dominio del archivo, que se va a duplicar, entre todos los nodos de almacenamiento del sistema.

La monitorización puede incluir además monitorizar operaciones de lectura llevadas a cabo en el sistema de almacenamiento de datos.

La presente exposición se refiere además a un nodo de almacenamiento de datos, para llevar a cabo mantenimiento de datos, correspondiente al método. En este caso, el nodo de almacenamiento comprende en general medios para llevar a cabo las acciones del método.

El objetivo también se alcanza por medio de un método para escribir datos en un sistema de almacenamiento de datos del tipo mencionado en el inicio, lo cual se materializa en un servidor que ejecuta una aplicación que accede a datos en el sistema de almacenamiento de datos. El método comprende: enviar una consulta de almacenamiento de multidifusión a una pluralidad de nodos de almacenamiento, recibir una pluralidad de respuestas de un subconjunto de dichos nodos de almacenamiento, incluyendo las respuestas datos geográficos referentes a la posición geográfica de cada servidor, seleccionar por lo menos dos nodos de almacenamiento del subconjunto, basándose en dichas respuestas, y enviar datos y un identificador de datos, correspondiente a los datos, a los nodos de almacenamiento seleccionados.

Este método logra una escritura robusta de datos en el sentido que se materializa una diversidad geográfica de una manera eficiente.

La posición geográfica puede incluir latitud y longitud del nodo de almacenamiento en cuestión, y las respuestas pueden incluir además carga del sistema y/o antigüedad del sistema para el nodo de almacenamiento en cuestión.

La consulta de almacenamiento de multidifusión puede incluir un identificador de datos, que identifica los datos a almacenar.

Típicamente, se pueden seleccionar por lo menos tres nodos para el almacenamiento, y se puede enviar una lista de nodos de almacenamiento, que almacenan satisfactoriamente los datos, a los nodos de almacenamiento seleccionados.

La presente exposición se refiere además a un servidor, para llevar a cabo escritura de datos, correspondiente al método. En este caso, el servidor comprende en general medios para llevar a cabo las acciones del método.

Breve descripción de los dibujos

La figura 1 ilustra un sistema de almacenamiento distribuido de datos.
Las figuras 2A a 2C y la figura 3 ilustran un proceso de lectura de datos.
Las figuras 4A a 4C y la figura 5 ilustran un proceso de escritura de datos.
La figura 6 ilustra esquemáticamente una situación en la que una serie de archivos se almacena entre una serie de nodos de almacenamiento de datos.
La figura 7 ilustra la transmisión de señales de latidos.
La figura 8 es una vista general de un proceso de mantenimiento de datos.

Descripción detallada

La presente exposición se refiere a un sistema de almacenamiento distribuido de datos que comprende una pluralidad de nodos de almacenamiento. La estructura del sistema y el contexto en que se utiliza se esbozan de manera general en la figura 1.

Un ordenador 1 de usuario accede, por medio de Internet 3, a una aplicación 5 que se ejecuta en un servidor 7. El contexto de usuario, según se ilustra en la presente, es por lo tanto una configuración habitual de cliente-servidor, la cual es bien conocida de por sí. Sin embargo, debe indicarse que el sistema de almacenamiento de datos que se va a dar a conocer también puede ser útil en otras configuraciones.

En el caso ilustrado, en el servidor 7 se ejecutan dos aplicaciones 5, 9. Sin embargo, evidentemente este número de aplicaciones puede ser diferente. Cada aplicación tiene una API (Interfaz de Programación de Aplicaciones) 11 que proporciona una interfaz con relación al sistema 13 de almacenamiento distribuido de datos y soporta solicitudes, típicamente solicitudes de escritura y lectura, provenientes de las aplicaciones que se ejecutan en el servidor. Desde el punto de vista de una aplicación, no es necesario que la lectura o escritura de información del/en el sistema 13 de almacenamiento de datos se manifieste como diferente con respecto al uso de cualquier otro tipo de solución de almacenamiento, por ejemplo, un servidor de archivos o simplemente una unidad de disco duro.

Cada API 11 se comunica con nodos 15 de almacenamiento del sistema 13 de almacenamiento de datos, y los nodos de almacenamiento se comunican entre sí. Estas comunicaciones están basadas en el TCP (Protocolo de Control de Transmisiones) y el UDP (Protocolo de Datagrama de Usuario). Estos conceptos son bien conocidos por aquellos versados en la materia, y no se explican de manera adicional.

Debe indicarse que APIs diferentes 11 en el mismo servidor 7 pueden acceder a conjuntos diferentes de nodos 15 de almacenamiento. Debe indicarse además que puede existir más de un servidor 7 que acceda a cada nodo 15 de almacenamiento. Sin embargo, esto no afecta en mayor medida a la manera en la que funcionan los nodos de almacenamiento, tal como se describirá posteriormente.

Los componentes del sistema de almacenamiento distribuido de datos son los nodos 15 de almacenamiento y las APIs 11, en el servidor 7 que accede a los nodos 15 de almacenamiento. Por lo tanto, la presente exposición se refiere a métodos llevados a cabo en el servidor 7 y en los nodos 15 de almacenamiento. Dichos métodos se materializarán principalmente en forma de implementaciones de software que se ejecutan en el servidor y en los nodos de almacenamiento, respectivamente, y son determinantes en conjunto para el funcionamiento y las propiedades del sistema de almacenamiento distribuido de datos global.

Típicamente, el nodo 15 de almacenamiento se puede materializar por medio de un servidor de archivos que está provisto de una serie de bloques funcionales. El nodo de almacenamiento puede comprender así un soporte 17 de almacenamiento, el cual típicamente está compuesto por una serie de unidades de disco duro, opcionalmente configuradas como un sistema RAID (Conjunto Redundante de Discos Independientes). Sin embargo, también son concebibles otros tipos de soportes de almacenamiento.

El nodo 15 de almacenamiento puede incluir además un directorio 19, el cual comprende listas de relaciones de entidades de datos/nodos de almacenamiento en forma de una lista de anfitriones, tal como se describirá posteriormente.

Además de la lista de anfitriones, cada nodo de almacenamiento contiene adicionalmente una lista de nodos que incluye las direcciones IP de todos los nodos de almacenamiento de su conjunto o grupo de nodos de almacenamiento. El número de nodos de almacenamiento de un grupo puede variar desde unos pocos hasta cientos de nodos de almacenamiento. La lista de nodos puede tener además un número de versión.

Adicionalmente, el nodo 15 de almacenamiento puede incluir un bloque 21 de duplicación y un bloque 23 de monitorización de agrupamientos. El bloque 21 de duplicación incluye una API 25 de nodo de almacenamiento, y está configurado para ejecutar funciones con el fin de identificar la necesidad de un proceso de duplicación y de llevar a cabo el mismo, tal como se describirá de forma detallada posteriormente. La API 25 de nodo de almacenamiento del bloque 21 de duplicación puede contener código que en gran medida se corresponda con el código de la API 11 de nodo de almacenamiento del servidor 7, ya que el proceso de duplicación comprende acciones que se corresponden en gran medida con las acciones llevadas a cabo por el servidor 7 durante operaciones de lectura y escritura que se van a describir. Por ejemplo, la operación de escritura llevada a cabo durante la duplicación se corresponde en gran medida con la operación de escritura llevada a cabo por el servidor 7. El bloque 23 de monitorización de agrupamientos está configurado para llevar a cabo la monitorización de otros nodos de almacenamiento en el sistema 13 de almacenamiento de datos, tal como se describirá de forma más detallada posteriormente.

Puede considerarse que los nodos 15 de almacenamiento del sistema de almacenamiento distribuido de datos existen en el mismo nivel jerárquico. No hay necesidad de designar ningún nodo de almacenamiento maestro que sea responsable del mantenimiento de un directorio de entidades de datos almacenadas y de la monitorización de la consistencia de datos, etcétera. Por el contrario, todos los nodos 15 de almacenamiento se pueden considerar como iguales y, en ocasiones, pueden llevar a cabo operaciones de gestión de datos con respecto a otros nodos de almacenamiento del sistema. Esta igualdad garantiza que el sistema es robusto. En caso de un funcionamiento deficiente de un nodo de almacenamiento, otros nodos del sistema cubrirán al nodo que funciona deficientemente y garantizarán un almacenamiento de datos fiable.

Se describirá el funcionamiento del sistema en el siguiente orden: lectura de datos, escritura de datos y mantenimiento de datos. Aun cuando estos métodos funcionan muy bien juntos, debe indicarse que, en principio, también pueden llevarse a cabo de manera independientemente mutua. Es decir, por ejemplo, el método de lectura de datos puede proporcionar propiedades excelentes incluso si no se utiliza el método de escritura de datos de la presente exposición y viceversa.

A continuación se describe el método de lectura en referencia a las figuras 2A a 2C y 3, siendo esta última un diagrama de flujo que ilustra el método.

La lectura, así como otras funciones en el sistema, utilizan una comunicación de multidifusión para comunicarse simultáneamente con una pluralidad de nodos de almacenamiento. Con multidifusión o multidifusión IP se pretende

significar en este caso una comunicación de punto-a-multipunto, que se logra mediante el envío de un mensaje a una dirección IP que está reservada para aplicaciones de multidifusión.

Por ejemplo, se envía un mensaje, típicamente una solicitud, a una dirección IP del tipo mencionado (por ejemplo, 244.0.0.1), y una serie de servidores destinatarios se registra como abonados a esa dirección IP. Cada uno de los servidores destinatarios tiene su propia dirección IP. Cuando un conmutador de la red recibe el mensaje dirigido a 244.0.0.1, el conmutador reenvía el mensaje a las direcciones IP de cada servidor registrado como abonado.

En principio, solamente un servidor puede registrarse como abonado a una dirección de multidifusión, en cuyo caso se logra una comunicación de punto-a-punto. No obstante, en el contexto de esta exposición, dicha comunicación se considera sin embargo una comunicación de multidifusión, ya que se utiliza un esquema de multidifusión.

También se utiliza una comunicación de unidifusión en relación con una comunicación con un solo destinatario.

En referencia a la figura 2A y la figura 3, el método para recuperar datos de un sistema de almacenamiento de datos comprende el envío 31 de una consulta de multidifusión a una pluralidad de nodos 15 de almacenamiento. En el caso ilustrado, hay cinco nodos de almacenamiento que tienen, cada uno de ellos, una dirección IP (Protocolo de Internet) 192.168.1.1, 192.168.1.2, etcétera. Huelga decir que el número de nodos de almacenamiento es solamente un ejemplo. La consulta contiene un identificador de datos "2B9B4A97-76E5-499E-A21A6D7932DD7927", que puede ser, por ejemplo, un Identificador Universalmente Único, UUID, el cual es bien conocido per se.

Los propios nodos de almacenamiento buscan datos correspondientes al identificador. Si se encuentran dichos datos, un nodo de almacenamiento envía una respuesta, que es recibida 33 por el servidor 7, consúltese la figura 2B. Tal como se ilustra, la respuesta puede contener opcionalmente otra información además de una indicación de que el nodo de almacenamiento tiene una copia de los datos pertinentes. Específicamente, la respuesta puede contener información del directorio de nodos de almacenamiento, sobre otros nodos de almacenamiento que contienen los datos, información referente a qué versión de los datos está contenida en el nodo de almacenamiento e información en relación con a qué carga está expuesto el nodo de almacenamiento en ese momento.

Basándose en las respuestas, el servidor selecciona 35 uno o más nodos de almacenamiento a partir de los cuales se van a recuperar datos, y envía 37 a ese/esos nodos de almacenamiento una solicitud de datos en unidifusión, consúltese la figura 2C.

Como respuesta a la solicitud de datos, el nodo/nodos de almacenamiento envían los datos pertinentes por unidifusión al servidor el cual recibe 39 los datos. En el caso ilustrado, se selecciona solamente un nodo de almacenamiento. Aunque con esto es suficiente, es posible seleccionar más de un nodo de almacenamiento con el fin de recibir dos conjuntos de datos, lo cual posibilita una comprobación de consistencia. Si falla la transferencia de datos, el servidor puede seleccionar otro nodo de almacenamiento para la recuperación.

La selección de nodos de almacenamiento se puede basar en un algoritmo que tiene en cuenta varios factores con el fin de lograr un buen rendimiento global del sistema. Típicamente, se seleccionará el nodo de almacenamiento que tenga la última versión de los datos y la carga más baja, aunque son totalmente concebibles otros conceptos.

Opcionalmente, la operación se puede concluir por medio del envío, por parte del servidor, de una lista a todos los nodos de almacenamiento implicados, indicando qué nodos contienen los datos y con qué versión. Basándose en esta información, los propios nodos de almacenamiento pueden mantener los datos de manera apropiada mediante el proceso de duplicación que se describirá.

Las figuras 4A a 4C y la figura 5 ilustran un proceso de escritura de datos para el sistema de almacenamiento distribuido de datos.

En referencia a la figura 4A y la figura 5, el método comprende un servidor que envía 41 una consulta de almacenamiento de multidifusión a una pluralidad de nodos de almacenamiento. La consulta de almacenamiento comprende un identificador de datos y consiste básicamente en una pregunta sobre si los nodos de almacenamiento receptores pueden almacenar este archivo. Opcionalmente, los nodos de almacenamiento pueden comprobar con sus directorios internos si ya disponen de un archivo con este nombre, y, en el caso improbable de que así sea, pueden remitir una notificación al servidor 7, de tal manera que el servidor pueda renombrar el archivo.

En cualquier caso, por lo menos un subconjunto de los nodos de almacenamiento proporcionará respuestas mediante transmisión de unidifusión al servidor 7. Típicamente, los nodos de almacenamiento que tienen un espacio de disco libre mínimo predeterminado responderán a la consulta. El servidor 7 recibe 43 las respuestas las cuales incluyen datos geográficos referentes a la posición geográfica de cada servidor. Por ejemplo, tal como se indica en la figura 4B, los datos geográficos pueden incluir la latitud, la longitud y la altitud de cada servidor. Sin embargo, también pueden ser concebibles otros tipos de datos geográficos, tales como un código postal o similar.

Además de los datos geográficos, puede proporcionarse otra información que sirva como entrada a un proceso de selección de nodos de almacenamiento. En el ejemplo ilustrado, se proporciona la cantidad de espacio libre en cada nodo de almacenamiento junto con una indicación de la antigüedad del sistema del nodo de almacenamiento y una indicación de la carga que experimenta en ese momento el nodo de almacenamiento.

Basándose en las respuestas recibidas, el servidor selecciona 45 por lo menos dos, en una realización típica, tres, nodos de almacenamiento del subconjunto para almacenar los datos. La selección de nodos de almacenamiento se lleva a cabo por medio de un algoritmo que tiene en cuenta diferentes datos. La selección se lleva a cabo con el fin de lograr algún tipo de diversidad geográfica. Preferentemente, debe evitarse por lo menos que, como nodos de almacenamiento, se seleccionen solamente servidores de archivos del mismo bastidor. Típicamente, puede lograrse una gran diversidad geográfica, incluso seleccionando nodos de almacenamiento en continentes diferentes. Además de la diversidad geográfica, en el algoritmo de selección pueden incluirse otros parámetros. Siempre que se alcance una diversidad geográfica mínima, también pueden tenerse en cuenta el espacio libre, la antigüedad del sistema y la carga actual.

Cuando se han seleccionado los nodos de almacenamiento, se envían a cada nodo seleccionado los datos a almacenar y un identificador de datos correspondiente, típicamente usando una transmisión de unidifusión.

Opcionalmente, la operación puede concluir por medio del envío de un acuse de recibo al servidor, por parte de cada nodo de almacenamiento que ha llevado a cabo satisfactoriamente la operación de escritura. A continuación, el servidor envía una lista a todos los nodos de almacenamiento implicados, indicando qué nodos han escrito satisfactoriamente los datos y cuáles no. Basándose en esta información, los propios nodos de almacenamiento pueden mantener correctamente los datos mediante el proceso de duplicación que se describirá. Por ejemplo, si fallase la escritura de un nodo de almacenamiento, se produce la necesidad de duplicar el archivo en otro nodo de almacenamiento más, con el fin de alcanzar el número deseado de nodos de almacenamiento que almacenan ese archivo.

El método de escritura de datos permite en sí mismo que una API en un servidor 7 almacene datos de una manera muy robusta, en la medida en la que se puede proporcionar una diversidad geográfica excelente.

Además de las operaciones de escritura y lectura, la API en el servidor 7 puede llevar a cabo operaciones que eliminen archivos y actualicen archivos. Estos procesos se describirán a continuación con respecto al proceso de mantenimiento de datos.

La finalidad del proceso de mantenimiento de datos es garantizar que un número razonable de nodos de almacenamiento que no funcionan deficientemente almacena, cada uno de ellos, la última versión de cada archivo. Adicionalmente, puede proporcionar la función de que no se almacenen archivos eliminados en ningún nodo de almacenamiento. El mantenimiento lo llevan a cabo los propios nodos de almacenamiento. Por lo tanto, no existe la necesidad de un nodo "maestro" dedicado que cargue con la responsabilidad del mantenimiento del almacenamiento de datos. Esto garantiza una fiabilidad mejorada puesto que, si no, el nodo "maestro" constituiría un punto débil del sistema.

La figura 6 ilustra esquemáticamente una situación en la que una serie de archivos se almacena entre una serie de nodos de almacenamiento de datos. En el caso ilustrado, se representan con fines ilustrativos doce nodos, que tienen direcciones IP numeradas consecutivamente de 192.168.1.1 a 192.168.1.12. No obstante, huelga decir que no es necesario que los números de direcciones IP se encuentren en el mismo intervalo en absoluto. Los nodos se sitúan en sucesión circular únicamente para simplificar la descripción, es decir, no es necesario que los nodos presenten ningún orden particular. Para simplificar, cada nodo almacena uno o dos archivos identificados con las letras A a F.

En referencia a la figura 8, el método para el mantenimiento de datos comprende la detección 51 de condiciones en el sistema de almacenamiento de datos que implican la necesidad de duplicación de datos entre los nodos en el sistema de almacenamiento de datos, y un proceso de duplicación 53. El resultado del proceso de detección 51 es una lista 55 de archivos para los cuales se ha identificado la necesidad de duplicación. La lista puede incluir además datos referentes a la prioridad de las diferentes necesidades de duplicación. Basándose en esta lista, se lleva a cabo el proceso de duplicación 53.

La robustez del almacenamiento distribuido se fundamenta en que, en el sistema, se almacene un número razonable de copias de cada archivo, versiones correctas. En el caso ilustrado, se almacenan tres copias de cada archivo. No obstante, si por ejemplo, el nodo de almacenamiento con la dirección 192.168.1.5 falla, no se alcanzará el número deseado de copias almacenadas correspondientes a los archivos "B" y "C".

Por lo tanto, uno de los eventos que da como resultado una necesidad de duplicación es el funcionamiento deficiente de un nodo de almacenamiento en el sistema.

Cada nodo de almacenamiento del sistema puede monitorizar el estado de otros nodos de almacenamiento en el sistema. Esto se puede llevar a cabo dejando que cada nodo de almacenamiento emita una denominada señal de latido, a intervalos regulares, tal como se ilustra en la figura 7. En el caso ilustrado, el nodo de almacenamiento con la dirección 192.168.1.7 emite una señal de multidifusión 57 hacia los otros nodos de almacenamiento en el sistema, indicando que está funcionando correctamente. Esta señal puede ser recibida por la totalidad del resto de nodos de almacenamiento en funcionamiento en el sistema que llevan a cabo la monitorización de latidos 59 (consúltese la figura 8) o un subconjunto de los mismos. No obstante, en el caso del nodo de almacenamiento con la dirección 192.168.1.5, este nodo está funcionando deficientemente y no emite ninguna señal de latido. Por lo tanto, los otros nodos de almacenamiento percibirán que este nodo no ha emitido ninguna señal de latido durante mucho tiempo, lo cual indica que el nodo de almacenamiento en cuestión está fuera de servicio.

La señal de latido puede incluir, además de la dirección del nodo de almacenamiento, el número de versión de su lista de nodos. Otro nodo de almacenamiento, que esté a la escucha de la señal de latido y que averigüe que el nodo de almacenamiento transmisor tiene una lista de nodos de una versión posterior, puede solicitar entonces que el nodo de almacenamiento transmisor transfiera su lista de nodos. Esto significa que la adición y la eliminación de nodos de almacenamiento se pueden obtener simplemente añadiendo o eliminando un nodo de almacenamiento y enviando una versión nueva de la lista de nodos a un nodo de almacenamiento individual. Esta lista de nodos se difundirá entonces por la totalidad del resto de nodos de almacenamiento del sistema.

Nuevamente en referencia a la figura 8, cada nodo de almacenamiento busca 61 en su directorio interno archivos que han sido almacenados por el nodo de almacenamiento que funciona de manera deficiente. Los nodos de almacenamiento que almacenan propiamente los archivos "B" y "C" hallarán el nodo de almacenamiento que funciona deficientemente, y por lo tanto pueden añadir el archivo correspondiente en sus listas 55.

No obstante, el proceso de detección puede revelar también otras condiciones que impliquen la necesidad de duplicar un archivo. Típicamente, dichas condiciones pueden ser inconsistencias, es decir, que uno o más nodos de almacenamiento tenga una versión obsoleta del archivo. Una operación de eliminación implica también un proceso de duplicación en la medida en la que este proceso puede llevar a cabo la eliminación física real del archivo. En ese caso, la operación de eliminación del servidor únicamente necesita garantizar que los nodos de almacenamiento activen una bandera de eliminación para el archivo en cuestión. Por lo tanto, cada nodo puede monitorizar operaciones de lectura y escritura llevadas a cabo en el sistema de almacenamiento de datos. La información proporcionada por el servidor 7 en la conclusión de operaciones de lectura y escritura, respectivamente, puede indicar que un nodo de almacenamiento contiene una versión obsoleta de un archivo (en el caso de una operación de lectura) o que un nodo de almacenamiento no llevó a cabo satisfactoriamente una operación de escritura. En ambos casos, se produce una necesidad de mantener datos por duplicación, de tal manera que se cumplen los objetivos globales del proceso de mantenimiento.

Además de las operaciones básicas de lectura y escritura 63, 65, por lo menos dos procesos adicionales pueden proporcionar indicaciones de que existe una necesidad de duplicación, concretamente los procesos de eliminación 67 y actualización 69 sobre los cuales se ofrece a continuación una breve explicación.

El proceso de eliminación es iniciado por el servidor 7 (consúltese la figura 1). De manera similar al proceso de lectura, el servidor envía una consulta por multidifusión a todos los nodos de almacenamiento, con el fin de hallar qué nodos de almacenamiento tienen datos con un identificador específico de datos. Los propios nodos de almacenamiento buscan datos con el identificador pertinente, y responden por medio de una transmisión de unidifusión en caso de que tengan los datos en cuestión. La respuesta puede incluir una lista, del directorio de nodos de almacenamiento, de otros nodos de almacenamiento que contengan los datos. A continuación, el servidor 7 envía a los nodos de almacenamiento que se considera que almacenan el archivo, una solicitud de unidifusión para que el archivo sea eliminado. Cada nodo de almacenamiento activa una bandera referente al archivo y que indica que el mismo se debería eliminar. El archivo se añade entonces a la lista de duplicación, y se envía un acuse de recibo al servidor. A continuación, tal como se describirá, el proceso de duplicación elimina físicamente el archivo.

El proceso de actualización tiene una función de búsqueda, similar a la del proceso de eliminación, y una función de escritura, similar a la llevada a cabo en el proceso de escritura. El servidor envía una consulta por multidifusión a todos los nodos de almacenamiento, para averiguar qué nodos de almacenamiento tienen datos con un identificador específico de datos. Los propios nodos de almacenamiento buscan datos con el identificador pertinente, y responden por medio de una transmisión de unidifusión en caso de que tengan los datos en cuestión. La respuesta puede incluir una lista, del directorio de nodos de almacenamiento, de otros nodos de almacenamiento que contengan los datos. El servidor 7 envía a continuación una solicitud de unidifusión, comunicando a los nodos de almacenamiento que actualicen los datos. Naturalmente, la solicitud contiene los datos actualizados. Los nodos de almacenamiento que actualizan los datos envían un acuse de recibo al servidor, el cual responde enviando una transmisión de unidifusión que contiene una lista con los nodos de almacenamiento que actualizaron satisfactoriamente los datos, y los nodos de almacenamiento que no lo hicieron. Nuevamente, esta lista puede ser usada por el proceso de mantenimiento.

Nuevamente en referencia a la figura 8, las operaciones de lectura 63, escritura 65, eliminación 67 y actualización 69 pueden indicar todas ellas que existe una necesidad de duplicación. Se aplica lo mismo para la monitorización de latidos 59. Por lo tanto, el proceso de detección global 51 genera datos referentes a qué archivos es necesario duplicar. Por ejemplo, una operación de lectura o actualización puede revelar que un nodo de almacenamiento específico contiene una versión obsoleta de un archivo. Un proceso de eliminación puede activar una bandera de eliminación para un archivo específico. La monitorización de latidos puede revelar que es necesario duplicar en un nodo de almacenamiento nuevo una serie de archivos almacenados en un nodo de almacenamiento que funciona deficientemente.

Cada nodo de almacenamiento monitoriza la necesidad de duplicación de todos los archivos que almacena y mantiene una lista de duplicación 55. La lista de duplicación 55 contiene por tanto una serie de archivos que es necesario duplicar. Los archivos se pueden ordenar en correspondencia con la prioridad para cada duplicación. Típicamente, puede haber tres niveles de prioridad diferentes. El nivel más alto se reserva para archivos cuya última copia en línea está contenida en el nodo de almacenamiento. Es necesario duplicar rápidamente dicho archivo en otros nodos de almacenamiento, de tal manera que se pueda alcanzar un nivel razonable de redundancia. Un nivel medio de prioridad puede referirse a archivos donde las versiones son inconsistentes entre los nodos de almacenamiento. Un nivel más bajo de prioridad puede referirse a archivos que están almacenados en un nodo de almacenamiento que está funcionando de manera deficiente.

El nodo de almacenamiento trata los archivos de la lista de duplicación 55 de acuerdo con su nivel de prioridad. A continuación se describe el proceso de duplicación para un nodo de almacenamiento al que en la presente se le denomina nodo de almacenamiento operativo, aunque todos los nodos de almacenamiento pueden funcionar de esta manera.

La parte de duplicación 53 del proceso de mantenimiento comienza con el nodo de almacenamiento operativo intentando 71 convertirse en el nodo maestro para el archivo que desea duplicar. Los nodos de almacenamiento operativos envían una solicitud de unidifusión para convertirse en nodo maestro a otros nodos de almacenamiento que se sabe que almacenan el archivo en cuestión. El directorio 19 (consúltese la figura 1) proporciona una lista de anfitriones que comprende información referente a qué nodos de almacenamiento preguntar. En la circunstancia de que uno de los nodos de almacenamiento no responda afirmativamente, por ejemplo, en caso de una colisión de solicitudes, el archivo se devuelve por el momento a la lista, y en su lugar se realiza un intento con el siguiente archivo de la lista. En cualquier otro caso, se considera que el nodo de almacenamiento operativo es el nodo maestro de este archivo y los otros nodos de almacenamiento activan una bandera que indica que el nodo de almacenamiento operativo es nodo maestro para el archivo en cuestión.

La siguiente etapa consiste en hallar 73 todas las copias del archivo en cuestión en el sistema de almacenamiento distribuido. Esto se puede llevar a cabo por medio del envío, por parte del nodo de almacenamiento operativo, de una consulta de multidifusión a todos los nodos de almacenamiento, preguntando cuáles de ellos tienen el archivo. Los nodos de almacenamiento que tienen el archivo presentan respuestas a la consulta, que contienen la versión del archivo que guardan así como sus listas de anfitriones, es decir, la lista de nodos de almacenamiento que contienen el archivo pertinente que se guarda en el directorio de cada nodo de almacenamiento. Estas listas de anfitriones son a continuación fusionadas 75 por el nodo de almacenamiento operativo, de tal manera que se forma una lista de servidores maestra que se corresponde con la unión de todas las listas de anfitriones recuperadas. Si se hallan nodos de almacenamiento adicionales, a los cuales no se les preguntó cuando el nodo de almacenamiento operativo intentó convertirse en nodo maestro, esa etapa puede repetirse en este momento para los nodos de almacenamiento adicionales. La lista de anfitriones maestra contiene información referente a qué versiones del archivo guardan los diferentes nodos de almacenamiento e ilustran el estado del archivo dentro del sistema de almacenamiento completo.

En caso de que el nodo de almacenamiento operativo no tuviese la última versión del archivo en cuestión, entonces este archivo se recupera 77 de uno de los nodos de almacenamiento que sí tienen la última versión.

A continuación el nodo de almacenamiento operativo decide 79 si es necesario cambiar la lista de anfitriones, típicamente si deberían añadirse nodos de almacenamiento adicionales. En caso afirmativo, el nodo de almacenamiento operativo puede llevar a cabo un proceso muy similar al proceso de escritura que lleva a cabo el servidor y que se describe con relación a las figuras 4A a 4C y 5. El resultado de este proceso es que el archivo se escribe en un nodo de almacenamiento nuevo.

En caso de inconsistencias de la versión, el nodo de almacenamiento operativo puede actualizar, 81, copias del archivo que están almacenadas en otros nodos de almacenamiento, de tal manera que todos los archivos almacenados tengan la versión correcta.

Las copias superfluas del archivo almacenado se pueden eliminar 83. Si el proceso de duplicación se inicia mediante una operación de eliminación, el proceso puede saltar directamente a esta etapa. A continuación, en cuanto todos los nodos de almacenamiento hayan aceptado la eliminación del archivo, el nodo de almacenamiento operativo

simplemente solicita, haciendo uso de la unidifusión, que todos los nodos de almacenamiento eliminen físicamente el archivo en cuestión. Los nodos de almacenamiento acusan recibo de que el archivo se ha eliminado.

Además, se actualiza el estado, es decir, la lista de anfitriones maestra del archivo. A continuación, es posible opcionalmente repetir las etapas 73 a 83 para garantizar que ya no existe necesidad de duplicación. Esta repetición debería dar como resultado una lista de anfitriones maestra consistente que no es necesario actualizar en la etapa 85.

Seguidamente, concluye el proceso de duplicación para ese archivo, y el nodo de almacenamiento operativo puede liberar 87 el estado de nodo maestro del archivo enviando un mensaje correspondiente a la totalidad del resto de nodos de almacenamiento de la lista de anfitriones.

Este sistema en el que cada nodo de almacenamiento asume la responsabilidad de mantener todos los archivos que almacena entre todo el conjunto de nodos de almacenamiento, proporciona un sistema auto-reparable (en caso de funcionamiento deficiente de un nodo de almacenamiento) y auto-limpiable (en caso de inconsistencias de archivos o archivos a eliminar) con una fiabilidad excelente. Es fácilmente escalable y puede almacenar archivos para un número elevado de aplicaciones diferentes simultáneamente.

La invención no queda limitada a los ejemplos específicos dados a conocer y se puede variar y modificar de diferentes maneras dentro del alcance de las reivindicaciones adjuntas.

REIVINDICACIONES

1. Método para mantener datos en un sistema de almacenamiento de datos que comprende una pluralidad de nodos de almacenamiento de datos, utilizándose el método en un nodo de almacenamiento en el sistema de almacenamiento de datos y comprendiendo:

- monitorizar el estado (59) de otros nodos de almacenamiento en el sistema así como operaciones (65, 67, 69) de escritura llevadas a cabo en el sistema de almacenamiento de datos, de manera que el nodo de almacenamiento tiene acceso, para una entidad de datos, a una lista de anfitriones, que incluye nodos de almacenamiento que almacenan la entidad de datos;
- detectar (51), basándose en la monitorización, condiciones en el sistema de almacenamiento de datos, que indican que un nodo de almacenamiento del sistema de almacenamiento de datos está funcionando de manera deficiente;
- determinar entidades de datos almacenadas que fueron almacenadas también por el nodo de almacenamiento que funciona de manera deficiente, sobre la base de la información existente en una pluralidad de listas de anfitriones, en donde la lista de anfitriones para cada entidad de datos comprende una lista asociada que identifica un subconjunto de nodos de almacenamiento dentro del sistema de almacenamiento de datos, que almacenan la entidad de datos; e
- iniciar un proceso (53) de duplicación para entidades de datos que fueron almacenadas por el nodo de almacenamiento que funciona deficientemente en caso de que se detecten las condiciones, en donde el proceso de duplicación incluye enviar un mensaje de multidifusión, a una pluralidad de nodos de almacenamiento, de modo que el mensaje consulta cuáles de dichos nodos de almacenamiento almacenan datos específicos.

2. Método según la reivindicación 1, en el que la monitorización incluye escuchar (59) señales de latido provenientes de otros nodos de almacenamiento en el sistema, y en donde una condición que implica necesidad de duplicación es un nodo de almacenamiento que funciona de manera deficiente.

3. Método según la reivindicación 1 ó 2, en el que los datos incluyen archivos y una condición es una de entre una eliminación de archivo o una inconsistencia de archivo.

4. Método según cualquiera de las reivindicaciones anteriores, en el que se mantiene una lista de duplicación, que incluye archivos que requieren duplicación.

5. Método según la reivindicación 4, en el que la lista de duplicación incluye prioridades.

6. Método según cualquiera de las reivindicaciones anteriores, en el que el proceso de duplicación incluye además:

- recibir respuestas de aquellos nodos de almacenamiento que contienen dichos datos específicos;
- determinar si dichos datos específicos están almacenados en un número suficiente de nodos de almacenamiento; y
- si no, seleccionar por lo menos un nodo de almacenamiento adicional y transmitir dichos datos específicos a ese nodo de almacenamiento.

7. Método según la reivindicación 6, que comprende además actualizar dichos datos específicos en nodos de almacenamiento que contienen versiones obsoletas de los mismos.

8. Método según la reivindicación 6 ó 7, en el que el proceso de duplicación comienza con el intento, por parte del nodo de almacenamiento, de obtener dominio del archivo a duplicar entre todos los nodos de almacenamiento del sistema.

9. Método según cualquiera de las reivindicaciones anteriores, en el que la monitorización incluye además monitorizar operaciones (63) de lectura llevadas a cabo en el sistema de almacenamiento de datos.

10. Nodo de almacenamiento de datos para mantener datos en un sistema de almacenamiento de datos que comprende una pluralidad de nodos de almacenamiento de datos, comprendiendo el nodo de almacenamiento de datos:

- medios para monitorizar el estado de otros nodos de almacenamiento en el sistema así como operaciones de escritura llevadas a cabo en el sistema de almacenamiento de datos, en donde la monitorización para una entidad de datos está basada en una lista de anfitriones, que incluye nodos de almacenamiento que almacenan la entidad de datos;

- medios para detectar, basándose en la monitorización, condiciones en el sistema de almacenamiento de datos que indican que un nodo de almacenamiento del sistema de almacenamiento de datos está funcionando de manera deficiente;
- medios para determinar entidades de datos almacenadas que fueron almacenadas también por el nodo de almacenamiento que funciona de manera deficiente, sobre la base de la información existente en una pluralidad de listas de anfitriones, en donde la lista de anfitriones para cada entidad de datos comprende una lista asociada que identifica un subconjunto de nodos de almacenamiento dentro del sistema de almacenamiento de datos, que almacenan la entidad de datos; y
- medios para iniciar un proceso de duplicación para entidades de datos que fueron almacenadas por el nodo de almacenamiento que funciona deficientemente en caso de que se detecten las condiciones, en donde el proceso de duplicación incluye enviar un mensaje de multidifusión, a una pluralidad de nodos de almacenamiento, de modo que el mensaje consulta cuáles de dichos nodos de almacenamiento almacenan datos específicos.

11. Método para escribir datos en un sistema de almacenamiento de datos que comprende una pluralidad de nodos de almacenamiento de datos, utilizándose el método en un servidor que ejecuta una aplicación que accede a datos en el sistema de almacenamiento de datos, y comprendiendo:

- enviar (41) una consulta de almacenamiento de multidifusión a una pluralidad de dichos nodos de almacenamiento;
- recibir (43) una pluralidad de respuestas de un subconjunto de dichos nodos de almacenamiento, incluyendo las respuestas datos geográficos referentes a la posición geográfica de cada nodo de almacenamiento;
- seleccionar (45) por lo menos dos nodos de almacenamiento del subconjunto, sobre la base de dichas respuestas; y
- enviar (47) a los nodos de almacenamiento seleccionados datos, un identificador de datos, correspondiente a los datos, y una lista de anfitriones, que es una lista de nodos de almacenamiento que almacenan satisfactoriamente los datos.

12. Método según la reivindicación 11, en el que la posición geográfica incluye latitud y longitud del nodo de almacenamiento en cuestión.

13. Método según la reivindicación 12, en el que las respuestas incluyen además antigüedad del sistema para el nodo de almacenamiento en cuestión.

14. Método según la reivindicación 12 ó 13, en el que las respuestas incluyen además carga del sistema para el nodo de almacenamiento en cuestión.

15. Método según cualquiera de las reivindicaciones 12 a 14, en el que la consulta de almacenamiento de multidifusión incluye un identificador de datos, que identifica los datos a almacenar.

16. Método según cualquiera de las reivindicaciones 12 a 15, en el que se seleccionan por lo menos tres nodos.

17. Servidor adaptado para escribir datos en un sistema de almacenamiento de datos que comprende una pluralidad de nodos de almacenamiento de datos, comprendiendo el servidor:

- medios para enviar una consulta de almacenamiento de multidifusión a una pluralidad de dichos nodos de almacenamiento;
- medios para recibir una pluralidad de respuestas de un subconjunto de dichos nodos de almacenamiento, incluyendo las respuestas datos geográficos referentes a la posición geográfica de cada nodo de almacenamiento;
- medios para seleccionar por lo menos dos nodos de almacenamiento del subconjunto, sobre la base de dichas respuestas; y
- medios para enviar a los nodos de almacenamiento seleccionados datos y un identificador de datos, correspondiente a los datos, y una lista de anfitriones, que es una lista de nodos de almacenamiento que almacenan satisfactoriamente los datos.

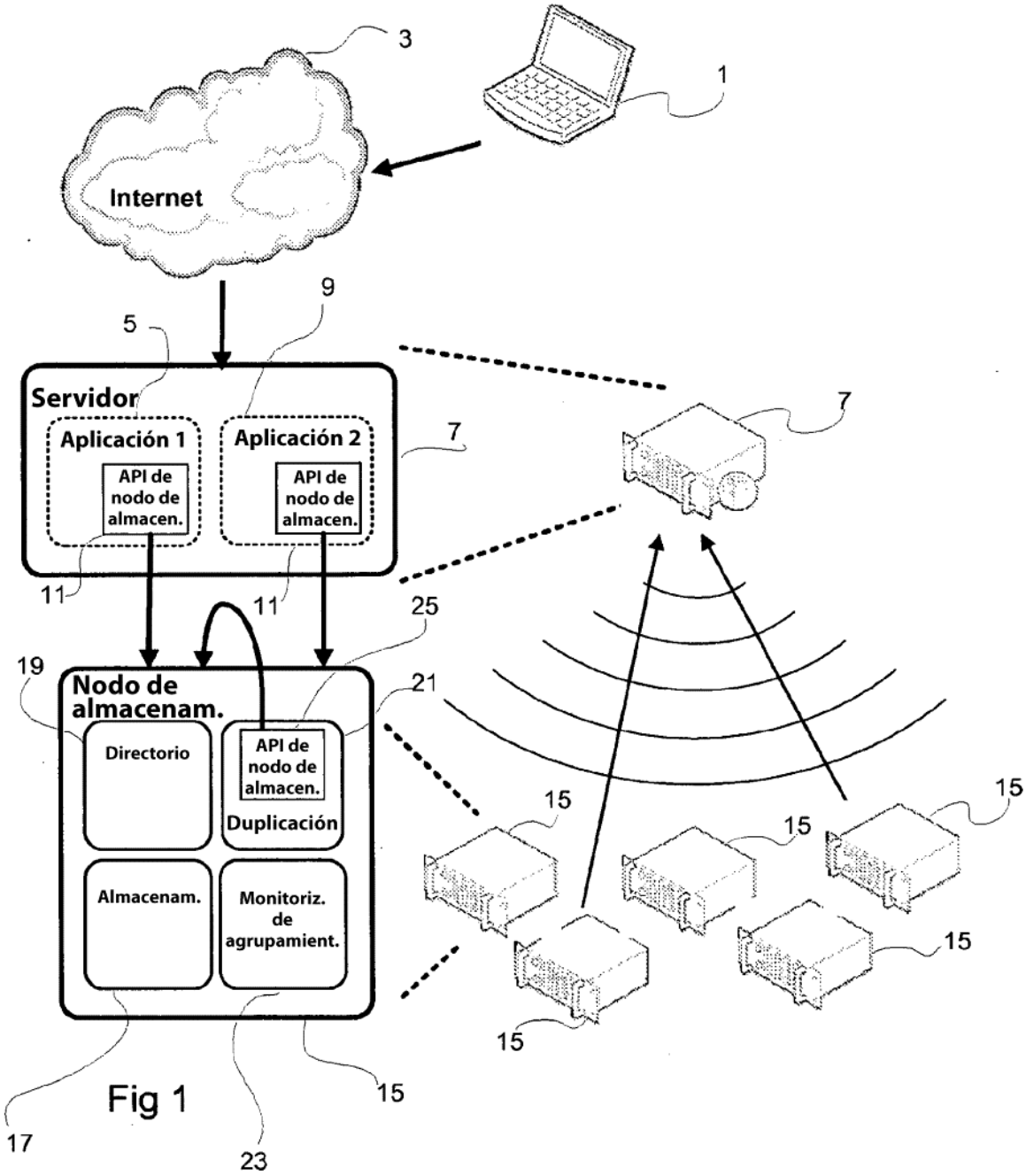
3
1
Internet
9
5
Servidor
Aplicación 1
API de nodo de almacen.
Aplicación 2
API de nodo de almacen.
7
7
11
11
25
19
Nodo de almacenam.
21
Directorio
API de nodo de almacen.
Duplicación
Almacenam.
Monitoriz. de agrupamient.
15
15
15
15
15
15
17
23

Fig 1

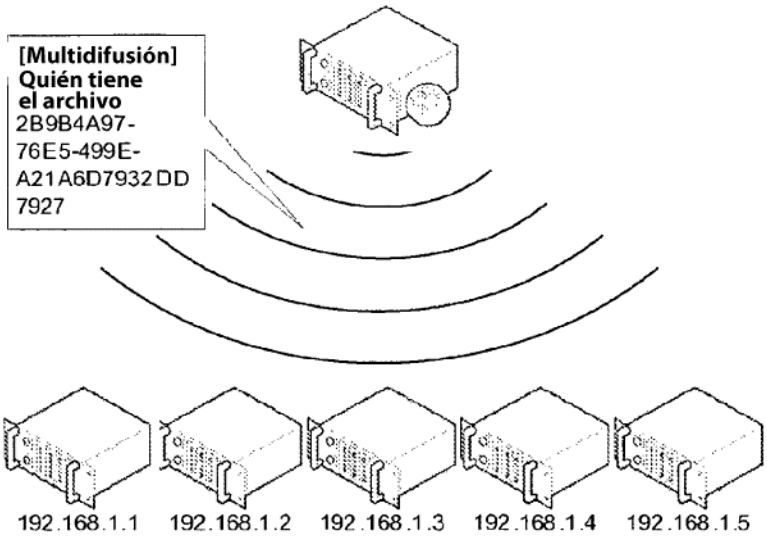

Fig 2A

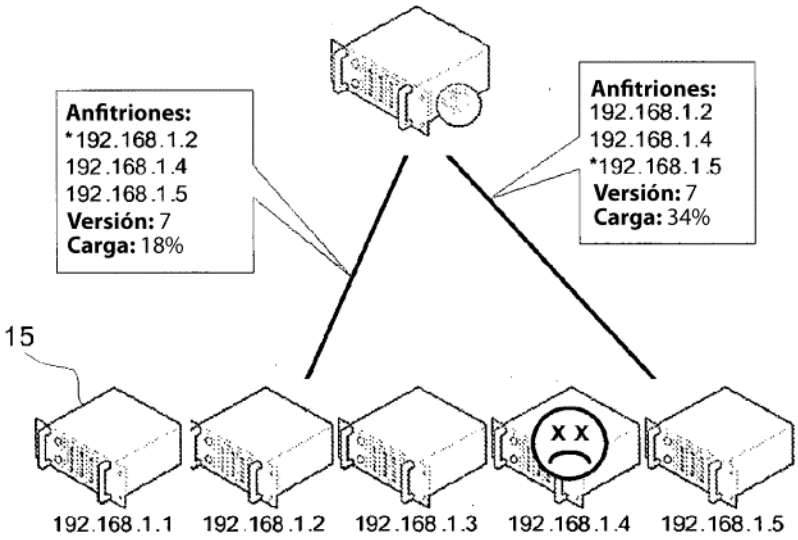

Fig 2B

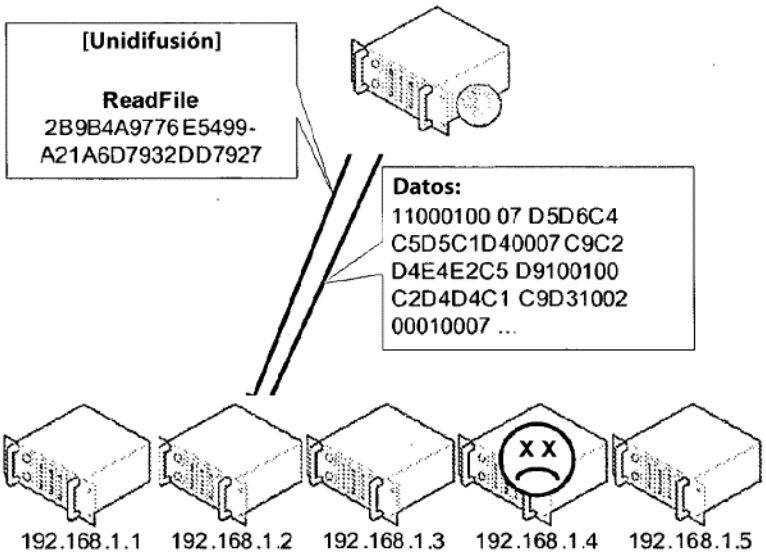

Fig 2C

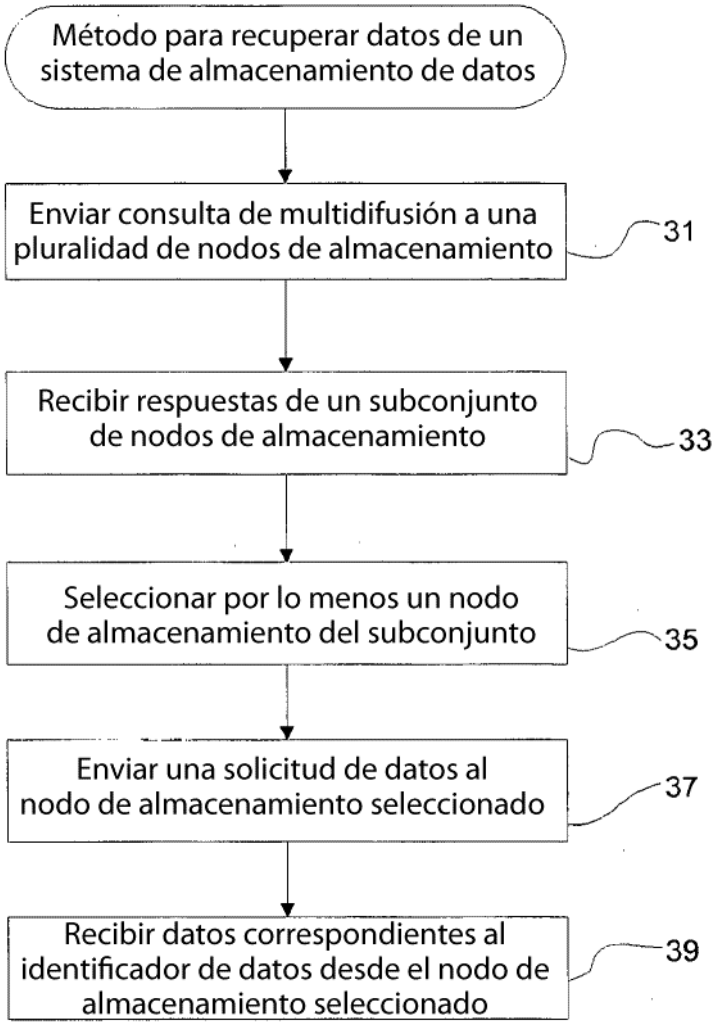
Método para recuperar datos de un sistema de almacenamiento de datos
Enviar consulta de multidifusión a una pluralidad de nodos de almacenamiento
31
Recibir respuestas de un subconjunto de nodos de almacenamiento
33
Seleccionar por lo menos un nodo de almacenamiento del subconjunto
35
Enviar una solicitud de datos al nodo de almacenamiento seleccionado
37
Recibir datos correspondientes al identificador de datos desde el nodo de almacenamiento seleccionado
39

Fig 3

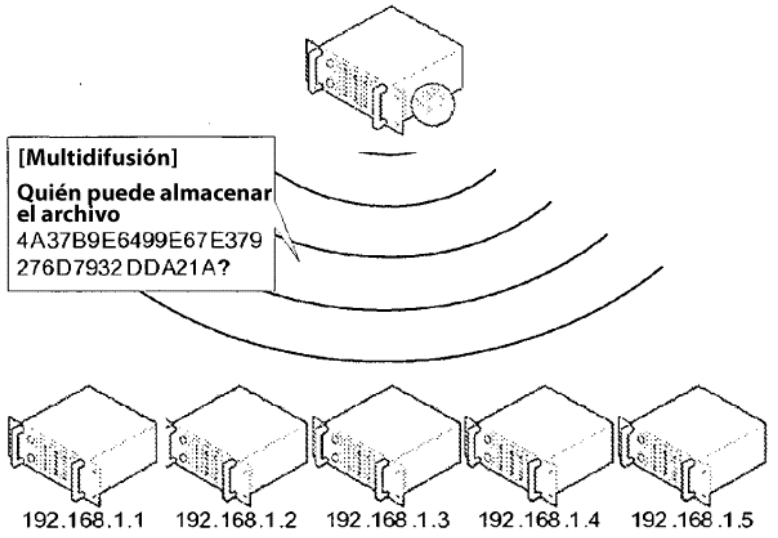

Fig 4A

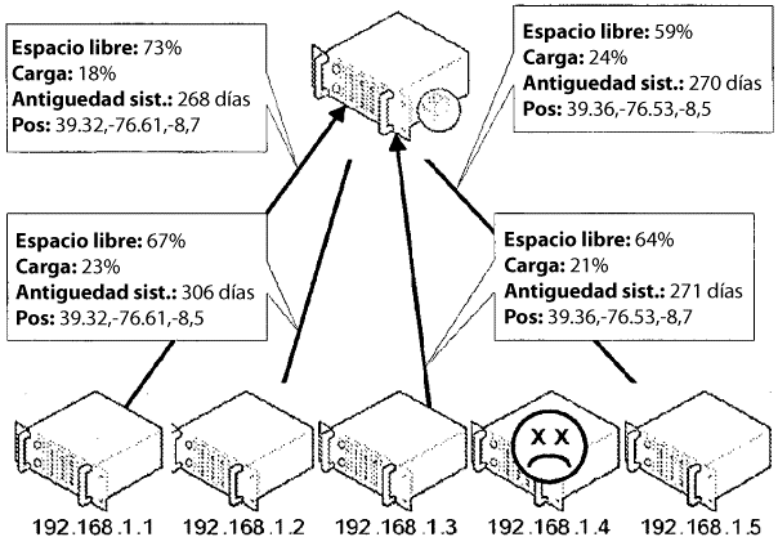

Fig 4B

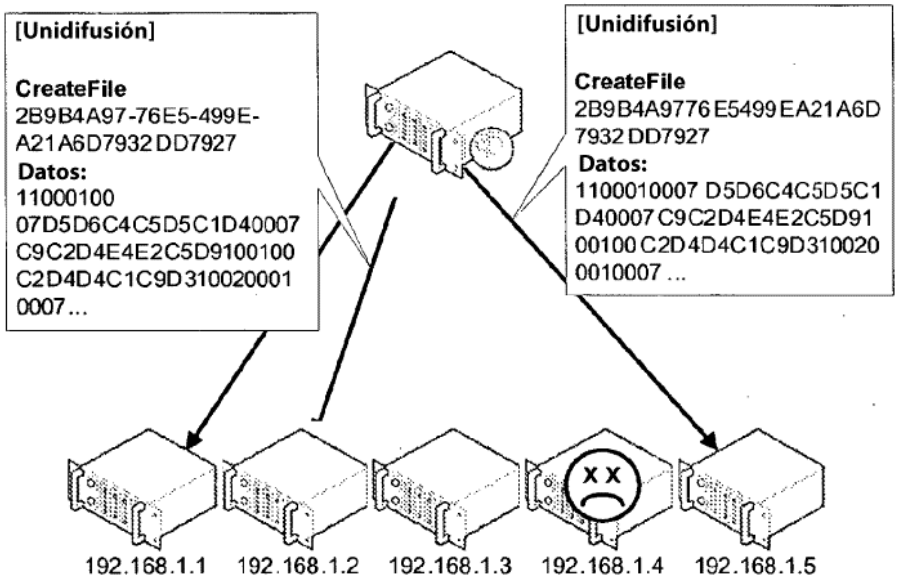

Fig 4C

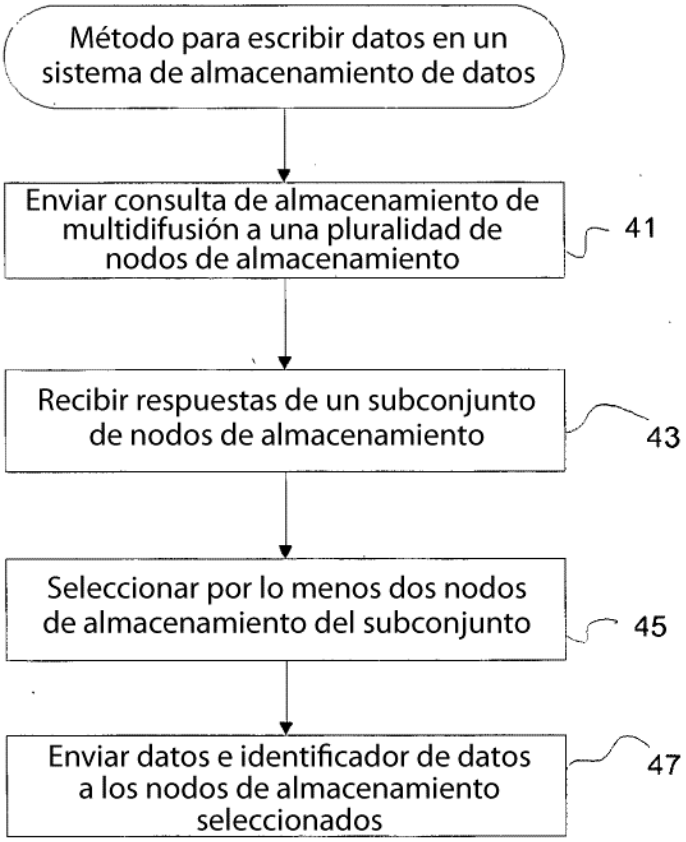
Método para escribir datos en un sistema de almacenamiento de datos
Enviar consulta de almacenamiento de multidifusión a una pluralidad de nodos de almacenamiento
41
Recibir respuestas de un subconjunto de nodos de almacenamiento
43
Seleccionar por lo menos dos nodos de almacenamiento del subconjunto
45
Enviar datos e identificador de datos a los nodos de almacenamiento seleccionados
47

Fig 5

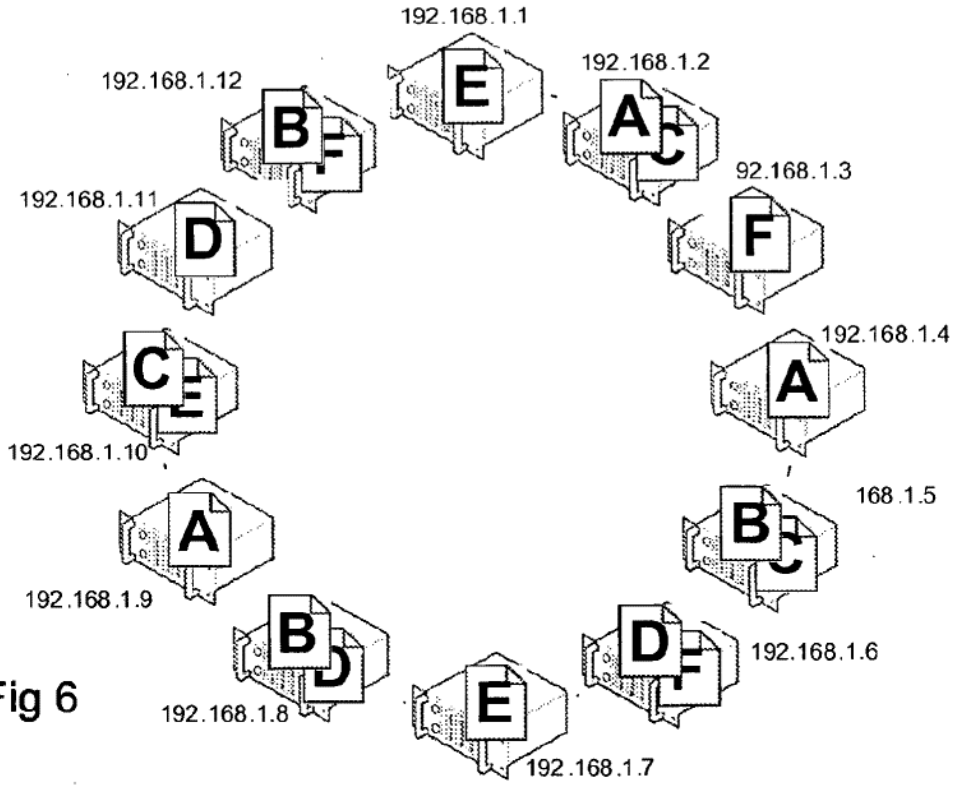
192.168.1.1
192.168.1.2
192.168.1.12
92.168.1.3
192.168.1.11
192.168.1.4
192.168.1.10
168.1.5
192.168.1.9
192.168.1.6
192.168.1.8
192.168.1.7

Fig 6

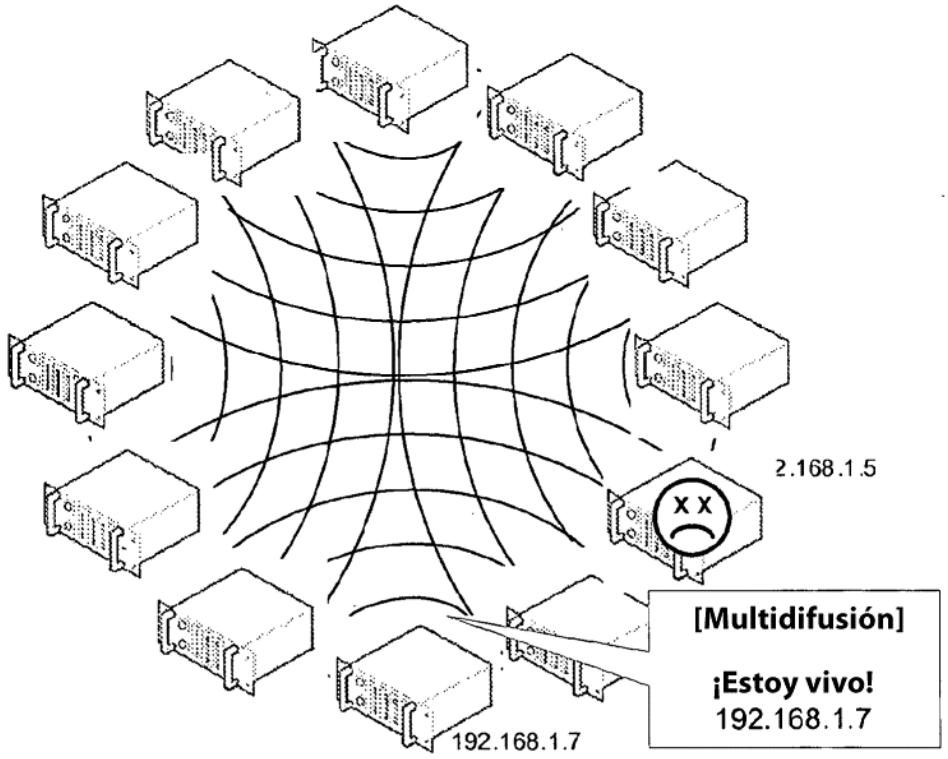
2.168.1.5
[Multidifusión]
¡Estoy vivo!
192.168.1.7

Fig 7

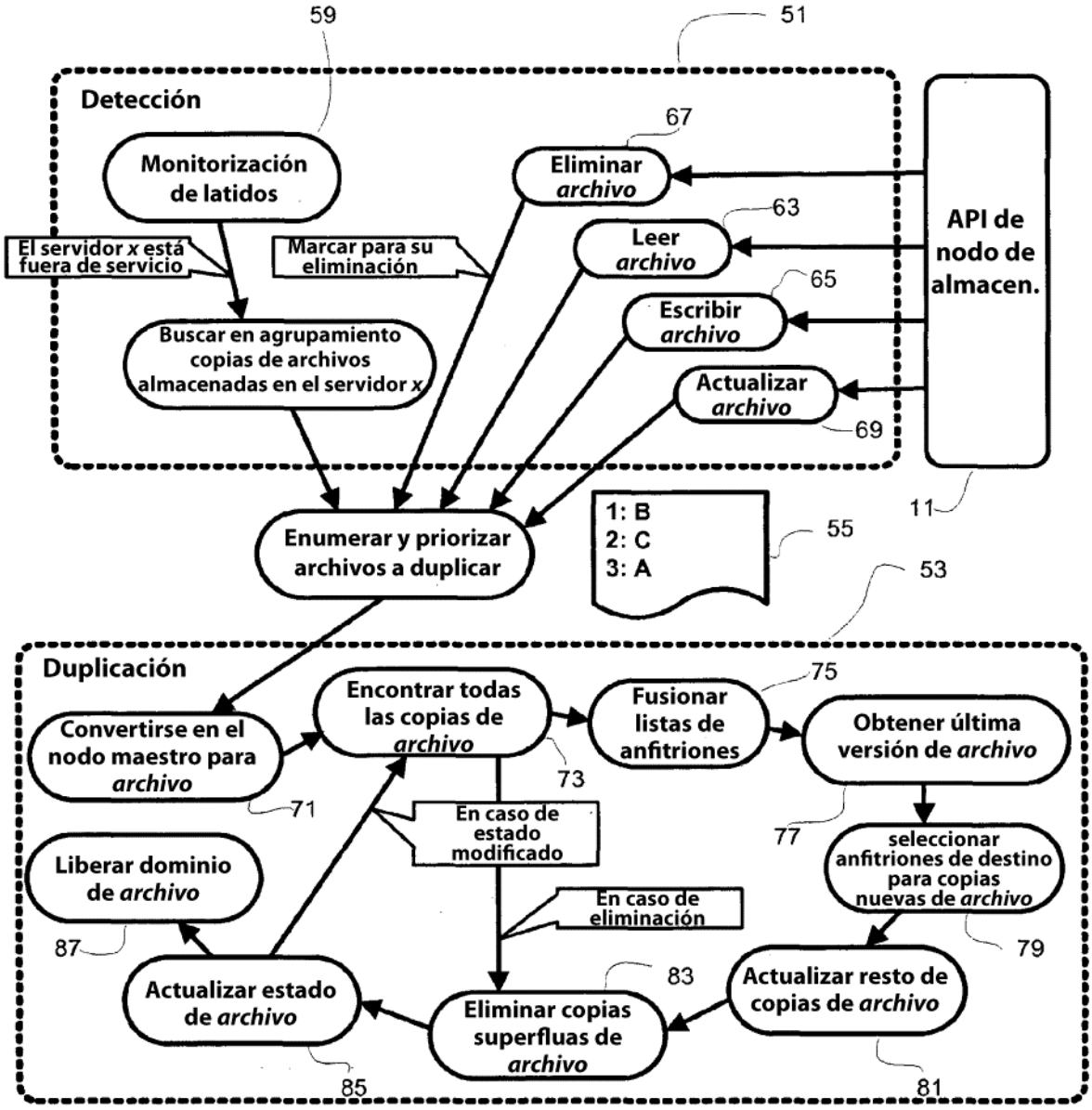
59
51
Detección
Monitorización de latidos
El servidor x está fuera de servicio
Buscar en agrupamiento copias de archivos almacenadas en el servidor x
Marcar para su eliminación
67
Eliminar archivo
63
Leer archivo
65
Escribir archivo
Actualizar archivo
69
API de nodo de almacen.
11
Enumerar y priorizar archivos a duplicar
1: B
2: C
3: A
55
53
Duplicación
Convertirse en el nodo maestro para archivo
71
Encontrar todas las copias de archivo
73
Fusionar listas de anfitriones
75
Obtener última versión de archivo
77
seleccionar anfitriones de destino para copias nuevas de archivo
79
Actualizar resto de copias de archivo
81
Eliminar copias superfluas de archivo
83
Actualizar estado de archivo
85
Liberar dominio de archivo
87
En caso de estado modificado
En caso de eliminación

Fig 8