(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 624 412**

(51) Int. CI.:

H04L 29/08 (2006.01)
G06F 3/06 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **31.12.2013 PCT/CN2013/091253**

(87) Fecha y número de publicación internacional: **03.07.2014 WO14101896**

(96) Fecha de presentación y número de la solicitud europea: **31.12.2013 E 13869766 (9)**

(97) Fecha y número de publicación de la concesión europea: **22.02.2017 EP 2930910**

(54) Título: **Procedimiento y sistema para compartir recursos de almacenamiento**

(30) Prioridad:

31.12.2012 WO PCT/CN2012/088109

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
14.07.2017

(73) Titular/es:

HUAWEI TECHNOLOGIES CO., LTD. (100.0%)
Huawei Administration Building, Bantian Longgang District , Shenzhen, Guangdong 518129, CN

(72) Inventor/es:

GU, JIONGJIONG;
MIN, XIAOYONG y
WANG, DAOHUI

(74) Agente/Representante:

LEHMANN NOVO, María Isabel

DESCRIPCIÓN

Procedimiento y sistema para compartir recursos de almacenamiento

SECTOR TÉCNICO

La presente invención se refiere al sector de las tecnologías de comunicaciones, y en particular, a un procedimiento y un sistema para compartir un recurso de almacenamiento.

ANTECEDENTES

En una aplicación informática en la nube, un sistema de grupo de servidores integra recursos informáticos, recursos de almacenamiento y recursos de red, y proporciona, utilizando una tecnología tal como virtualización y por medio de una red, los recursos a los usuarios para su utilización. Una forma de aplicación es, por ejemplo, una máquina virtual (Virtual Machine, "VM" para abreviar), el alquiler de la capacidad informática, el alquiler de la capacidad de almacenamiento o similares.

Actualmente, por razones tales como los diferentes tipos de demandas de recursos, un sistema de grupo de servidores utiliza generalmente diferentes dispositivos para proporcionar dispositivos de almacenamiento, y las fuentes de recursos de almacenamiento están diversificadas, por ejemplo, un recurso de almacenamiento incorporado en un nodo servidor, y un recurso de almacenamiento desplegado independientemente, donde el recurso de almacenamiento desplegado independientemente puede ser, por ejemplo, una matriz de almacenamiento o un servidor de almacenamiento dedicados, tal como una red de área de almacenamiento (Storage Area Network, "SAN" por brevedad).

En la técnica anterior, los dispositivos de almacenamiento de un sistema de grupo de servidores proporcionan independientemente servicios de almacenamiento externos, lo que tiene como resultado una baja utilización combinada de los recursos de almacenamiento. Además, los recursos de almacenamiento de dispositivo de almacenamiento en red, acaparados primitivamente por empresas, no pueden ser utilizados por el sistema de grupo de servidores, provocando un tremendo derroche.

El documento US 2011/087833 A1 se refiere a un servidor de datos, a un sistema de adaptador de anfitrión para el servidor de datos y a procedimientos de funcionamiento relacionados, que facilitan las operaciones de escritura y lectura de datos para un almacenamiento de datos basado en red que está acoplado remotamente al servidor de datos. El sistema de adaptador de anfitrión incluye un módulo de controlador de almacenamiento local y un módulo de controlador de almacenamiento en red. El módulo de controlador de almacenamiento local se utiliza para un dispositivo de memoria caché, no volátil, de escritura simultánea, conectado localmente, del servidor de datos. El módulo de controlador de almacenamiento en red se utiliza para una arquitectura de almacenamiento de datos basada en red, del servidor de datos. Los módulos de controlador de almacenamiento soportan escritura simultáneamente de datos en el almacenamiento en memoria caché local y la arquitectura de almacenamiento basada en red. Los módulos de controlador de almacenamiento soportan asimismo la lectura de datos mantenidos por servidor desde el almacenamiento en memoria caché local y la arquitectura de almacenamiento basado en red.

RESUMEN

Las realizaciones de la presente invención dan a conocer un procedimiento y un sistema para compartir un recurso de almacenamiento, que se utilizan para integrar, compartir y utilizar recursos de almacenamiento heterogéneos, mejorando de ese modo la utilización de los recursos de almacenamiento.

De acuerdo con un primer aspecto, una realización de la presente invención da a conocer un procedimiento para compartir un recurso de almacenamiento, aplicado a un sistema de grupo de servidores, donde el sistema de grupo de servidores incluye un nodo servidor y un nodo de almacenamiento en red, el nodo servidor incluye un disco duro, el nodo de almacenamiento en red incluye una matriz de almacenamiento, un controlador de almacenamiento distribuido se ejecuta en el nodo servidor, y el controlador de almacenamiento distribuido comprende un controlador de metadatos, un módulo de servicio de bloque virtual y un módulo de control de lectura-escritura, el módulo de control de lectura-escritura comprende un delegado de almacenamiento de objetos y un agente de almacenamiento SAN, y el procedimiento incluye:

determinar, mediante el controlador de metadatos, estados de despliegue del delegado de almacenamiento de objetos y del agente de almacenamiento SAN en el nodo servidor; generar información de visualización del módulo de control de lectura-escritura de acuerdo con los estados de despliegue determinados, y suministrar la información de visualización del módulo de control de lectura-escritura al módulo de servicio de bloque virtual, la información de visualización de dichos por lo menos dos módulos de control de lectura-escritura se utiliza para indicar información de encaminamiento a cada módulo de control de lectura-escritura;

dividir, mediante el controlador de metadatos, los recursos de almacenamiento del disco duro y la matriz de almacenamiento en múltiples particiones de almacenamiento, donde las múltiples particiones de almacenamiento forman un conjunto de recursos de almacenamiento compartidos;

asignar, mediante el controlador de metadatos, un módulo de control de lectura-escritura a cada partición de almacenamiento;

generar, mediante el controlador de metadatos, información de partición global y suministrar la información de partición global al módulo de servicio de bloque virtual, donde la información de partición global registra una correspondencia entre cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos y el módulo de control de lectura-escritura asignado, de cada partición de almacenamiento;

recibir, mediante el módulo de servicio de bloque virtual, un mensaje de solicitud de almacenamiento, y determinar una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con la información de partición global, un módulo de control de lectura-escritura que es correspondiente a la partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento; determinar, mediante el módulo de servicio de bloque virtual, información de encaminamiento del módulo de control de lectura-escritura determinado, de acuerdo con la información de visualización del módulo de control de lectura-escritura; y

enviar, mediante el módulo de servicio de bloque virtual, el mensaje de solicitud de almacenamiento al módulo de control de lectura-escritura determinado, de acuerdo con la información de encaminamiento del módulo de control de lectura-escritura determinado, de tal modo que el módulo de control de lectura-escritura determinado lleva a cabo una operación solicitada por el mensaje de solicitud de almacenamiento.

Haciendo referencia al primer aspecto, en un primer posible modo de implementación, la determinación de una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento incluye:

determinar, mediante el módulo de servicio de bloque virtual, un ID de un volumen de usuario en el que están localizados los datos a gestionar de acuerdo con el mensaje de solicitud de almacenamiento y una dirección de bloque lógico LBA de por lo menos un bloque de datos de los datos a gestionar; y

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, una partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

Haciendo referencia al primer posible modo de implementación del primer aspecto, en un segundo posible modo de implementación, el procedimiento incluye además:

configurar, mediante el módulo de servicio de bloque virtual, metadatos de cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos, donde los metadatos de cada partición de almacenamiento registran una correspondencia entre un ID de la partición de almacenamiento y un ID de cada bloque de datos asignado a la partición de almacenamiento, donde

la determinación, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, de una partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos, incluye:

determinar, mediante el módulo de servicio de bloque virtual, un ID de dicho por lo menos un bloque de datos de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, consultar los metadatos de cada partición de almacenamiento y determinar un ID de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos

Haciendo referencia al primer aspecto o con el primer posible modo de implementación del primer aspecto, en un tercer posible modo de implementación, la determinación, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, de una partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos, incluye:

formar un valor de clave de cada bloque de datos utilizando el ID del volumen de usuario y una LBA de cada bloque de datos, calcular un valor valor correspondiente al valor de clave de cada bloque de datos y determinar, de acuerdo con el valor valor, una partición de almacenamiento correspondiente a cada bloque de datos.

Haciendo referencia al primer posible modo de implementación del primer aspecto, al segundo posible modo de implementación del primer aspecto o al tercer posible modo de implementación del tercer aspecto, en un cuarto posible modo de implementación, la recepción, mediante el módulo de servicio de bloque virtual, de un mensaje de solicitud de almacenamiento incluye: recibir un comando de creación de volumen de usuario, donde el comando de creación de volumen de usuario indica el tamaño del volumen de usuario, donde

la determinación, mediante el módulo de servicio de bloque virtual, de un ID de un volumen de usuario en el que están situados los datos a gestionar de acuerdo con el mensaje de solicitud de almacenamiento y de una dirección de bloque lógico LBA de por lo menos un bloque de datos de los datos a gestionar, y la determinación, de acuerdo

con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos, incluye:

asignar, mediante el módulo de servicio de bloque virtual, el ID del volumen de usuario al volumen de usuario;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el tamaño del volumen de usuario, el tamaño de un recurso de almacenamiento inicial asignado al volumen de usuario, y determinar la LBA de dicho por lo menos un bloque de datos de acuerdo con el tamaño del recurso de almacenamiento inicial; y

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

Haciendo referencia al primer posible modo de implementación del primer aspecto, al segundo posible modo de implementación del primer aspecto o al tercer posible modo de implementación del primer aspecto, en un quinto posible modo de implementación, la recepción, mediante el módulo de servicio de bloque virtual, de un mensaje de solicitud de almacenamiento incluye: recibir, mediante el módulo de servicio de bloque virtual, una solicitud de operación de escritura;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con un nombre de archivo contenido en la solicitud de operación de escritura de datos, un ID de un volumen de usuario correspondiente a una operación de escritura actual;

dividir, mediante el módulo de servicio de bloque virtual, los datos a escribir en múltiples bloques de datos a escribir, y asignar una LBA a cada bloque de datos a escribir;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario correspondiente a la operación de escritura actual, y con la LBA de cada bloque de datos a escribir, la partición de almacenamiento correspondiente a cada bloque de datos a escribir;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con la información de partición global, el módulo de control de lectura-escritura que es correspondiente a la partición de almacenamiento correspondiente a cada bloque de datos a escribir;

generar, mediante el módulo de servicio de bloque virtual, comandos de escritura de bloque de datos, donde cada comando de escritura de bloque de datos es correspondiente a un bloque de datos a escribir, y cada comando de escritura de bloque de datos lleva el bloque de datos a escribir y un ID de los datos a escribir; y

enviar por separado cada comando de escritura de bloque de datos al módulo de control de lectura-escritura determinado correspondiente a cada bloque de datos a escribir, de tal modo que el módulo de control de lectura-escritura determinado correspondiente a cada bloque de datos a escribir escribe cada bloque de datos a escribir en un recurso de hardware de almacenamiento.

Haciendo referencia al primer posible modo de implementación del primer aspecto, al segundo posible modo de implementación del primer aspecto o al tercer posible modo de implementación del primer aspecto, en un quinto posible modo de implementación, la recepción, mediante el módulo de servicio de bloque virtual, de un mensaje de solicitud de almacenamiento incluye: recibir, mediante el módulo de servicio de bloque virtual, una solicitud de operación de lectura de datos, donde la solicitud de operación de lectura de datos lleva un nombre de archivo y un desplazamiento de los datos a leer;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el nombre de archivo contenido en la solicitud de operación de lectura de datos, un ID de un volumen de usuario correspondiente a una operación de lectura actual;

determinar, mediante el módulo de servicio de bloque virtual, LBAs de múltiples bloques de datos a leer, de acuerdo con información de desplazamiento de los datos a leer;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario correspondiente a la operación de lectura actual, y con la LBA de cada bloque de datos a leer, la partición de almacenamiento correspondiente a cada bloque de datos a leer;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con la información de partición global, módulos de control de lectura-escritura que son correspondientes a las particiones de almacenamiento correspondientes a los múltiples bloques de datos a leer;

generar, mediante el módulo de servicio de bloque virtual, múltiples comandos de lectura de bloque de datos, donde cada comando de lectura de bloque de datos es correspondiente a un bloque de datos a leer, y cada comando de lectura de bloque de datos lleva el bloque de datos a leer y un ID del bloque de datos a leer; y

enviar por separado, mediante el módulo de servicio de bloque virtual, cada comando de lectura de bloque de datos al módulo de control de lectura-escritura correspondiente a cada bloque de datos a leer, de tal modo que el módulo de control de lectura-escritura determinado correspondiente a cada bloque de datos a leer lee cada bloque de datos a leer.

De acuerdo con un segundo aspecto, una realización de la presente invención da a conocer un sistema de grupo de servidores, donde el sistema de grupo de servidores incluye un nodo servidor y un nodo de almacenamiento en red, el nodo servidor incluye un disco duro, el nodo de almacenamiento en red incluye una matriz de almacenamiento, se ejecuta un controlador de almacenamiento distribuido en el nodo servidor, y el controlador de almacenamiento distribuido incluye un controlador de metadatos, un módulo de servicio de bloque virtual y un módulo de control de lectura-escritura, comprendiendo el módulo de control de lectura-escritura un delegado de almacenamiento de objetos y un agente de almacenamiento SAN;

en el que

el controlador de metadatos, configurado para: determinar estados de despliegue del delegado de almacenamiento de objetos y del agente de almacenamiento SAN en el nodo servidor; generar información de visualización del módulo de control de lectura-escritura de acuerdo con los estados de despliegue determinados y suministrar la información de visualización del módulo de control de lectura-escritura al módulo de servicio de bloque virtual, utilizándose la información de visualización del módulo de control de lectura-escritura para indicar información de encaminamiento de cada módulo de control de lectura-escritura;

el controlador de metadatos está configurado además para dividir: recursos de almacenamiento del disco duro y de la matriz de almacenamiento en múltiples particiones de almacenamiento, donde las múltiples particiones de almacenamiento forman un conjunto de recursos de almacenamiento compartidos; asignar un módulo de control de lectura-escritura a cada partición de almacenamiento; generar información de partición global, donde la información de partición global registra una correspondencia entre cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos y el módulo de control de lectura-escritura asignado, de cada partición de almacenamiento; y suministrar la información de partición global a un módulo de servicio de bloque virtual;

el módulo de servicio de bloque virtual, configurado para: estar frente a una capa de servicio, y recibir un mensaje de solicitud de almacenamiento; determinar una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento; determinar, de acuerdo con la información de partición global, un módulo de control de lectura-escritura que es correspondiente a la partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento; determinar información de encaminamiento del módulo de control de lectura-escritura determinado, de acuerdo con la información de visualización del módulo de control de lectura-escritura, y enviar el mensaje de solicitud de almacenamiento al módulo de control de lectura-escritura determinado; y

el módulo de control de lectura-escritura, configurado para estar frente al disco duro o al nodo de almacenamiento en red, y llevar a cabo una operación solicitada por el mensaje de solicitud de almacenamiento.

Haciendo referencia al segundo aspecto, en un primer posible modo de implementación, donde:

el controlador de metadatos está configurado específicamente para asignar el delegado de almacenamiento de objetos a una partición de almacenamiento del disco duro local, como un módulo de control de lectura-escritura, y asignar el agente de almacenamiento SAN a una partición de almacenamiento de la matriz de almacenamiento, como un módulo de control de lectura-escritura;

el delegado de almacenamiento de objetos está configurado para: recibir el mensaje de solicitud de almacenamiento, determinar una dirección física correspondiente al mensaje de solicitud de almacenamiento, y llevar a cabo, en función de la dirección física, la operación solicitada por el mensaje de solicitud de almacenamiento en el disco duro; y

el agente de almacenamiento SAN está configurado para: recibir el mensaje de solicitud de almacenamiento, determinar una dirección lógica que es del nodo de almacenamiento en red y corresponde al mensaje de solicitud de almacenamiento, y llevar a cabo, de acuerdo con la dirección lógica, la operación solicitada por el mensaje de solicitud de almacenamiento en la matriz de almacenamiento.

Haciendo referencia al segundo aspecto, en un segundo posible modo de implementación, el módulo de servicio de bloque virtual está configurado específicamente para determinar un ID de un volumen de usuario en el que están situados datos a gestionar de acuerdo con el mensaje de solicitud de almacenamiento y una dirección de bloque lógico LBA de por lo menos un bloque de datos de los datos a gestionar, y determinar, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, una partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

Haciendo referencia al segundo posible modo de implementación del segundo aspecto, en un tercer posible modo de implementación, el módulo de servicio de bloque virtual está configurado específicamente para: configurar metadatos de cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos, donde los metadatos de cada partición de almacenamiento registran una correspondencia entre un ID del partición de

almacenamiento y un ID de un bloque de datos asignado a la partición de almacenamiento; determinar un ID de dicho por lo menos un bloque de datos de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos; consultar los metadatos de cada partición de almacenamiento; y determinar un ID de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

Haciendo referencia al segundo posible modo de implementación del segundo aspecto, en un cuarto posible modo de implementación, el módulo de servicio de bloque virtual está configurado específicamente para formar un valor de clave de cada bloque de datos utilizando el ID del volumen de usuario y una LBA de cada bloque de datos, calcular un valor valor correspondiente al valor de clave de cada bloque de datos y determinar, de acuerdo con el valor valor, una partición de almacenamiento correspondiente a cada bloque de datos.

Haciendo referencia al segundo aspecto y a cualquiera de los posibles modos de implementación del segundo aspecto, en un quinto posible modo de implementación, el módulo de servicio de bloque virtual está configurado específicamente para: recibir un comando de creación de volumen de usuario, donde el comando de creación de volumen de usuario indica el tamaño del volumen de usuario; asignar el ID del volumen de usuario al volumen de usuario; determinar, de acuerdo con el tamaño del volumen de usuario, el tamaño de un recurso de almacenamiento inicial asignado al volumen de usuario, y determinar la LBA de dicho por lo menos un bloque de datos de acuerdo con el tamaño del recurso de almacenamiento inicial; y determinar, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

Haciendo referencia al segundo aspecto y a cualquiera de los posibles modos de implementación del segundo aspecto, en un sexto posible modo de implementación, el módulo de servicio de bloque virtual está configurado específicamente para: recibir una solicitud de operación de escritura de datos; determinar, de acuerdo con un nombre de archivo contenido en la solicitud de operación de escritura de datos, un ID de un volumen de usuario correspondiente a una operación de escritura actual; dividir los datos a escribir en múltiples bloques de datos a escribir, y asignar una LBA a cada bloque de datos a escribir; determinar, de acuerdo con el ID del volumen de usuario correspondiente a la operación de escritura actual, y a la LBA de cada bloque de datos a escribir, una partición de almacenamiento correspondiente a cada bloque de datos a escribir; determinar, de acuerdo con la información de partición global, un módulo de control de lectura-escritura que es correspondiente a la partición de almacenamiento correspondiente a cada bloque de datos a escribir; generar múltiples comandos de escritura de bloque de datos, donde cada comando de escritura de bloque de datos es correspondiente a un bloque de datos a escribir, y cada comando de escritura de bloque de datos lleva el bloque de datos a escribir y un ID de los datos a escribir, y enviar por separado cada comando de escritura de bloque de datos al módulo de control de lectura-escritura correspondiente a cada bloque de datos a escribir.

Haciendo referencia al segundo aspecto y a cualquiera de los posibles modos de implementación del segundo aspecto, en un séptimo posible modo de implementación, el módulo de servicio de bloque virtual está configurado específicamente para: recibir una solicitud de operación de lectura de datos, donde la solicitud de operación de lectura de datos lleva un nombre de archivo y un desplazamiento de los datos a leer; determinar, de acuerdo con el nombre de archivo contenido en la solicitud de operación de lectura de datos, un ID de un volumen de usuario correspondiente a una actual operación de lectura; determinar LBAs de múltiples bloques de datos a leer de acuerdo con la información de desplazamiento de los datos a leer; determinar, de acuerdo con el ID del volumen de usuario correspondiente a la operación de lectura actual y a una LBA de cada bloque de datos a leer, una partición de almacenamiento correspondiente a cada bloque de datos a leer; determinar, de acuerdo con la información de partición global, módulos de control de lectura-escritura que son correspondientes a las particiones de almacenamiento correspondientes a los múltiples bloques de datos a leer; generar múltiples comandos de lectura de bloque de datos, donde cada comando de lectura de bloque de datos es correspondiente a un bloque de datos a leer, y cada comando de lectura de bloque de datos lleva el bloque de datos a leer y un ID del bloque de datos a leer; y enviar independientemente cada comando de lectura de bloque de datos al módulo de control de lectura-escritura correspondiente a cada bloque de datos a leer.

Haciendo referencia al octavo posible modo de implementación del segundo aspecto, en un noveno posible modo de implementación, el controlador de metadatos está configurado específicamente para: determinar desplegar el delegado de almacenamiento de objetos en un nodo servidor que tiene un recurso de disco duro en el sistema de grupo de servidores, y determinar desplegar el agente de almacenamiento SAN en un nodo servidor con una carga pequeña en el sistema de grupo de servidores.

Haciendo referencia al octavo posible modo de implementación del segundo aspecto, en un décimo posible modo de implementación, el controlador de metadatos está configurado además para reunir un recurso de almacenamiento disponible del disco duro del nodo servidor y un recurso de almacenamiento disponible de la matriz de almacenamiento del nodo de almacenamiento en red, y dividir los recursos de almacenamiento disponibles del disco duro y de la matriz de almacenamiento en múltiples particiones de almacenamiento.

De acuerdo con un tercer aspecto, una realización de la presente invención da a conocer un ordenador.

De acuerdo con un cuarto aspecto, una realización de la presente invención da a conocer un medio de almacenamiento informático.

A partir de las soluciones técnicas anteriores se puede ver que, en las realizaciones de la presente invención, los recursos de almacenamiento de un disco duro y de una matriz de almacenamiento se dividen en múltiples particiones de almacenamiento, y las múltiples particiones de almacenamiento forman un conjunto de recursos de almacenamiento compartidos, un módulo de control de lectura-escritura es asignado a cada partición de almacenamiento, y se genera información de partición global para registrar una correspondencia entre cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos y el módulo de control de lectura-escritura. De este modo, cuando se recibe posteriormente un mensaje de solicitud de almacenamiento, se puede determinar una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento, se puede determinar en función de la información de partición global un módulo de control de lectura-escritura que se corresponde con la partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento, y finalmente el mensaje de solicitud de almacenamiento puede ser enviado al módulo de control de lectura-escritura determinado, de tal modo que el módulo de control de lectura-escritura lleva a cabo una operación solicitada por el mensaje de solicitud de almacenamiento. Las realizaciones de la presente invención implementan una integración rápida y simple de recursos de almacenamiento heterogéneos, y por lo tanto pueden utilizar eficientemente diversos recursos de almacenamiento, ahorrar costes y evitar el derroche de recursos.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Para describir más claramente las soluciones técnicas en las realizaciones de la presente invención o de la técnica anterior, a continuación se introducen brevemente los dibujos adjuntos necesarios para describir las realizaciones o la técnica anterior. Obviamente, los dibujos adjuntos en la siguiente descripción muestran algunas realizaciones de la presente invención.

La figura 1 es un diagrama de bloques esquemático de un sistema de grupo de servidores, de acuerdo con una realización de la presente invención;

la figura 2 es un diagrama esquemático que muestra la división de recursos de almacenamiento compartidos de acuerdo con una realización de la presente invención;

la figura 3 es un diagrama de flujo de utilización de un recurso de almacenamiento compartido de acuerdo con una realización de la presente invención;

la figura 4 es otro diagrama de flujo de utilización de un recurso de almacenamiento compartido de acuerdo con una realización de la presente invención;

la figura 5 es otro diagrama de flujo de utilización de un recurso de almacenamiento compartido de acuerdo con una realización de la presente invención;

la figura 6 es otro diagrama de bloques esquemático de un sistema de grupo de servidores, de acuerdo con una realización de la presente invención; y

la figura 7 es un diagrama de composición de un ordenador, de acuerdo con una realización de la presente invención.

DESCRIPCIÓN DE REALIZACIONES

Para aclarar los objetivos, soluciones técnicas y ventajas de las realizaciones de la presente invención, a continuación se describen de manera clara y completa las soluciones técnicas de las realizaciones de la presente invención haciendo referencia a los dibujos adjuntos de las realizaciones de la presente invención. Evidentemente, las realizaciones descritas son algunas, pero no la totalidad de las realizaciones de la presente invención.

Además, el término "y/o" en esta descripción, describe solamente una relación de asociación para describir objetos asociados, y representa que pueden existir tres relaciones. Por ejemplo, A y/o B puede representar lo siguientes tres casos: solamente existe A, existen tanto A como B, y existe solamente B. Además, en esta descripción el símbolo "/" representa en general que los objetos asociados antes y después del símbolo están en una relación "o".

En las soluciones técnicas dadas a conocer por las realizaciones de la presente invención, un controlador distribuido se despliega en un servidor para implementar la integración de recursos de almacenamiento heterogéneos, de tal modo que la integración y utilización de los recursos de almacenamiento heterogéneos se puede implementar sin necesidad de adquirir adicionalmente un dispositivo de integración y almacenamiento heterogéneo, mejorando de ese modo la relación precio/rendimiento del sistema.

En las realizaciones de la presente invención, después de que sean integrados verticalmente los recursos informáticos y los recursos de almacenamiento, se integran entonces horizontalmente varios recursos de almacenamiento, y en particular, son integrados y utilizados recursos de almacenamiento heterogéneos. En las realizaciones de la presente invención, se despliega un controlador de almacenamiento distribuido en un servidor, y se utilizan varios recursos de almacenamiento heterogéneos para formar un conjunto de recursos de almacenamiento de compartición en grupo con el fin de asignar y gestionar los recursos de almacenamiento de manera unificada. El procedimiento puede implementar una integración rápida y simple de los recursos de

almacenamiento heterogéneos, y por lo tanto puede utilizar eficientemente varios recursos de almacenamiento, ahorrar costes y evitar el derroche de recursos.

Los recursos de almacenamiento heterogéneos mencionados en las realizaciones de la presente invención se refieren a dos o más tipos diferentes de dispositivos de almacenamiento. Específicamente, un dispositivo de almacenamiento de un primer tipo se refiere a un disco duro local incorporado en un nodo servidor, tal como un disco de estado sólido (Solid State Disk, SSD), un disco duro mecánico (Hard Disk, HD) o un disco duro híbrido (Hybrid Hard Disk, HHD); un dispositivo de almacenamiento de un segundo tipo se refiere a un nodo de almacenamiento en red, donde el nodo de almacenamiento en red puede ser un dispositivo de almacenamiento de red de área de almacenamiento (Storage Area Network, SAN), y puede ser asimismo un dispositivo de almacenamiento de almacenamiento conectado en red (Network Attached Storage, NAS) y el nodo de almacenamiento en red es un dispositivo de hardware externo de un servidor, pero no es un dispositivo incorporado en el servidor.

La figura 1 es un diagrama de composición del sistema de grupo de servidores, de acuerdo con una realización de la presente invención. El sistema de grupo de servidores comunica con un cliente de aplicación o un centro de gestión de almacenamiento utilizando una capa de red. El sistema de grupo de servidores incluye un nodo servidor y un nodo de almacenamiento en red (esta realización utiliza, como ejemplo, un dispositivo de almacenamiento SAN), puede haber uno o varios nodos servidores y uno o varios nodos de almacenamiento en red, y esta realización utiliza como ejemplo 2 nodos de almacenamiento SAN. Un dispositivo físico de cada nodo servidor incluye una CPU, una memoria, una red, un disco duro y similares, y un dispositivo físico del nodo de almacenamiento en red incluye una matriz de almacenamiento y un controlador de la matriz de almacenamiento. En esta realización, los dispositivos físicos, tales como la CPU y la memoria del nodo servidor, que se utilizan para proporcionar recursos informáticos para un programa de aplicación conectado al sistema de grupo de servidores, se denominan colectivamente recursos informáticos del sistema de grupo de servidores y proporcionan la base para formar una capa informática; el disco duro del nodo servidor y la matriz de almacenamiento del nodo de almacenamiento en red que están en una capa de recursos de almacenamiento se denominan colectivamente recursos de almacenamiento del sistema de grupo de servidores.

El sistema de grupo de servidores se utiliza para proporcionar externamente los recursos informáticos a diferentes programas de aplicación para su utilización. Por ejemplo, una aplicación web o un sistema de grupo distribuido HADOOP se pueden ejecutar en el sistema de grupo de servidores. Los recursos informáticos del sistema de grupo de servidores se pueden abstraer adicionalmente en múltiples máquinas virtuales; diferentes programas de aplicación se ejecutan en cada máquina virtual, o las múltiples máquinas virtuales forman un grupo de máquinas virtuales para proporcionar servicios para el mismo programa de aplicación. Esta realización no impone una limitación sobre una forma de implementación específica. Cuando un programa de aplicación se ejecuta en el sistema de grupo de servidores, los datos relacionados del programa de aplicación se pueden almacenar en un recurso de almacenamiento del sistema de grupo de servidores, es decir, almacenar en el disco duro del nodo servidor, o en la matriz de almacenamiento del nodo SAN, o se pueden almacenar tanto en el disco duro del nodo servidor como en la matriz de almacenamiento del nodo SAN.

El sistema de grupo de servidores en ésta realización de la presente invención ejecuta además un controlador de almacenamiento distribuido, donde el controlador de almacenamiento distribuido está configurado para dividir recursos de almacenamiento del disco duro del nodo servidor y de la matriz de almacenamiento proporcionada por un nodo de almacenamiento en red (tal como una SAN) en múltiples particiones de almacenamiento, donde las múltiples particiones de almacenamiento forman un conjunto de recursos de almacenamiento compartidos del sistema de grupo de servidores, de tal modo que un programa de aplicación que se ejecuta en el sistema de grupo de servidores puede obtener un bloque de recursos de almacenamiento distribuidos a partir del conjunto de recursos de almacenamiento compartidos y utilizar este bloque de recursos de almacenamiento distribuidos, garantizando de ese modo una mayor utilización de los recursos de almacenamiento y una distribución homogénea del almacenamiento, y mejorar la eficiencia de lectura-escritura de los recursos de almacenamiento. En esta realización de la presente invención, el controlador de almacenamiento distribuido se implementa utilizando un módulo de software instalado en un dispositivo de hardware de un servidor, y por lo tanto se puede evitar la adquisición de un dispositivo de hardware adicional como dispositivo de control de almacenamiento. La solución es más económica y ahorra costes.

El controlador de almacenamiento distribuido descrito en esta realización de la presente invención es un término general para un módulo funcional de control de almacenamiento que se ejecuta en cada nodo servidor, y el controlador de almacenamiento distribuido proporcionado por la solución puede incluir diferentes módulos funcionales. Sin embargo, durante el despliegue real, cada nodo servidor puede ejecutar diferentes módulos funcionales del controlador de almacenamiento distribuido, según la función y la política de despliegue del nodo servidor. Es decir, de acuerdo con la política de despliegue del sistema de grupo de servidores, diferentes módulos funcionales del controlador de almacenamiento distribuido se pueden ejecutar en diferentes nodos servidores, y cada nodo servidor puede ejecutar la totalidad de los módulos funcionales del controlador de almacenamiento distribuido, o puede ejecutar algunos módulos funcionales del controlador de almacenamiento distribuido. A continuación se describen en detalle modos de despliegue específicos.

El controlador de almacenamiento distribuido está configurado principalmente para proporcionar una interfaz de acceso de datos para los recursos informáticos del sistema de grupo de servidores, y para llevar a cabo la gestión y el control de lectura-escritura sobre los recursos de almacenamiento compartidos del sistema de grupo de servidores.

Específicamente, el controlador de almacenamiento distribuido se puede dividir funcionalmente en los módulos siguientes:

un controlador de metadatos MDC, configurado para: adquirir un recurso de almacenamiento del disco duro local del nodo servidor y un recurso de almacenamiento de la matriz de almacenamiento del nodo de almacenamiento en red, dividir el recurso de almacenamiento del nodo servidor y el recurso de almacenamiento del nodo de almacenamiento en red en múltiples particiones de almacenamiento (partición), asignar un identificador de partición de almacenamiento a cada partición de almacenamiento, y a continuación formar un conjunto de recursos de almacenamiento compartidos utilizando las múltiples particiones de almacenamiento, de tal modo que un programa de aplicación que se ejecuta en el sistema de grupo de servidores utiliza los recursos de almacenamiento compartidos.

Específicamente, el MDC puede llevar a cabo en primer lugar una revisión del estado del recurso de disco duro del nodo servidor y de la matriz de almacenamiento del nodo de almacenamiento en red, y recopilar los recursos de almacenamiento disponibles de los mismos para formar el conjunto de recursos de almacenamiento compartidos. Durante la división de la partición, el MDC puede dividir los recursos en particiones de almacenamiento de un mismo tamaño, por ejemplo, división en unidades de 10 GB. La información sobre los recursos de almacenamiento recopilada por el MDC puede incluir: la capacidad y un ID de cada disco duro, un ID de un servidor en el que está emplazado cada disco duro, la capacidad y un ID de cada unidad de almacenamiento lógica LUN incluida en cada matriz de almacenamiento, y un ID de un nodo de almacenamiento en red en el que está situada cada LUN.

La información sobre los recursos de almacenamiento recopilados por el MDC es, por ejemplo, la siguiente:

ID del disco = 1, capacidad del disco = 50 GB, ID del servidor = 1

ID del disco = 2, capacidad del disco = 50 GB, ID del servidor = 1

ID del disco = 3, capacidad del disco = 50 GB, ID del servidor = 2

LUN = 1, capacidad de la LUN = 50 GB, ID de la SAN = 1;

LUN = 2, capacidad de la LUN = 50 GB, ID de la SAN = 1; y

LUN = 3, capacidad de la LUN = 50 GB, ID de la SAN = 1;

Después de reunir la información anterior sobre los recursos de almacenamiento, el MDC divide los recursos de almacenamiento de los discos 1 a 3 y de las LUN 1 a 3 en múltiples particiones de almacenamiento, donde las particiones de almacenamiento pueden ser de tamaño igual o diferente. Por ejemplo, los recursos de almacenamiento se dividen homogéneamente en unidades de 10 GB, y los recursos de almacenamiento de los discos y de las LUN se dividen en 30 particiones de almacenamiento, donde el tamaño de cada partición de almacenamiento es de 10 GB, y los identificadores de partición de las particiones de almacenamiento son 1 a 30, es decir, P1 a P30. El MDC forma un conjunto de recursos de almacenamiento compartidos utilizando las particiones de almacenamiento P1 a P30, donde las particiones de almacenamiento P1 a P15 incluyen el recurso de almacenamiento del disco duro incorporado en el nodo servidor, y las particiones de almacenamiento P16 a P30 incluyen el recurso de almacenamiento de la matriz de almacenamiento del nodo SAN. Es decir, los recursos de almacenamiento compartidos incluyen dos tipos de particiones de almacenamiento, donde las particiones de almacenamiento del primer tipo son P1 a P15, y las particiones de almacenamiento del segundo tipo son P16 a P30.

El controlador de almacenamiento distribuido incluye además un módulo de control de lectura-escritura, donde el módulo de control de lectura-escritura incluye en esta realización un delegado de almacenamiento de objetos (object storage delegate, OSD) y un agente de almacenamiento SAN (SAN storage agent, SSA). El OSD está configurado para llevar a cabo un control de lectura-escritura sobre el recurso de almacenamiento del disco duro incorporado en el nodo servidor, es decir, para implementar almacenamiento de datos en, y adquisición de datos desde el disco duro local del nodo servidor; por ejemplo, el OSD lleva a cabo un control de lectura-escritura sobre las particiones de almacenamiento P1 a P15 en esta realización. El SSA lleva a cabo un control de lectura-escritura sobre el recurso de almacenamiento de la matriz de almacenamiento del nodo SAN, es decir, implementa el almacenamiento de datos en, y adquisición de datos desde la matriz de almacenamiento del nodo SAN; por ejemplo, el SSA lleva a cabo un control de lectura-escritura sobre las particiones de almacenamiento P16 a P30 en esta realización. Tanto el OSD como el SSA son módulos funcionales del controlador de almacenamiento distribuido, y después de reunir la información sobre los recursos de almacenamiento del sistema de grupo de servidores, el MDC puede determinar además, de acuerdo con los estados de despliegue de los recursos de almacenamiento, cómo desplegar el OSD y la SSA en el sistema de grupo de servidores. Específicamente, el MDC puede ejecutar el OSD en cada nodo servidor que tiene un disco duro local en el sistema de grupo de servidores; el MDC puede ejecutar el SSA en cada nodo servidor en el sistema de grupo de servidores, y puede asimismo desplegar, en función de la situación de carga de

cada nodo servidor, el SSA en un nodo servidor con poca carga. Por ejemplo, el MDC puede calcular de manera unificada las condiciones de carga de los recursos informáticos en todos los nodos servidores, y generar información global de despliegue del SSA en función de la capacidad de la matriz de almacenamiento de cada nodo de almacenamiento SAN y de la carga. En esta realización, por ejemplo, el MDC ejecuta un OSD1 en un nodo servidor 1, ejecuta un OSD2 en un nodo servidor 2 y ejecuta un SSA1 en el nodo servidor 2.

Después de determinar los estados de despliegue del OSD y del SSA, el MDC puede registrar además información de visualización del OSD e información de visualización del SSA. La información de visualización del OSD incluye un servidor en el que está desplegado correspondientemente el OSD, que se utiliza para indicar información de encaminamiento del OSD; además, la visualización del OSD puede incluir además cada OSD, un estado correspondiente a cada OSD y un DISCO gestionado correspondientemente por cada OSD. La información de visualización del SSA incluye un servidor en el que está desplegado correspondientemente el SSA, que se utiliza para indicar información de encaminamiento del SSA; además, la información de visualización del SSA incluye además el estado de cada SSA y una LUN, de la matriz de almacenamiento de la SAN, que está gestionada correspondientemente por cada SSA. Por ejemplo, las siguientes tabla 1 y tabla 2 muestran la información de visualización del OSD y la información de visualización del SSA, respectivamente:

Tabla 1: información de visualización del OSD

Información del OSD	Información del nodo de despliegue	Estado del OSD (opcional)	Información del dispositivo de almacenamiento (opcional)
OSD1	Servidor 1	IN	DISCO1 y DISCO2
OSD2	Servidor 2	IN	DISCO3

Tabla 2: información de visualización del SSA

Información del SSA	Información del nodo de despliegue	Estado del SSA (opcional)	Información del dispositivo de almacenamiento (opcional)
SSA1	Servidor 2	IN	LUN1, LUN2 y LUN3

Las anteriores tabla 1 y tabla 2 describen respectivamente información de visualización del OSD e información de visualización del SSA, y un experto en la materia puede combinar asimismo las anteriores tabla 1 y tabla 2 en información de visualización de un módulo de control de lectura-escritura.

Después de dividir los recursos de almacenamiento en particiones de almacenamiento y determinar el despliegue del módulo de control de lectura-escritura, el MDC puede configurar además un correspondiente módulo de control de lectura-escritura para cada partición de almacenamiento. El proceso de asignación puede ser relativamente flexible y puede ser determinado por el MDC en función de la situación de división de las particiones de almacenamiento y de la carga de ejecución real. Por ejemplo, las particiones de almacenamiento P1 a 10 se despliegan correspondientemente en el nodo servidor 1, y el OSD1 que se ejecuta en el nodo servidor 1 sirve como módulo de control de lectura-escritura para las particiones de almacenamiento; las particiones de almacenamiento P11 a 20 se despliegan correspondientemente en el nodo servidor 2, y el OSD2 que se ejecuta en el nodo servidor 2 sirve como un módulo de control de lectura-escritura para las particiones de almacenamiento; las particiones de almacenamiento P21 a 30 se despliegan correspondientemente en el nodo servidor 2, y el SSA1 que se ejecuta en el nodo servidor 2 sirve como módulo de control de lectura-escritura para las particiones de almacenamiento.

Además, el MDC puede generar adicionalmente información de partición global (se utiliza una tabla de información de partición global como ejemplo en esta realización de la presente invención), donde la tabla de información de partición global registra una situación de distribución de las particiones de almacenamiento en el sistema de grupo de servidores. Tal como se muestra en la figura 2 y en la tabla 3, la tabla de información de partición global registra un módulo de control de lectura-escritura (un OSD o un SSA) correspondiente a cada partición de almacenamiento. La tabla de información de partición global puede registrar además información sobre un dispositivo de almacenamiento de origen correspondiente a cada partición de almacenamiento, por ejemplo, un número de disco magnético o información de dirección física.

Tal como se muestra en la tabla 3, el módulo de control de lectura-escritura correspondiente a P1 es el OSD1, una unidad de almacenamiento de origen correspondiente a P1 es un DISCO1 en un SERVIDOR1, y una dirección física de origen correspondiente a P1 es de 100 a 199.

Tabla 3: tabla de información de partición global

Almacenamiento	Lectura-escritura	Información del dispositivo de almacenamiento de origen		
información de partición	módulo de control	unidad de origen (opcional)	nodo de origen (opcional)	dirección física de origen (opcional)
P1	OSD1	DISCO1	Servidor1	100-199
P2	OSD1	DISCO3	Servidor2	300-399
P3	SSA1	LUN3	SAN1	1000-1999
...	...	...	...	...
P30	SSA1	DISCO2	Servidor1	200-299

El controlador de almacenamiento distribuido incluye además un servicio de bloque virtual VBS. Después de completar el despliegue de las particiones de almacenamiento y del módulo de control de lectura-escritura, el MDC puede suministrar además la anterior tabla de información de partición global e información de visualización del módulo de control de lectura-escritura al VBS. El VBS obtiene una visualización de E/S de acuerdo con la información suministrada por el MDC, donde la visualización de E/S es una tabla secundaria de la tabla de información de partición global, se utiliza para indicar un módulo de control de lectura-escritura actual correspondiente a cada partición de almacenamiento, e incluye una correspondencia entre la partición de almacenamiento y el módulo de control de lectura-escritura. La visualización de E/S puede ser suministrada directamente por el MDC al VBS, y puede ser asimismo generada por el VBS en función de la tabla de información de partición global suministrada por el módulo de MDC.
El VBS se puede ejecutar en cada nodo servidor en el sistema de grupo de servidores, y como capa de controlador de almacenamiento, está configurado para proporcionar una interfaz de acceso de bloques para un módulo de aplicación del sistema de grupo de servidores, por ejemplo, una interfaz de acceso de dispositivos de bloque basada en la SCSI. Después de recibir una solicitud de lectura-escritura suministrada por una aplicación de capa superior, el VBS determina una partición de almacenamiento que tiene que ser leída y escrita según la solicitud de lectura-escritura de datos, determina, de acuerdo con una regla de visualización de la visualización de E/S, un módulo de control de lectura-escritura (un OSD y un SSA) correspondiente a la partición de almacenamiento solicitada por la actual solicitud de lectura-escritura de datos, y suministra la solicitud de datos de lectura-escritura al correspondiente módulo de control de lectura-escritura, para completar la lectura y escritura de datos.

Específicamente, el VBS puede soportar además la gestión de metadatos globales, donde los metadatos globales registran un estado de utilización global de las particiones de almacenamiento en el conjunto de recursos de almacenamiento compartidos en el sistema de grupo de servidores, y metadatos de cada partición de almacenamiento. El estado de utilización global incluye información sobre una partición de almacenamiento ocupada e información sobre una partición de almacenamiento libre. Los metadatos de cada partición de almacenamiento se utilizan para indicar el estado de asignación de cada partición de almacenamiento. En esta realización de la presente invención, las particiones de almacenamiento se asignan a modo de asignación de almacenamiento de datos de bloque. Es decir, cada partición de almacenamiento se utiliza en una unidad de un bloque de datos, y la utilización de la partición de almacenamiento incluye lectura, escritura, asignación o similares. Por ejemplo, cuando se asigna una partición de almacenamiento a un volumen de usuario, se utiliza un bloque de datos como unidad para asignación. Por ejemplo, en esta realización de la presente invención, cada partición de almacenamiento tiene 10 GB de tamaño, y la partición de almacenamiento de 10 GB se puede dividir homogéneamente en 10.240 bloques de datos (bloque). Cuando se leen datos desde cada partición de almacenamiento o se escriben datos en cada partición de almacenamiento, se utiliza un bloque de datos como unidad para lectura y escritura. Por lo tanto, los metadatos de cada partición de almacenamiento incluyen específicamente una correspondencia entre IDs de bloque asignados a cada partición de almacenamiento, donde múltiples bloques de datos son asignados a cada partición de almacenamiento. Los bloques de datos pueden ser de tamaño uniforme, o pueden no estar limitados en tamaño. Esta realización de la presente invención utiliza un ejemplo en el que el tamaño de cada bloque de datos es de 1 MB. Además, en esta realización de la presente invención, un ID de cada bloque de datos puede incluir un ID de un volumen de usuario correspondiente al bloque de datos, o puede incluir un ID de un volumen de usuario correspondiente al bloque de datos y una dirección de bloque lógico (logical block address, LBA).

Los metadatos de cada partición de almacenamiento se muestran, por ejemplo, en la tabla 4:

Tabla 4: metadatos de una partición de almacenamiento

ID de partición de almacenamiento	Bloque de datos asignado	Dirección física de un bloque de datos asignado (opcional)	Tamaño de un bloque de datos asignado (opcional)
P1	Bloque 1 (volumen 1 + LBA 1)	×××-×××	1 MB
P2	Bloque 2 (volumen 1 + LBA 2)	×××-×××	1 MB
P3	Bloque 3 (volumen 1 + LBA 3)	×××-×××	1 MB
...	...	...	...
P30	Bloque 100 (volumen 2 + LBA 70)	×××-×××	1 MB

Una correspondencia entre una partición de almacenamiento y un bloque de datos asignado puede estar en forma de índice de Valor de clave, donde un ID del bloque de datos es un valor de clave. Por ejemplo, el valor de clave está relacionado con un identificador de un volumen de usuario y una dirección de bloque lógico del bloque de datos, y un ID de la partición de almacenamiento es un valor Valor. Se debe observar que, si se utiliza la forma de índice de Valor de clave, el VBS puede asimismo determinar directamente la correspondencia utilizando un algoritmo sin mantener la tabla 4 anterior. Al ser inicializado, el VBS puede adquirir información de asignación de recursos de almacenamiento recorriendo el disco duro del nodo servidor y un disco magnético de la matriz de almacenamiento del nodo SAN, e inicializa metadatos almacenados de acuerdo con la tabla de información de partición global suministrada por el MDC.
El controlador de almacenamiento distribuido incluye además un módulo de control de lectura-escritura, por ejemplo, un OSD que lleva a cabo control de lectura-escritura sobre un recurso de disco duro del nodo servidor, y un SSA que lleva a cabo control de lectura-escritura sobre un recurso de almacenamiento de la matriz de almacenamiento del nodo de almacenamiento en red.

Específicamente, el OSD recibe principalmente un comando de lectura-escritura desde el VBS, y completa el almacenamiento de datos en, y la adquisición de datos desde el disco duro del nodo servidor. El SSA recibe principalmente un comando de lectura-escritura desde el VBS, y completa el almacenamiento de datos en, y la adquisición de datos desde el disco duro del nodo SAN. El SSA está configurado para implementar un agente de un dispositivo SAN en un anfitrión. Se configura una vista, en el SSA, para información de almacenamiento de cada dispositivo SAN físico, se lleva a cabo el acceso a cada dispositivo SAN/NAS físico utilizando el agente del SSA, y se añade una función de interfaz iSCSI al SSA.

Además, si se asigna una dirección física uniforme a las particiones de almacenamiento en el conjunto de recursos de almacenamiento compartidos, el SSA puede mantener además una correspondencia entre la dirección física uniforme y la dirección LUN original en el nodo SAN, donde el SSA puede determinar además, de acuerdo con la correspondencia, una dirección de una LUN original correspondiente a la solicitud de lectura-escritura.

El anterior sistema de grupo de servidores ejecuta el controlador de almacenamiento distribuido, donde el MDC, el VBS, el OSD y el SSA en el controlador de almacenamiento distribuido pueden integrar y utilizar recursos de almacenamiento heterogéneos, formar un conjunto de recursos de compartición en grupo utilizando varios recursos de almacenamiento heterogéneos, y asignar y gestionar la totalidad de los recursos de almacenamiento de manera unificada, mejorando de ese modo la utilización de los recursos de almacenamiento. Además, las múltiples particiones de almacenamiento se pueden leer o escribir simultáneamente, mejorando de ese modo el rendimiento de lectura-escritura y aumentando el grado de interés del sistema.

Haciendo referencia a la figura 1, la figura 3 es un proceso de procesamiento de creación de un volumen de usuario en un sistema de grupo de servidores que utiliza integración de recursos de almacenamiento heterogéneos, de acuerdo con una realización de la presente invención.

S301: un VBS desplegado en un nodo servidor en el sistema de grupo de servidores recibe un comando de creación de volumen de usuario enviado por un extremo de aplicación.

Específicamente, un programa de aplicación (por ejemplo, una máquina virtual) del extremo de aplicación que se ejecuta en el sistema de grupo de servidores inicializa el comando de creación de volumen de usuario, donde el comando es enviado mediante un gestor de aplicación a un VBS desplegado en cualquier nodo servidor en el sistema de grupo de servidores (de manera preferente, un VBS en un nodo servidor recibe el comando de creación de volumen de usuario, donde un recurso informático de la máquina virtual que inicializa el comando está situado en el nodo servidor). Preferentemente, si el sistema de grupo de servidores en esta realización de la presente invención proporciona además una función de VBS primario y secundario, después de recibir el comando de creación de volumen de usuario, el VBS puede determinar además si el VBS es un VBS primario en el sistema de grupo de

servidores. Si no lo es, el VBS envía el comando de creación de volumen de usuario al VBS primario. De hecho, el despliegue del VBS es relativamente flexible. Los VBS instalados en todos los nodos servidores en el sistema de grupo de servidores pueden ser de igual rango, y en este caso, las configuraciones y funciones de todos los VBS son idénticas. Se puede tener asimismo que un VBS en el sistema de grupo de servidores está seleccionado como el VBS primario, y los otros VBS se utilizan como VBS secundarios. El VBS primario está configurado para asignar un volumen de usuario/bloque de datos y gestionar metadatos de una partición de almacenamiento, y el VBS secundario está configurado para consultar los metadatos en el VBS primario, y llevar a cabo una operación de acuerdo con un comando del VBS primario. Esta realización de la presente invención utiliza un ejemplo en el que un sistema de grupo de servidores implementa VBSs primario y secundarios.

S302: un VBS primario consulta metadatos globales según la información del tamaño de un volumen indicado por el comando de creación de volumen de usuario; determina si los recursos restantes en un conjunto de recursos de almacenamiento compartidos satisfacen un requisito; y en caso afirmativo, crea un volumen de usuario, es decir, determina un identificador de volumen (ID) del volumen de usuario, asigna una partición de almacenamiento inicial al volumen de usuario y registra en los metadatos de la partición de almacenamiento inicial el identificador del volumen de usuario y la información sobre la partición de almacenamiento inicial asignada.

Específicamente, si el comando de creación de volumen de usuario especifica el ID del volumen de usuario, el VBS primario utiliza directamente el ID del volumen de usuario en el comando de creación de volumen de usuario; si el comando de creación de volumen de usuario no especifica el ID del volumen de usuario, el VBS asigna al volumen de usuario un ID del volumen de usuario.

En un proceso de creación del volumen de usuario, el VBS puede asignar además la partición de almacenamiento inicial al volumen de usuario, es decir, algunas particiones de almacenamiento se seleccionan a partir de particiones de almacenamiento libres como particiones de almacenamiento iniciales del volumen de usuario. El tamaño de un recurso de almacenamiento inicial del volumen de usuario se puede asignar flexiblemente según la capacidad del volumen de usuario, especificada por el comando de creación de volumen de usuario. Toda la capacidad del volumen de usuario especificada por el comando de creación de volumen de usuario puede ser utilizada como capacidad de la partición de almacenamiento inicial. Por ejemplo, el comando de creación de volumen de usuario solicita crear un volumen de usuario de 5 GB, y el VBS puede asignar la totalidad de los 5 GB al volumen de usuario, como partición de almacenamiento inicial. Es decir, los 5 GB se dividen en 5120 bloques de datos con un tamaño de 1 MB, los 5120 bloques de datos se despliegan en las particiones de almacenamiento P1 a P30 de manera distribuida y, en este caso, el tamaño de la partición de almacenamiento inicial es de 5 GB. El VBS puede utilizar asimismo un modo de asignación reducida para asignar una parte del recurso de almacenamiento al volumen de usuario, según el estado real del conjunto de recursos de almacenamiento compartidos, por ejemplo, asignar un recurso de almacenamiento inicial de 1 GB al volumen de usuario. El recurso de almacenamiento inicial de 1 GB se divide en 1024 bloques de datos con un tamaño de 1 MB, los 1024 bloques de datos se despliegan en las particiones de almacenamiento P1 a P30 de manera distribuida y, en este caso, el tamaño de la partición de almacenamiento inicial es de 1 GB.

El VBS registra el ID del volumen de usuario y la información sobre la partición de almacenamiento inicial asignada en información de metadatos de cada partición de almacenamiento inicial en los metadatos globales.

Cuando se asigna la partición de almacenamiento inicial al volumen de usuario, el VBS asigna asimismo una correspondiente dirección física de origen a cada bloque de datos de cada volumen de usuario.

S303: el VBS primario monta el volumen de usuario, y genera un dispositivo de almacenamiento virtual después de que el montaje sea satisfactorio.

S304: el VBS primario devuelve información sobre el dispositivo de almacenamiento virtual al extremo de aplicación.

S305: el VBS primario devuelve los metadatos globales a un MDC en el sistema de grupo de servidores, de tal modo que el MDC actualiza una tabla de información de partición global de acuerdo con los metadatos globales.

La etapa 305 es una etapa opcional, y se puede implementar en una secuencia flexible.

Haciendo referencia a la figura 1, la figura 4 es un proceso de procesamiento en el que un usuario escribe datos en un sistema de grupo de servidores que utiliza integración de recursos de almacenamiento heterogéneos, de acuerdo con una realización de la presente invención.

S401: después de que algún programa de aplicación que se ejecuta en el sistema de grupo de servidores inicie una operación de escritura de datos, un VBS en el sistema de grupo de servidores recibe una solicitud de operación de escritura de datos.

La solicitud de operación de escritura de datos lleva un nombre de archivo y datos a escribir.

S402: el VBS determina, de acuerdo con un nombre de archivo contenido en la solicitud de operación de escritura de datos, un ID de un volumen de usuario correspondiente a la actual operación de escritura.

El VBS puede calcular además, de acuerdo con los datos a escribir, el tamaño de los datos a escribir.

El VBS asigna una LBA a los datos a escribir (la asignación de una LBA es opcional en esta etapa, y el VBS puede asimismo no asignar una LBA a los datos a escribir en esta etapa).

Por ejemplo, el VBS determina el ID de volumen 1, tamaño = 1 GB, y una LBA: 001×-221× para la actual operación de escritura.

S403: el VBS segmenta los datos a escribir en múltiples bloques de datos, y asigna una LBA a cada bloque de datos.

El VBS puede segmentar homogéneamente los datos a escribir según una unidad, por ejemplo, segmentar los datos según 1 MB, es decir, segmentar los datos según una unidad de utilización de cada partición de almacenamiento, cada vez. En esta realización, el VBS segmenta los datos a escribir con el tamaño de 1 GB en 1024 bloques de datos, donde el tamaño de cada bloque de datos es de 1 MB. Si los restantes datos a escribir son menores de 1 MB, el tamaño del último bloque de datos es el tamaño real de los restantes datos a escribir. El VBS asigna además la correspondiente LBA a cada bloque de datos.

Por ejemplo:

Bloque 1 LBA:0000-1024

Bloque 2 LBA:1025-2048

...

S404: el VBS determina una correspondiente partición de almacenamiento para cada bloque de datos.

Específicamente, el VBS determina en primer lugar la dirección de bloque lógico (LBA) de cada bloque de datos a escribir, combina a continuación el ID del volumen de usuario y la LBA de cada bloque de datos en un valor de clave de cada bloque de datos, y determina, de acuerdo con un algoritmo de almacenamiento distribuido, tal como un algoritmo hash, la partición de almacenamiento correspondiente a cada bloque de datos. La LBA en la presente memoria puede ser un valor después de que se haya procesado una LBA original, por ejemplo, la LBA 0000-1024 correspondiente al bloque 1 corresponde a 1, y la LBA 1025-2048 correspondiente al bloque 2 corresponde a 2.

S405: el VBS genera múltiples comandos de escritura de bloque de datos, donde cada bloque de datos corresponde a un comando de escritura de bloque de datos, y cada comando de escritura de bloque de datos lleva el bloque de datos a escribir, y un ID del bloque de datos a escribir (por ejemplo, un ID de bloque incluye el ID del volumen de usuario y la LBA del bloque de datos a escribir).

Esta etapa se puede llevar a cabo asimismo después de que se hayan completado las etapas siguientes, y la implementación específica no se limita a ninguna secuencia temporal.

S406: el VBS determina, de acuerdo con la partición de almacenamiento correspondiente a cada bloque de datos, un módulo de control de lectura-escritura correspondiente a cada bloque de datos.

Específicamente, el VBS determina, de acuerdo con una tabla de información de partición global, el módulo de control de lectura-escritura correspondiente a cada bloque de datos.

S407: el VBS envía cada comando de escritura de bloque de datos al módulo de control de lectura-escritura correspondiente a cada bloque de datos, de manera que el módulo de control de lectura-escritura correspondiente a cada bloque de datos escribe cada bloque de datos en un recurso de hardware de almacenamiento.

Específicamente, si un OSD recibe el comando de escritura de bloque de datos, el OSD consulta, de acuerdo con el ID del bloque de datos a escribir, los metadatos de bloque de datos almacenados por el OSD, y determina si la operación a llevar a cabo por el OSD sobre el ID del bloque de datos es una primera operación. Si se trata de la primera operación, el OSD asigna una dirección física real al bloque de datos a escribir, escribe el bloque de datos a escribir en un disco magnético correspondiente a la dirección física, actualiza los metadatos de bloque de datos almacenados por el OSD y registra la correspondencia entre el ID del bloque de datos a escribir y la dirección física. Si no es la primera operación, el OSD consulta, de acuerdo con el ID del bloque de datos a escribir, los metadatos de bloque de datos almacenados por el OSD, determina una dirección física correspondiente al bloque de datos a escribir y escribe el bloque de datos a escribir en la dirección física obtenida mediante la consulta.

Si un SSA recibe el comando de escritura de bloque de datos, el SSA consulta, de acuerdo con el ID del bloque de datos a escribir, los metadatos de bloque de datos almacenados por el SSA, y determina si la operación a llevar a cabo por el SSA sobre el ID del bloque de datos es una primera operación. Si es la primera operación, el SSA asigna una dirección lógica en una matriz de almacenamiento de un nodo de almacenamiento SAN real al bloque de datos a escribir, es decir, una dirección LUN, escribe el bloque de datos a escribir en un disco magnético correspondiente a la dirección LUN, actualiza los metadatos de bloque de datos almacenados por el SSA y registra la correspondencia entre el ID del bloque de datos a escribir y la dirección LUN. Si no es la primera operación, el

OSD consulta, de acuerdo con el ID del bloque de datos a escribir, los metadatos de bloque de datos almacenados por el OSD, determina una dirección LUN correspondiente al bloque de datos a escribir y escribe el bloque de datos a escribir en la dirección LUN obtenida mediante la consulta.

Durante una operación de escritura, el OSD o la SSA pueden escribir en primer lugar el bloque de datos en una capa caché local, es decir, devolver un mensaje de respuesta, mejorando de ese modo la eficiencia del almacenamiento.

Haciendo referencia a la figura 1, la figura 5 es un proceso de procesamiento en el que un usuario lee datos en un sistema de grupo de servidores que utiliza integración de recursos de almacenamiento heterogéneos, de acuerdo con una realización de la presente invención.

S501: después de que algún programa de aplicación que se ejecuta en el sistema de grupo de servidores inicie una operación de lectura de datos, un VBS en el sistema de grupo de servidores recibe una solicitud de operación de lectura de datos.

La solicitud de operación de lectura de datos lleva un nombre de archivo e información de desplazamiento de los datos a leer.

S502: el VBS determina, de acuerdo con un nombre de archivo contenido en la solicitud de operación de lectura de datos, un ID de un volumen de usuario correspondiente a la operación de lectura actual, y determina, de acuerdo con información de desplazamiento de los datos a leer, una LBA de los datos a leer.

S503: el VBS determina múltiples bloques de datos a leer de acuerdo con el ID del volumen de usuario y la LBA de los datos a leer.

Específicamente, un ID de cada bloque de datos a leer incluye el volumen de usuario y una LBA de cada bloque de datos, donde la LBA de cada bloque de datos se puede determinar en función de la cantidad de datos a leer y de un desplazamiento de los datos a leer.

S504: el VBS determina una correspondiente partición de almacenamiento para cada bloque de datos a leer.

Específicamente, el VBS determina en primer lugar la dirección de bloque lógico (LBA) de cada bloque de datos a leer, combina a continuación el ID del volumen de usuario y la LBA de cada bloque de datos en un valor de clave de cada bloque de datos, y determina, de acuerdo con un algoritmo de almacenamiento distribuido, tal como un algoritmo hash, la partición de almacenamiento correspondiente a cada bloque de datos.

S505: el VBS genera múltiples comandos de lectura de bloque de datos, donde cada bloque de datos corresponde a un comando de lectura de bloque de datos, y cada comando de lectura de bloque de datos lleva un ID del bloque de datos a leer (por ejemplo, un ID de bloque incluye el ID del volumen de usuario y una LBA del bloque de datos a leer).

S506: el VBS determina, de acuerdo con la partición de almacenamiento correspondiente a cada bloque de datos, un módulo de control de lectura-escritura correspondiente a cada bloque de datos.

Específicamente, el VBS determina, de acuerdo con una tabla de información de partición global, el módulo de control de lectura-escritura correspondiente a cada bloque de datos.

S507: el VBS envía cada comando de lectura de bloque de datos al módulo de control de lectura-escritura correspondiente a cada bloque de datos, de tal modo que el módulo de control de lectura-escritura correspondiente a cada bloque de datos lee cada bloque de datos a leer desde un recurso de hardware de almacenamiento.

Específicamente, si un OSD recibe el comando de lectura de bloque de datos, el OSD consulta, de acuerdo con el ID del bloque de datos a leer, metadatos de bloque de datos almacenados por el OSD, determina una dirección física real asignada al bloque de datos a leer y lee el bloque de datos a escribir desde un disco magnético correspondiente a la dirección física.

Si un SSA recibe el comando de escritura de bloque de datos, el SSA consulta, de acuerdo con el ID del bloque de datos a escribir, metadatos de bloque de datos almacenados por el SSA, determina una dirección lógica real en una matriz de almacenamiento de un nodo de almacenamiento SAN, es decir, una dirección LUN, que está asignada al bloque de datos a leer, y lee el bloque de datos a leer desde un disco magnético correspondiente a la dirección LUN.

Utilizando un sistema de grupo para cálculo, almacenamiento e integración según esta realización de la presente invención, los problemas de funcionamiento complicado y coste elevado debidos a la utilización de una SAN dedicada en la técnica anterior se resuelven en términos de hardware. Puede haber múltiples dispositivos de almacenamiento, y se puede desplegar una memoria caché en cada dispositivo de almacenamiento, incrementando sensiblemente de ese modo la capacidad de expansión de una memoria caché en un extremo de almacenamiento, en términos de hardware. Un recurso de almacenamiento no depende de un recurso informático, y se puede aumentar y reducir independientemente, mejorando de ese modo la escalabilidad de un sistema. Un disco

magnético persistente y un recurso de memoria caché en el sistema se virtualizan en un conjunto de recursos compartidos, y son compartidos por toda la informática, toda la informática y el almacenamiento pueden participar en lectura de datos y escritura de datos, y se mejora el rendimiento de almacenamiento del sistema por medio de una mejora de la concurrencia. Además, un sistema de grupo para cálculo, almacenamiento e integración según esta realización de la presente invención utiliza una red de intercambio de datos de alta velocidad para llevar a cabo la comunicación, y por lo tanto se incrementa adicionalmente la velocidad de intercambio de datos.

La figura 6 es otro diagrama de composición de un sistema de grupo de servidores, de acuerdo con una realización de la presente invención. El sistema de grupo de servidores incluye nodos servidores 1 y 2, y un nodo de almacenamiento en red, es decir, un dispositivo SAN de un fabricante A, donde el nodo servidor 1 incluye discos duros 1 y 2, el nodo servidor 2 incluye un disco duro 3, el nodo de almacenamiento en red incluye una matriz de almacenamiento, es decir, LUN1 y LUN2, un controlador de almacenamiento distribuido se ejecuta en los nodos servidores, y el controlador de almacenamiento distribuido incluye:

controladores de metadatos, desplegados en los dos nodos servidores en esta realización, donde el controlador de metadatos desplegado en el nodo servidor 1 es un MDC primario, el controlador de metadatos desplegado en el nodo servidor 2 es un MDC secundario y los controladores de metadatos están configurados para: dividir recursos de almacenamiento de los discos duros y de la matriz de almacenamiento en múltiples particiones de almacenamiento, donde la partición de almacenamiento múltiple forma un conjunto de recursos de almacenamiento compartidos; asignar un módulo de control de lectura-escritura a cada partición de almacenamiento; generar información de partición global, donde la información de partición global registra una correspondencia entre cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos y el módulo de control de lectura-escritura; y suministrar la información de partición global a un módulo de servicio de bloque virtual;

el módulo de servicio de bloque virtual, donde un VBS es desplegado en cada nodo servidor en esta realización, configurado para: estar situado frente a una capa de servicio, y recibir un mensaje de solicitud de almacenamiento; determinar una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento; determinar, de acuerdo con la información de partición global, un módulo de control de lectura-escritura que es correspondiente a la partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento; y enviar el mensaje de solicitud de almacenamiento al módulo de control de lectura-escritura determinado; y

el módulo de control de lectura-escritura, configurado para estar situado frente a los discos duros o el nodo de almacenamiento en red, y llevar a cabo una operación solicitada por el mensaje de solicitud de almacenamiento, donde el módulo de control de lectura-escritura en esta realización incluye un OSD1 y un OSD2 que están desplegados en el nodo servidor 1, y un OSD3, un SSA1 y un SSA2 que están desplegados en el nodo servidor 2, donde el OSD1 está configurado para llevar a cabo control de lectura-escritura sobre el disco duro 1, el OSD2 está configurado para llevar a cabo control de lectura-escritura sobre el disco duro 2, el OSD3 está configurado para llevar a cabo control de lectura-escritura sobre el disco duro 3, el SSA1 está configurado para llevar a cabo control de lectura-escritura sobre la LUN1 y el SSA2 está configurado para llevar a cabo control de lectura-escritura sobre la LUN2.

El controlador de metadatos está configurado además para: determinar independientemente estados de despliegue del delegado de almacenamiento de objetos y del agente de almacenamiento SAN en el nodo servidor; generar información de visualización del módulo de control de lectura-escritura de acuerdo con los estados de despliegue determinados, donde la información de visualización del módulo de control de lectura-escritura se utiliza para indicar información sobre un nodo servidor en el que se despliega cada módulo de control de lectura-escritura; y suministrar la información de visualización del módulo de control de lectura-escritura al módulo de servicio de bloque virtual.

Además, el controlador de metadatos está configurado específicamente para: determinar desplegar el delegado de almacenamiento de objetos en un nodo servidor que tiene un recurso de disco duro en el sistema de grupo de servidores, y determinar desplegar el agente de almacenamiento SAN en un nodo servidor con poca carga en el sistema de grupo de servidores.

Por ejemplo, en esta realización, el controlador de metadatos despliega el SSA1 y el SSA2 en el nodo servidor 2.

El módulo de servicio de bloque virtual está configurado específicamente para determinar información de encaminamiento del módulo de control de lectura-escritura de acuerdo con la información de visualización del módulo de control de lectura-escritura, y enviar el mensaje de solicitud de almacenamiento al módulo de control de lectura-escritura determinado.

En esta realización mostrada en la figura 6, se puede ejecutar adicionalmente el procedimiento descrito en cualquiera de las figuras 3 a 5, lo que no se describe de nuevo en detalle en esta realización de la presente invención.

La figura 7 es un diagrama esquemático de composición de una estructura de un ordenador según una realización de la presente invención. El ordenador de esta realización de la presente invención puede incluir:

un procesador 701, una memoria 702, un bus del sistema 704 y una interfaz de comunicaciones 705, donde la CPU 701, la memoria 702 y la interfaz de comunicaciones 705 están conectadas y completan la comunicación mutua utilizando el bus del sistema 704.

El procesador 701 puede ser una unidad central de procesamiento de un solo núcleo o de múltiples núcleos, o un circuito integrado específico, o puede estar configurado como uno o varios circuitos integrados que implementan esta realización de la presente invención.

La memoria 702 puede ser una memoria RAM de alta velocidad, o puede ser una memoria no volátil (memoria no volátil), por ejemplo, por lo menos una memoria de disco magnético.

La memoria 702 está configurada para almacenar una instrucción de ejecución por ordenador 703. Específicamente, la instrucción de ejecución por ordenador 703 puede incluir código de programa.

Cuando el ordenador está en funcionamiento, el procesador 701 ejecuta la instrucción de ejecución por ordenador 703, y puede ejecutar el procedimiento proporcionado por cualquier realización de las realizaciones de la presente invención. Más específicamente, si un controlador de almacenamiento distribuido descrito en las realizaciones de la presente invención se implementa utilizando código informático, el ordenador lleva a cabo una función del controlador de almacenamiento distribuido de las realizaciones de la presente invención.

Se debería entender que, en las realizaciones de la presente invención, "B corresponde a A" indica B está asociado con A, y B se puede determinar en función de A. Sin embargo, se debería entender además que determinar B según A no significa que B se determine solamente según A; B se puede determinar asimismo según A y/u otra información.

Un experto en la materia puede ser consciente de que, en combinación con los ejemplos descritos en las realizaciones dadas a conocer en esta descripción, las unidades y etapas de algoritmo se pueden implementar mediante hardware electrónico, software informático o una combinación de los mismos. Para describir claramente el carácter intercambiable entre el hardware y el software, en lo anterior se han descrito generalmente composiciones y etapas de cada ejemplo según funciones. Que las funciones se lleven a cabo mediante hardware o software depende de aplicaciones particulares y de condiciones de límites de diseño de las soluciones técnicas. Un experto en la materia puede utilizar diferentes procedimientos para implementar las funciones descritas para cada aplicación particular, pero no se deberá considerar que la implementación va más allá del alcance de la presente invención.

En las realizaciones dadas a conocer en la presente solicitud, se deberá comprender que el sistema dado a conocer se puede implementar de otros modos. Por ejemplo, la realización de sistema descrita es tan solo a modo de ejemplo. Por ejemplo, la división en unidades es simplemente una división de funciones lógicas y puede ser otra división en una implementación real. Por ejemplo, se pueden combinar o integrar una serie de unidades o componentes en otro sistema, o algunas características se pueden ignorar o no llevar a cabo. Además, los acoplamientos mutuos o los acoplamientos directos o conexiones de comunicación, visualizados o discutidos, se pueden implementar por medio de algunas interfaces. Las conexiones de comunicación o acoplamientos indirectos entre los aparatos o unidades se pueden implementar en forma electrónica, mecánica u otras.

Las unidades descritas como partes independientes pueden o no ser físicamente independientes, y las partes visualizadas comunidades pueden o no ser unidades físicas, pueden estar emplazadas en una posición, o pueden estar distribuidas en una serie de unidades de red. Parte o la totalidad de las unidades se pueden seleccionar en función de las necesidades reales para conseguir los objetivos de las soluciones de las realizaciones de la presente invención.

Además, las unidades funcionales en las realizaciones de la presente invención pueden estar integradas en una unidad de proceso, o cada una de las unidades pueden existir por separado físicamente, o dos o más unidades estar integradas en una unidad. La unidad integrada puede estar implementada en forma de hardware, o puede estar implementada en forma de una unidad funcional de software.

Cuando la unidad integrada se implementa en forma de una unidad funcional de software y se vende o se utiliza como un producto independiente, la unidad integrada se puede almacenar en un medio de almacenamiento legible por ordenador. En base a este concepto, las soluciones técnicas de la presente invención esencialmente, o la parte que contribuye a la técnica anterior, o la totalidad o parte de las soluciones técnicas, se pueden implementar en forma de un producto de software. El producto de software se almacena en un medio de almacenamiento e incluye varias instrucciones para ordenar a un dispositivo informático (que puede ser un ordenador personal, un servidor o un dispositivo de red) que lleve a cabo la totalidad o parte de las etapas de los procedimientos descritos en las realizaciones de la presente invención. El medio de almacenamiento anterior incluye: cualquier medio que pueda almacenar código de programa, tal como una unidad flash USB, un disco duro extraíble, memoria de sólo lectura (ROM, Read-Only Memory), una memoria de acceso aleatorio (RAM, Random Access Memory), un disco magnético o un disco óptico.

REIVINDICACIONES

1. Un procedimiento para compartir un recurso de almacenamiento, aplicado a un sistema de grupo de servidores, en el que el sistema de grupo de servidores comprende uno o varios nodos servidores y uno o varios nodos de almacenamiento en red, cada nodo servidor comprende un disco duro, cada nodo de almacenamiento en red comprende una matriz de almacenamiento, un controlador de almacenamiento distribuido se ejecuta en cada nodo servidor, y el controlador de almacenamiento distribuido comprende un controlador de metadatos, un módulo de servicio de bloque virtual y un módulo de control de lectura-escritura, el módulo de control de lectura-escritura comprende un delegado de almacenamiento de objetos configurado para realizar control de lectura-escritura sobre el recurso de almacenamiento del disco duro y un agente de almacenamiento SAN configurado para realizar control de lectura-escritura sobre el recurso de almacenamiento de la matriz de almacenamiento, y el procedimiento comprende:

determinar, mediante el controlador de metadatos, estados de despliegue del delegado de almacenamiento de objetos y el agente de almacenamiento SAN en el nodo servidor, generar información de visualización del módulo de control de lectura-escritura según los estados de despliegue determinados, incluyendo dicha información de visualización el servidor en el que está desplegado el delegado de almacenamiento de objetos o el agente de almacenamiento SAN, y suministrar la información de visualización del módulo de control de lectura-escritura al módulo de servicio de bloque virtual, utilizándose la información de visualización del módulo de control de lectura-escritura para indicar información de encaminamiento de cada módulo de control de lectura-escritura;

dividir, mediante el controlador de metadatos, los recursos de almacenamiento de los discos duros y las matrices de almacenamiento en múltiples particiones de almacenamiento, donde las múltiples particiones de almacenamiento forman un conjunto de recursos de almacenamiento compartidos;

asignar, mediante el controlador de metadatos, un módulo de control de lectura-escritura a cada partición de almacenamiento;

generar, mediante el controlador de metadatos, información de partición global y suministrar la información de partición global al módulo de servicio de bloque virtual, donde la información de partición global registra una correspondencia entre cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos y el módulo de control de lectura-escritura asignado, de cada partición de almacenamiento;

cuando se recibe, mediante el módulo de servicio de bloque virtual, un mensaje de solicitud de almacenamiento, llevar a cabo el proceso siguiente:

determinar, mediante el módulo de servicio de bloque virtual, una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento;

determinar, mediante el módulo de servicio de bloque virtual, según la información de partición global, un módulo de control de lectura-escritura que corresponde a la partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento;

determinar, mediante el módulo de servicio de bloque virtual, información de encaminamiento del módulo de control de lectura-escritura determinado, según la información de visualización del módulo de control de lectura-escritura; y

enviar, mediante el módulo de servicio de bloque virtual, el mensaje de solicitud de almacenamiento al módulo de control de lectura-escritura determinado, de acuerdo con la información de encaminamiento del módulo de control de lectura-escritura determinado, de tal modo que el módulo de control de lectura-escritura determinado lleva a cabo una operación solicitada por el mensaje de solicitud de almacenamiento.

2. El procedimiento según la reivindicación 1, en el que la determinación de una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento comprende:

determinar, mediante el módulo de servicio de bloque virtual, un ID de un volumen de usuario en el que están localizados los datos a gestionar de acuerdo con el mensaje de solicitud de almacenamiento y una dirección de bloque lógico, LBA, de por lo menos un bloque de datos de los datos a gestionar; y

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, una partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

3. El procedimiento según la reivindicación 2, en el que procedimiento comprende además:

configurar, mediante el módulo de servicio de bloque virtual, metadatos de cada partición de almacenamiento, en el que los metadatos de cada partición de almacenamiento registran una correspondencia entre un ID de la partición de almacenamiento y un ID de cada bloque de datos asignado a la partición de almacenamiento, en el que

la determinación, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos, comprende:

determinar, mediante el módulo de servicio de bloque virtual, un ID de dicho por lo menos un bloque de datos de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, consultar los metadatos de cada partición de almacenamiento y determinar un ID de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

4. El procedimiento según la reivindicación 2 ó 3, en el que la recepción, mediante el módulo de servicio de bloque virtual, de un mensaje de solicitud de almacenamiento comprende: recibir, mediante el módulo de servicio de bloque virtual, un comando de creación de volumen de usuario, en el que el comando de creación de volumen de usuario indica el tamaño del volumen de usuario, en el que

la determinación, mediante el módulo de servicio de bloque virtual, de un ID de un volumen de usuario en el que están situados los datos a gestionar de acuerdo con el mensaje de solicitud de almacenamiento y de una dirección de bloque lógico, LBA, de por lo menos un bloque de datos de los datos a gestionar, y la determinación, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos, comprende:

asignar, mediante el módulo de servicio de bloque virtual, el ID del volumen de usuario al volumen de usuario;

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el tamaño del volumen de usuario, el tamaño de un recurso de almacenamiento inicial asignado al volumen de usuario, y determinar la LBA de dicho por lo menos un bloque de datos de acuerdo con el tamaño del recurso de almacenamiento inicial; y

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

5. El procedimiento según la reivindicación 2 ó 3, en el que la recepción, mediante el módulo de servicio de bloque virtual, de un mensaje de solicitud de almacenamiento comprende: recibir, mediante el módulo de servicio de bloque virtual, una solicitud de operación de escritura de datos, en el que

la determinación, mediante el módulo de servicio de bloque virtual, de un ID de un volumen de usuario en el que están situados los datos a gestionar de acuerdo con el mensaje de solicitud de almacenamiento y de una dirección de bloque lógico, LBA, de por lo menos un bloque de datos de los datos a gestionar, y la determinación, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos, comprende:

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con un nombre de archivo contenido en la solicitud de operación de escritura de datos, un ID de un volumen de usuario correspondiente a una operación de escritura actual;

dividir, mediante el módulo de servicio de bloque virtual, los datos a escribir en múltiples bloques de datos a escribir, y asignar una LBA a cada bloque de datos a escribir; y

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario correspondiente a la operación de escritura actual, y con la LBA de cada bloque de datos a escribir, la partición de almacenamiento correspondiente a cada bloque de datos a escribir.

6. El procedimiento según la reivindicación 2 ó 3, en el que la recepción, mediante el módulo de servicio de bloque virtual, de un mensaje de solicitud de almacenamiento comprende: recibir, mediante el módulo de servicio de bloque virtual, una solicitud de operación de lectura de datos, en el que la solicitud de operación de lectura de datos contiene un nombre de archivo y un desplazamiento de los datos a leer, en el que

la determinación, mediante el módulo de servicio de bloque virtual, de un ID de un volumen de usuario en el que están situados los datos a gestionar de acuerdo con el mensaje de solicitud de almacenamiento y de una dirección de bloque lógico, LBA, de por lo menos un bloque de datos de los datos a gestionar, y la determinación, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos, comprende:

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el nombre de archivo contenido en la solicitud de operación de lectura de datos, un ID de un volumen de usuario correspondiente a una operación de lectura actual;

determinar, mediante el módulo de servicio de bloque virtual, LBAs de múltiples bloques de datos a leer, de acuerdo con el desplazamiento de los datos a leer; y

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con el ID del volumen de usuario correspondiente a la operación de lectura actual, y con la LBA de cada bloque de datos a leer, la partición de almacenamiento correspondiente a cada bloque de datos a leer;

7. El procedimiento según la reivindicación 6, en el que los múltiples bloques de datos a leer que son de un tamaño uniforme se determinan en función del desplazamiento de los datos a leer.

8. El procedimiento según la reivindicación 7, en el que la determinación, mediante el módulo de servicio de bloque virtual, según la información de partición global, de un módulo de control de lectura-escritura correspondiente al mensaje de solicitud de almacenamiento; y el envío del mensaje de solicitud de almacenamiento al módulo de control de lectura-escritura determinado, de tal modo que el módulo de control de lectura-escritura determinado lleva a cabo una operación solicitada por el mensaje de solicitud de almacenamiento, comprende:

determinar, mediante el módulo de servicio de bloque virtual, de acuerdo con la información de partición global, los módulos de control de lectura-escritura determinados que son correspondientes a las particiones de almacenamiento correspondientes a los múltiples bloques de datos a leer;

generar, mediante el módulo de servicio de bloque virtual, múltiples comandos de lectura de bloque de datos, donde cada comando de lectura de bloque de datos es correspondiente a un bloque de datos a leer, y cada comando de lectura de bloque de datos lleva el bloque de datos a leer y un ID del bloque de datos a leer; y

enviar por separado, mediante el módulo de servicio de bloque virtual, cada comando de lectura de bloque de datos al módulo de control de lectura-escritura determinado, correspondiente a cada bloque de datos a leer, de tal modo que el módulo de control de lectura-escritura determinado correspondiente a cada bloque de datos a leer lee cada bloque de datos a leer.

9. Un sistema de grupo de servidores, en el que el sistema de grupo de servidores comprende uno o varios nodos servidores y uno o varios nodos de almacenamiento en red, cada nodo servidor comprende un disco duro, cada nodo de almacenamiento en red comprende una matriz de almacenamiento, un controlador de almacenamiento distribuido se ejecuta en cada nodo servidor, y el controlador de almacenamiento distribuido comprende un controlador de metadatos, un módulo de servicio de bloque virtual y un módulo de control de lectura-escritura, el módulo de control de lectura-escritura comprende un delegado de almacenamiento de objetos configurado para realizar control de lectura-escritura sobre el recurso de almacenamiento del disco duro y un agente de almacenamiento SAN configurado para realizar control de lectura-escritura sobre el recurso de almacenamiento de la matriz de almacenamiento; en el que

el controlador de metadatos, configurado para determinar estados de despliegue del delegado de almacenamiento de objetos y el agente de almacenamiento SAN en el nodo servidor, generar información de visualización del módulo de control de lectura-escritura según los estados de despliegue determinados, incluyendo dicha información de visualización el servidor en el que está desplegado el delegado de almacenamiento de objetos o el agente de almacenamiento SAN, y suministrar la información de visualización del módulo de control de lectura-escritura al módulo de servicio de bloque virtual, la información de visualización del módulo de control de lectura-escritura se utiliza para indicar información de encaminamiento de cada módulo de control de lectura-escritura;

el controlador de metadatos está configurado además para: dividir recursos de almacenamiento de los discos duros y de las matrices de almacenamiento en múltiples particiones de almacenamiento, donde las múltiples particiones de almacenamiento forman un conjunto de recursos de almacenamiento compartidos; asignar un módulo de control de lectura-escritura a cada partición de almacenamiento; generar información de partición global, donde la información de partición global registra una correspondencia entre cada partición de almacenamiento en el conjunto de recursos de almacenamiento compartidos y el módulo de control de lectura-escritura asignado, de cada partición de almacenamiento; y suministrar la información de partición global a un módulo de servicio de bloque virtual;

el módulo de servicio de bloque virtual, configurado para: estar situado frente a la capa de servicio, y recibir un mensaje de solicitud de almacenamiento; determinar una partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento; determinar, según la información de partición global, un módulo de control de lectura-escritura que corresponde a la partición de almacenamiento correspondiente al mensaje de solicitud de almacenamiento; determinar información de encaminamiento del módulo de control de lectura-escritura determinado según la información de visualización del módulo de control de lectura-escritura, y enviar el mensaje de solicitud de almacenamiento al módulo de control de lectura-escritura determinado, de acuerdo con la información de encaminamiento del módulo de control de lectura-escritura determinado; y

el módulo de control de lectura-escritura, configurado para estar frente al disco duro o al nodo de almacenamiento en red, y llevar a cabo una operación solicitada por el mensaje de solicitud de almacenamiento.

10. El sistema según la reivindicación 9, en el que:

el controlador de metadatos está configurado específicamente para asignar el delegado de almacenamiento de objetos a una partición de almacenamiento del disco duro local, como un módulo de control de lectura-escritura, y

asignar el agente de almacenamiento SAN a una partición de almacenamiento de la matriz de almacenamiento, como un módulo de control de lectura-escritura;

el delegado de almacenamiento de objetos está configurado para: recibir el mensaje de solicitud de almacenamiento, determinar una dirección física correspondiente al mensaje de solicitud de almacenamiento, y llevar a cabo, en función de la dirección física, la operación solicitada por el mensaje de solicitud de almacenamiento en el disco duro; y

el agente de almacenamiento SAN está configurado para: recibir el mensaje de solicitud de almacenamiento, determinar una dirección lógica que es del nodo de almacenamiento en red y corresponde al mensaje de solicitud de almacenamiento, y llevar a cabo, de acuerdo con la dirección lógica, la operación solicitada por el mensaje de solicitud de almacenamiento en la matriz de almacenamiento.

11. El sistema según la reivindicación 9, en el que el módulo de servicio de bloque virtual está configurado específicamente para determinar un ID de un volumen de usuario en el que están situados datos a gestionar según el mensaje de solicitud de almacenamiento y una dirección de bloque lógico LBA de por lo menos un bloque de datos de los datos a gestionar, y determinar, de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos, una partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

12. El sistema según la reivindicación 11, en el que el módulo de servicio de bloque virtual está configurado específicamente para: configurar metadatos de cada partición de almacenamiento, en el que los metadatos de cada partición de almacenamiento registran una correspondencia entre un ID de la partición de almacenamiento y un ID de cada bloque de datos asignado a la partición de almacenamiento; determinar un ID dicho por lo menos un bloque de datos de acuerdo con el ID del volumen de usuario y con la LBA de dicho por lo menos un bloque de datos; consultar los metadatos de cada partición de almacenamiento; y determinar un ID de la partición de almacenamiento correspondiente a dicho por lo menos un bloque de datos.

13. El sistema según la reivindicación 9, en el que el controlador de metadatos está configurado específicamente para: determinar desplegar el delegado de almacenamiento de objetos en un nodo servidor que tiene un recurso de disco duro en el sistema de grupo de servidores, y determinar desplegar el agente de almacenamiento SAN en un nodo servidor con poca carga en el sistema de grupo de servidores.

14. El sistema según la reivindicación 9, en el que el controlador de metadatos está configurado además para recopilar un recurso de almacenamiento disponible del disco duro del nodo servidor y un recurso de almacenamiento disponible de la matriz de almacenamiento del nodo de almacenamiento en red, y dividir los recursos de almacenamiento disponibles del disco duro y de la matriz de almacenamiento en múltiples particiones de almacenamiento.

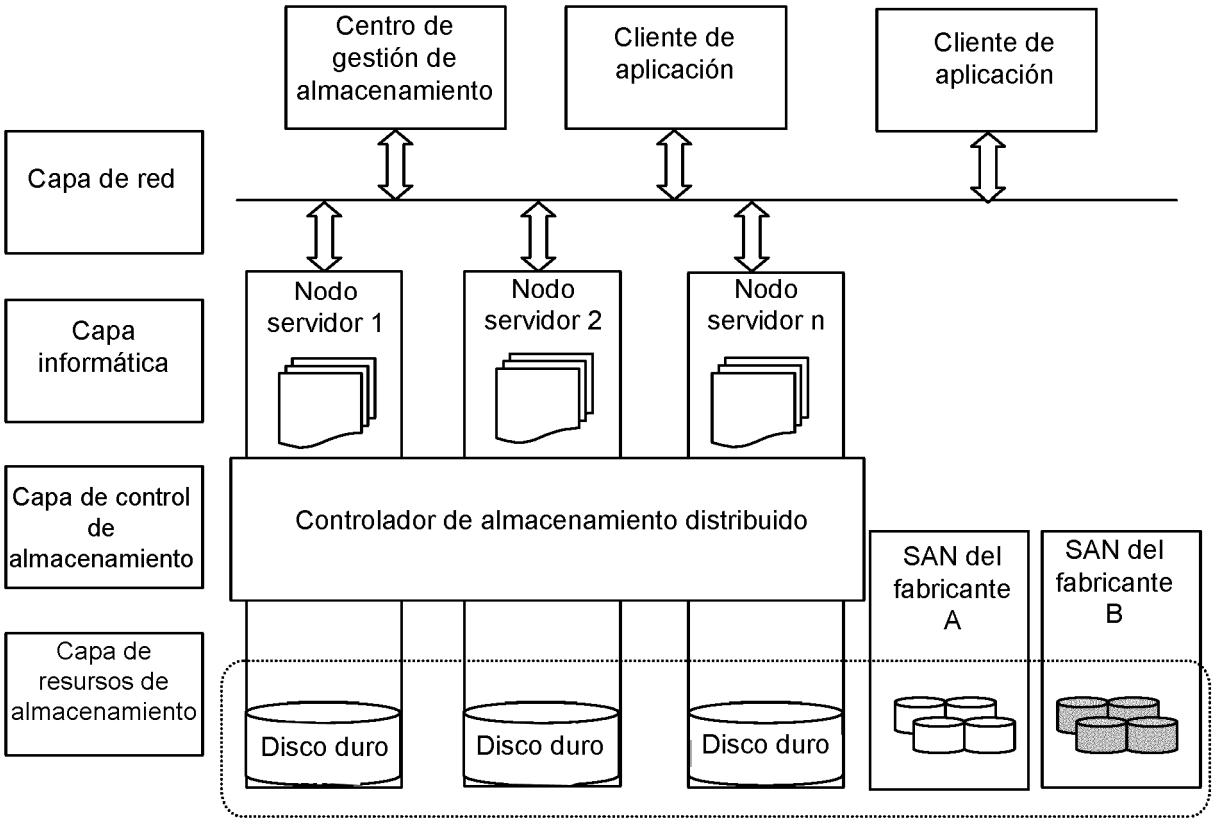
Centro de gestión de almacenamiento
Cliente de aplicación
Cliente de aplicación
Capa de red
Capa informática
Capa de control de almacenamiento
Capa de resursos de almacenamiento
Nodo servidor 1
Nodo servidor 2
Nodo servidor n
Controlador de almacenamiento distribuido
SAN del fabricante A
SAN del fabricante B
Disco duro
Disco duro
Disco duro

FIG. 1

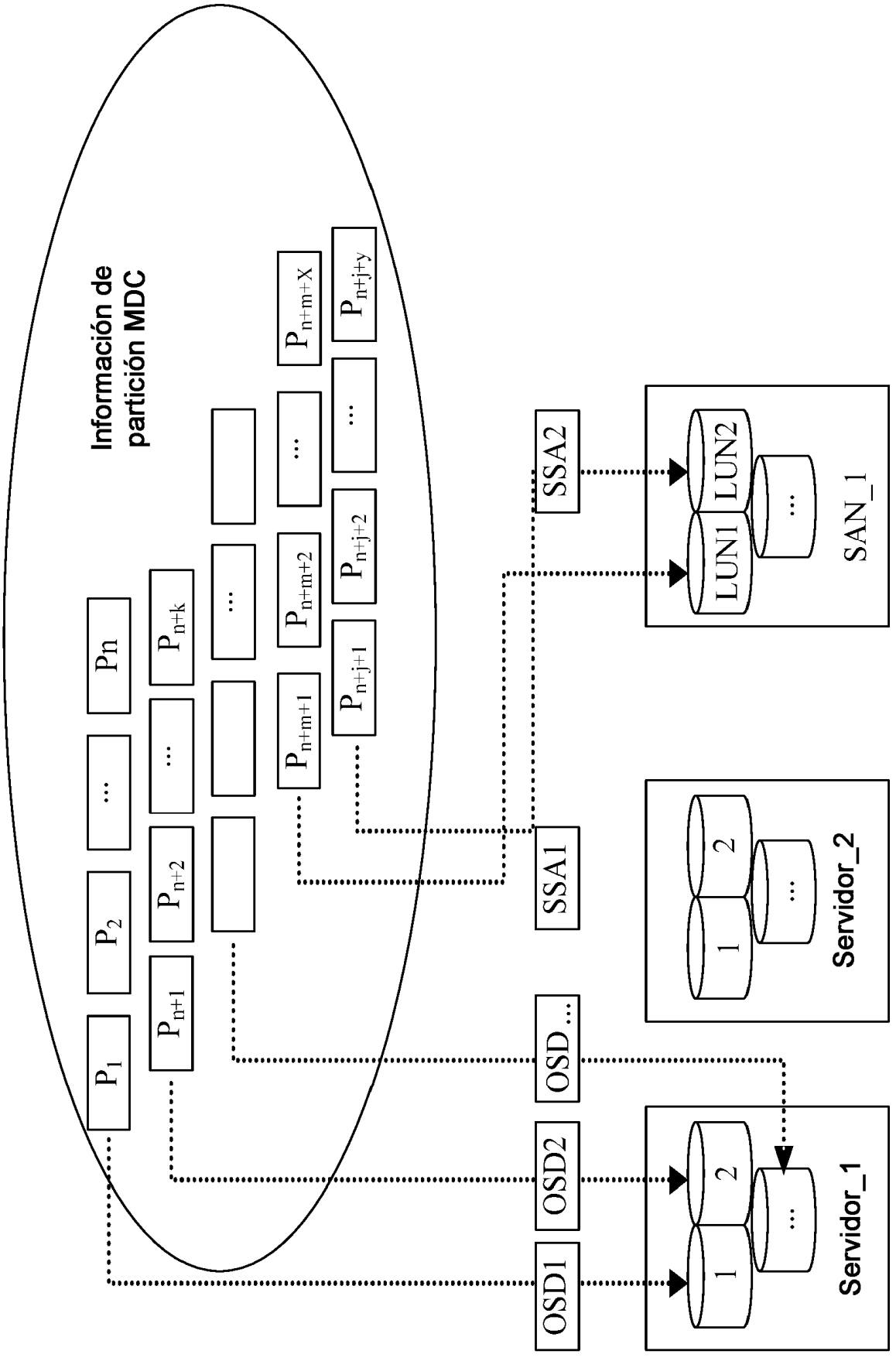
Información de partición MDC
P1
P2
...
Pn
Pn+1
Pn+2
...
Pn+k
Pn+m+1
Pn+m+2
...
Pn+m+X
Pn+j+1
Pn+j+2
...
Pn+j+y
OSD1
OSD2
OSD...
SSA1
SSA2
Servidor_1
Servidor_2
SAN_1
LUN1
LUN2
1
2
...

FIG. 2

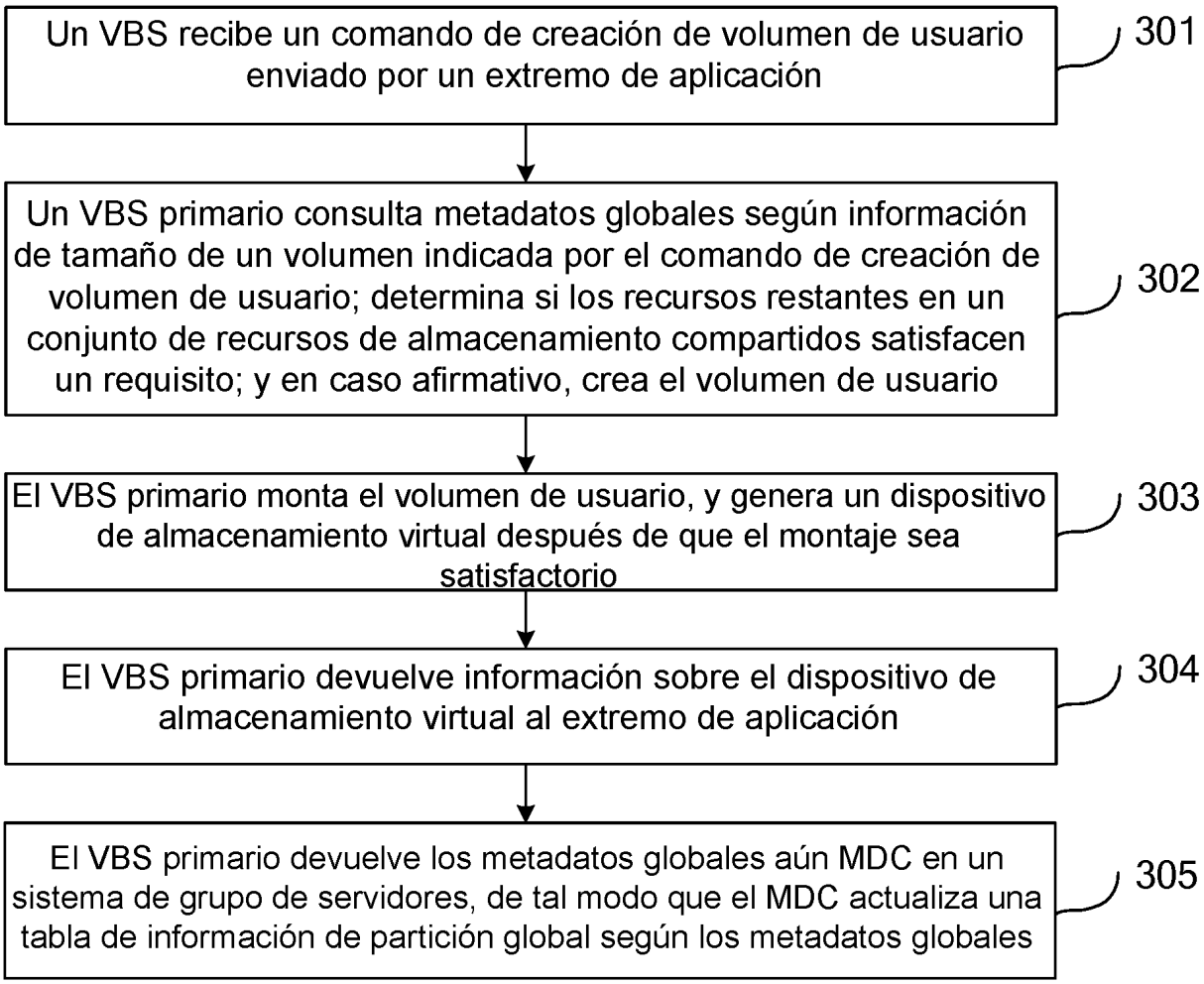
Un VBS recibe un comando de creación de volumen de usuario enviado por un extremo de aplicación
301
Un VBS primario consulta metadatos globales según información de tamaño de un volumen indicada por el comando de creación de volumen de usuario; determina si los recursos restantes en un conjunto de recursos de almacenamiento compartidos satisfacen un requisito; y en caso afirmativo, crea el volumen de usuario
302
El VBS primario monta el volumen de usuario, y genera un dispositivo de almacenamiento virtual después de que el montaje sea satisfactorio
303
El VBS primario devuelve información sobre el dispositivo de almacenamiento virtual al extremo de aplicación
304
El VBS primario devuelve los metadatos globales aún MDC en un sistema de grupo de servidores, de tal modo que el MDC actualiza una tabla de información de partición global según los metadatos globales
305

FIG. 3

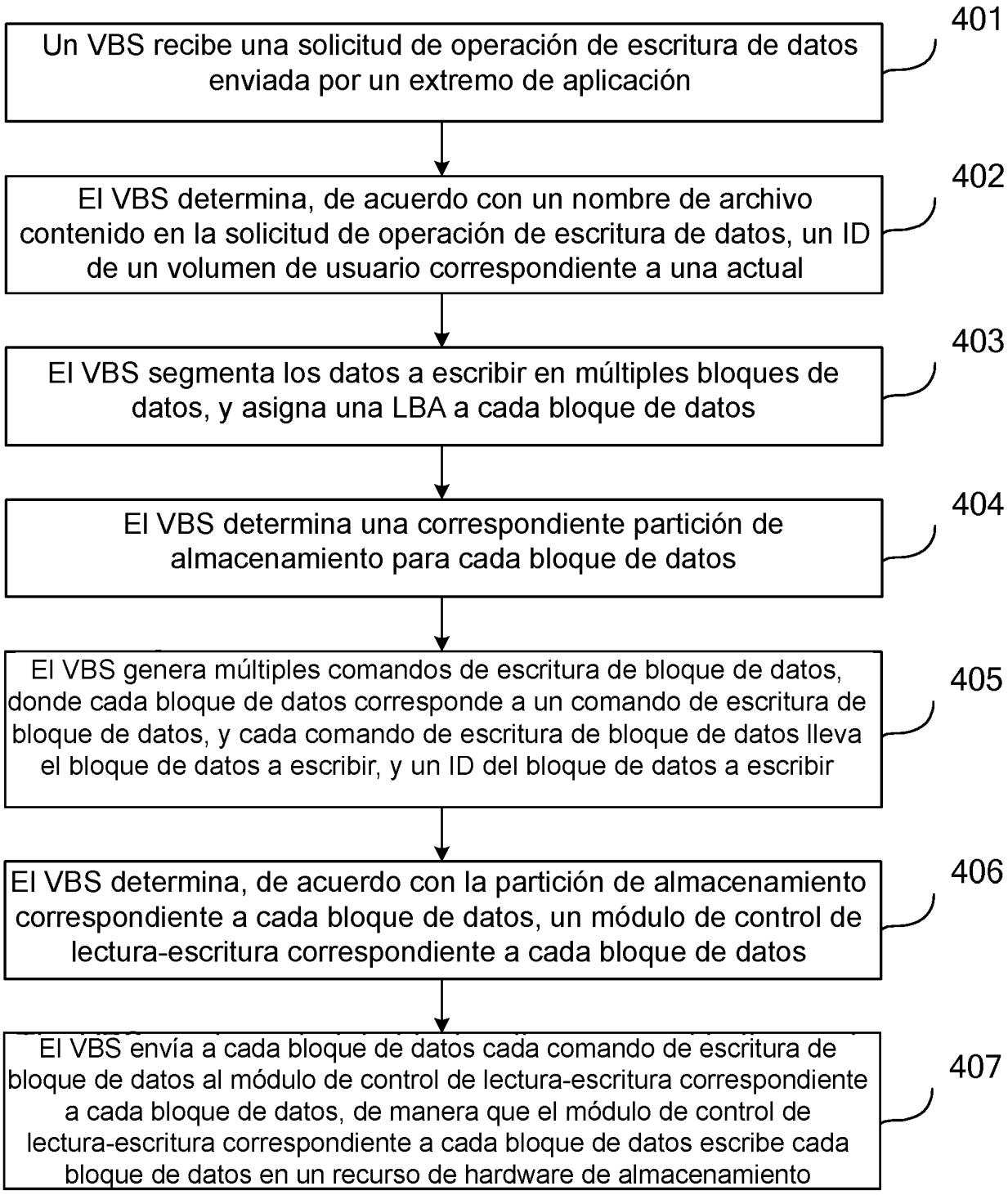
Un VBS recibe una solicitud de operación de escritura de datos enviada por un extremo de aplicación
401
El VBS determina, de acuerdo con un nombre de archivo contenido en la solicitud de operación de escritura de datos, un ID de un volumen de usuario correspondiente a una actual
402
El VBS segmenta los datos a escribir en múltiples bloques de datos, y asigna una LBA a cada bloque de datos
403
El VBS determina una correspondiente partición de almacenamiento para cada bloque de datos
404
El VBS genera múltiples comandos de escritura de bloque de datos, donde cada bloque de datos corresponde a un comando de escritura de bloque de datos, y cada comando de escritura de bloque de datos lleva el bloque de datos a escribir, y un ID del bloque de datos a escribir
405
El VBS determina, de acuerdo con la partición de almacenamiento correspondiente a cada bloque de datos, un módulo de control de lectura-escritura correspondiente a cada bloque de datos
406
El VBS envía a cada bloque de datos cada comando de escritura de bloque de datos al módulo de control de lectura-escritura correspondiente a cada bloque de datos, de manera que el módulo de control de lectura-escritura correspondiente a cada bloque de datos escribe cada bloque de datos en un recurso de hardware de almacenamiento
407

FIG. 4

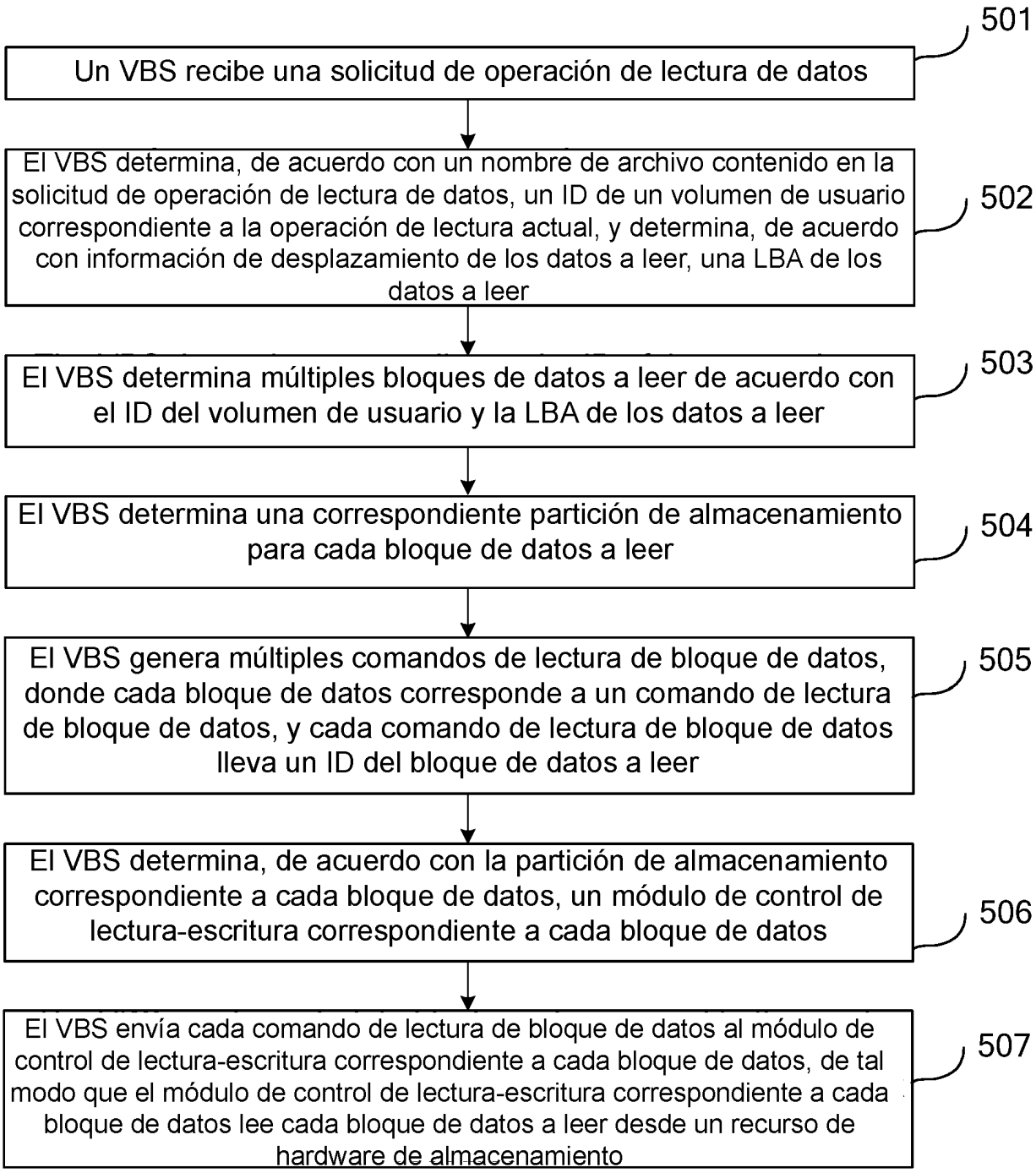
501
Un VBS recibe una solicitud de operación de lectura de datos
El VBS determina, de acuerdo con un nombre de archivo contenido en la solicitud de operación de lectura de datos, un ID de un volumen de usuario correspondiente a la operación de lectura actual, y determina, de acuerdo con información de desplazamiento de los datos a leer, una LBA de los datos a leer
502
El VBS determina múltiples bloques de datos a leer de acuerdo con el ID del volumen de usuario y la LBA de los datos a leer
503
El VBS determina una correspondiente partición de almacenamiento para cada bloque de datos a leer
504
El VBS genera múltiples comandos de lectura de bloque de datos, donde cada bloque de datos corresponde a un comando de lectura de bloque de datos, y cada comando de lectura de bloque de datos lleva un ID del bloque de datos a leer
505
El VBS determina, de acuerdo con la partición de almacenamiento correspondiente a cada bloque de datos, un módulo de control de lectura-escritura correspondiente a cada bloque de datos
506
El VBS envía cada comando de lectura de bloque de datos al módulo de control de lectura-escritura correspondiente a cada bloque de datos, de tal modo que el módulo de control de lectura-escritura correspondiente a cada bloque de datos lee cada bloque de datos a leer desde un recurso de hardware de almacenamiento
507

FIG. 5

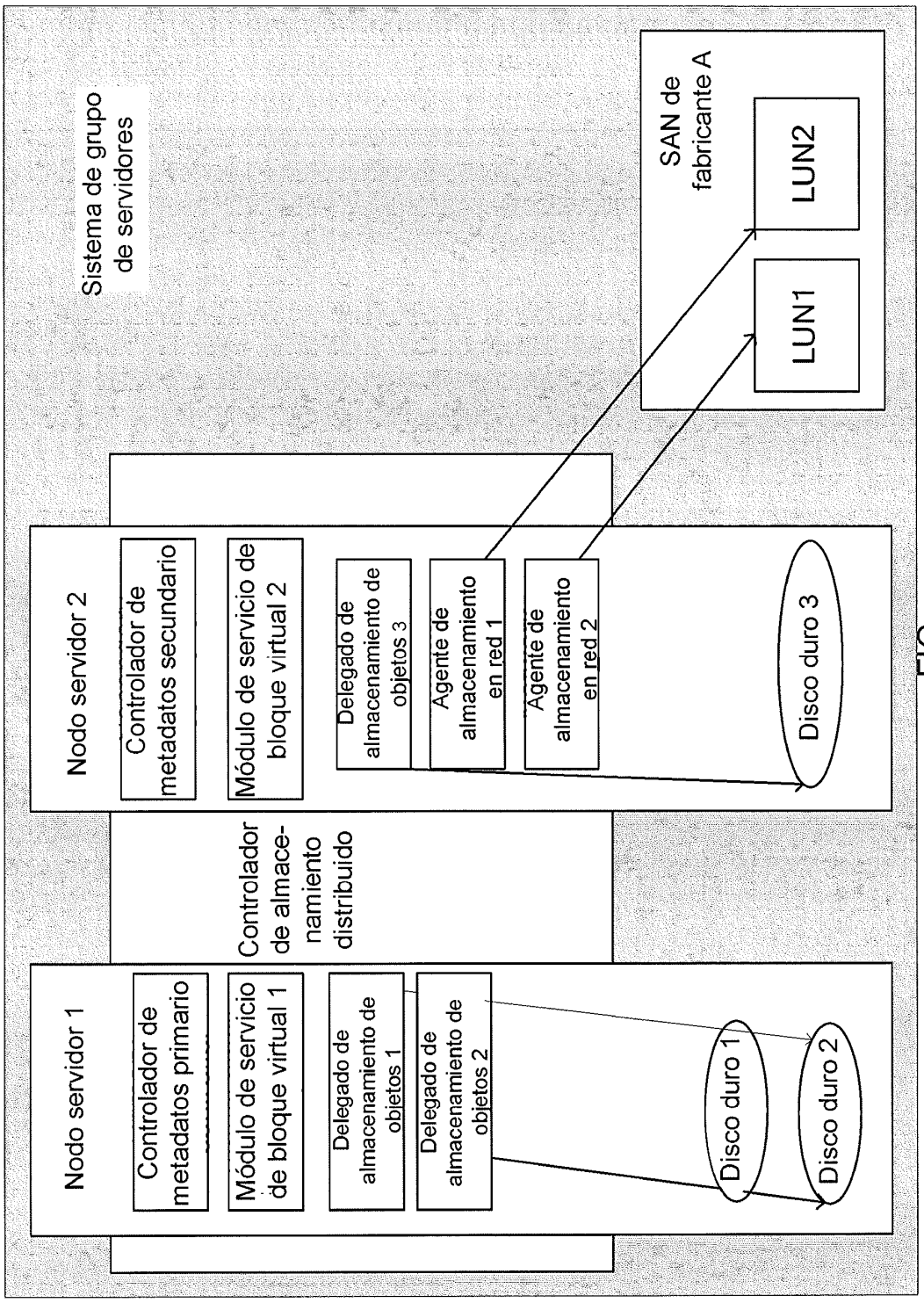
Sistema de grupo de servidores
Nodo servidor 1
Controlador de metadatos primario
Módulo de servicio de bloque virtual 1
Delegado de almacenamiento de objetos 1
Delegado de almacenamiento de objetos 2
Controlador de almace-namiento distribuido
Nodo servidor 2
Controlador de metadatos secundario
Módulo de servicio de bloque virtual 2
Delegado de almacenamiento de objetos 3
Agente de almacenamiento en red 1
Agente de almacenamiento en red 2
Disco duro 1
Disco duro 2
Disco duro 3
SAN de fabricante A
LUN1
LUN2

FIG. 6

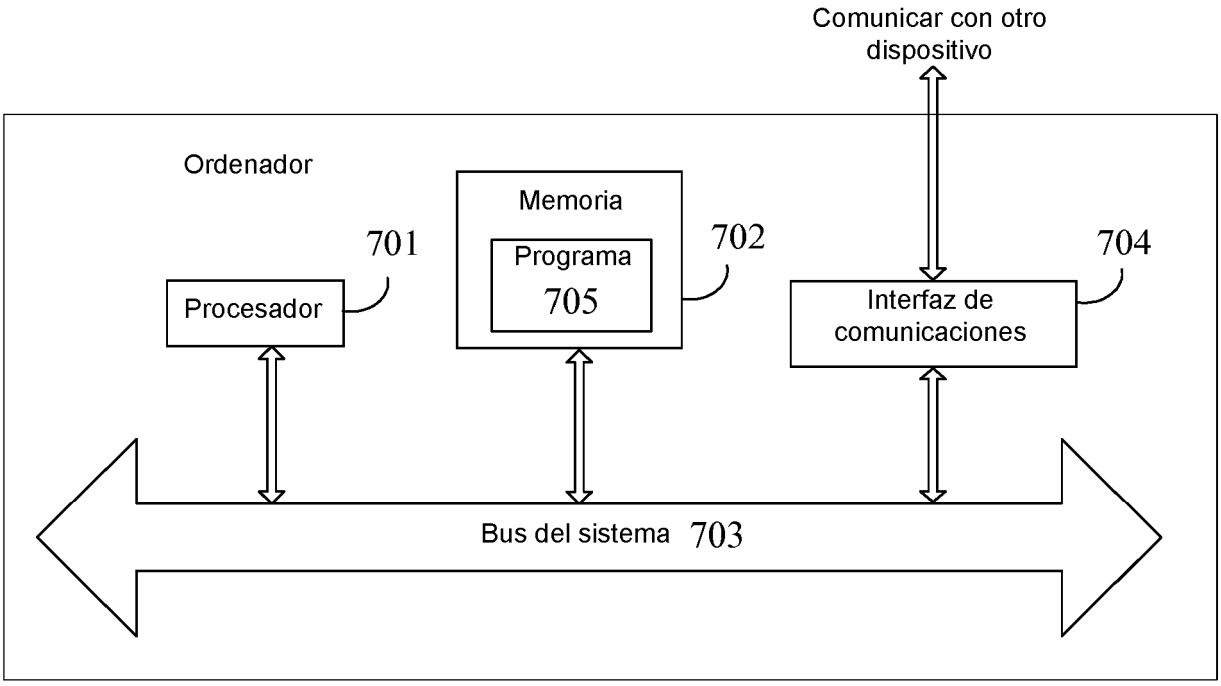
Comunicar con otro dispositivo
Ordenador
Memoria
Programa
705
701
702
704
Procesador
Interfaz de comunicaciones
Bus del sistema
703

FIG. 7