(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 507 072**

(51) Int. CI.:

G06F 12/08 (2006.01)
G06F 3/06 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **06.12.2007 E 07867661 (6)**

(97) Fecha y número de publicación de la concesión europea: **25.06.2014 EP 2109822**

(54) Título: **Aparato, sistema, y método para un almacenamiento de estado sólido como memoria caché para un almacenamiento no volátil de alta capacidad**

(30) Prioridad:

06.12.2006 US 873111 P
22.09.2007 US 974470 P

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
14.10.2014

(73) Titular/es:

FUSION-IO, INC. (100.0%)
2855 E Cottonwood Parkway Box 100
Salt Lake City, UT 84121, US

(72) Inventor/es:

FLYNN, DAVID;
STRASSER, JOHN;
THATCHER, JONATHAN y
ZAPPE, MICHAEL

(74) Agente/Representante:

ISERN JARA, Jorge

DESCRIPCIÓN

Aparato, sistema, y método para un almacenamiento de estado sólido como memoria caché para un almacenamiento no volátil de alta capacidad

Antecedentes de la invención

Campo de la invención

La presente invención se refiere a gestión de datos y más particularmente se refiere al uso de un almacenamiento de estado sólido como memoria caché para dispositivos de almacenamiento no volátiles de alta capacidad

Descripción de las técnicas relacionadas

En general, la memoria caché es ventajosa porque los datos que se acceden a menudo o que se cargan como parte de una aplicación o sistema operativo se pueden almacenar en memoria caché con un acceso posterior mucho más rápido que cuando los datos se tiene que acceder a través de un dispositivo de almacenamiento no volátil de alta capacidad ("HCNV"), tal como una unidad de disco duro ("HDD"), una unidad óptica, un almacenamiento de cinta, etc. La memoria caché está usualmente incluida en un ordenador. Los documentos WO 02/01365, US 2005/0177687 y "Windows PC Accelerators" de Microsoft muestran el uso de la memoria no volátil como caché.

Sumario de la invención

Algunos dispositivos de almacenamiento y sistemas incluyen memoria caché en los dispositivos de almacenamiento de HCHV. Algunos dispositivos de almacenamiento HCNV contienen memoria caché de estado de estado sólido no volátil, esta proporciona el beneficio de reducción de los tiempos de acceso pero solo puede proporcionar un funcionamiento consistente con la capacidad usualmente limitada de la interfaz del dispositivo de almacenamiento HCNV. Existen algunos dispositivos de almacenamiento de memoria caché de estado sólido no volátiles que están usualmente situados en la placa base; estos dispositivos no se pueden usar en entornos multi-usuario ya que no se proporciona la coherencia de la memoria caché. Algunos controladores de dispositivos HCNV también incluyen memoria caché. Cuando los controladores de la memoria caché de HCNV redundante se comparten entre múltiples clientes, se requieren algoritmos sofisticados de coherencia de la memoria caché para asegurar que no se corrompen los datos.

Usualmente, las memorias caché se implementan en DRAM, que tienen como objetivo la capacidad de la memoria caché y que requieren una potencia de funcionamiento relativamente alta. Si se pierde la potencia que soporta la memoria caché volátil, se pierden los datos almacenados en la memoria caché. Usualmente, se usa alguna batería de respaldo para evitar la pérdida de datos en el caso de fallo de potencia, con suficiente capacidad para traspasar la memoria caché a la memoria no volátil antes de que falle la batería de respaldo. Además, los sistemas de batería de respaldo consumen potencia, requieren redundancia, impactan negativamente en la fiabilidad y consumen espacio. Las baterías también se tienen que mantener en una base regular y la baterías de respaldo pueden ser relativamente cara.

A partir de la discusión anterior, debería ser evidente que existe una necesidad de un aparato, sistema y método que gestionen los datos usando un almacenamiento de estado sólido como memoria caché. Ventajosamente, tal aparato, sistema, y método proporcionarían una memoria caché no volátil que consume poca potencia, proporciona una capacidad significativamente mayor y no requiere una batería de respaldo para mantener los datos almacenados en la memoria caché.

La presente invención se ha desarrollado en respuesta al presente estado de la técnica, y en particular en respuesta a los problemas y necesidades en la técnica que no se han resuelto aun completamente por los sistemas disponibles actualmente para la gestión del almacenamiento de datos. Por consiguiente, la presente invención se ha desarrollado para proporcionar un aparato, sistema, y método para gestionar el almacenamiento de datos sobre uno o más dispositivos de almacenamiento no volátil de alta capacidad ("HCNV") que superen muchos o todos los inconvenientes tratados anteriormente en la técnica.

El aparato se proporciona, en una realización con una pluralidad de módulos incluyendo un módulo del extremo frontal de la memoria caché y un módulo del extremo posterior de la memoria caché. El módulo del extremo frontal de la memoria caché gestiona las transferencias de datos asociadas con una petición de almacenamiento. Las transferencias de datos son entre un dispositivo solicitante y una función de almacenamiento de estado sólido como memoria caché para uno o más dispositivos de almacenamiento HCNV, y las transferencias de datos pueden incluir uno o más de datos, metadatos, e índices de metadatos. El almacenamiento de estado sólido puede incluir una red de elementos de almacenamiento de datos de estado sólido, no volátil. El módulo del extremo posterior de la memoria caché gestiona las transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

En una realización del aparato el módulo del extremo frontal de la memoria caché y el módulo del extremo posterior de la memoria caché están localizadas conjuntamente con un controlador de almacenamiento de estado sólido que gestiona el almacenamiento de estado sólido. En una realización adicional, el módulo del extremo frontal de la memoria caché, el módulo del extremo posterior de la memoria caché y el controlador del almacenamiento de estado sólido pueden operar de forma autónoma desde el dispositivo solicitante.

En una realización, el aparato incluye un módulo de RAID de HCNV que almacena datos almacenados en memoria caché en el almacenamiento de estado sólido en dos o más dispositivos de almacenamiento HCNV en una red redundante de unidades independientes ("RAID") consistente con un nivel de RAID. Los datos pueden aparecer a un dispositivo solicitante como un conjunto. En otra realización, el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV pueden incluir un dispositivo de almacenamiento híbrido dentro de un conjunto de dispositivos de almacenamiento híbrido que está configurado como un grupo de RAID. Un segmento de datos almacenados en memoria caché en el almacenamiento de estado sólido y más tarde almacenados sobre un dispositivo de HCNV puede incluir uno de N segmentos de datos de una banda o un segmento de datos de paridad de la banda. El dispositivo de almacenamiento híbrido usualmente recibe peticiones de almacenamiento de uno o más clientes independientes de los segmentos de datos de una banda de RAID. En una realización adicional, el dispositivo de almacenamiento híbrido puede ser un dispositivo de almacenamiento de un grupo de RAID distribuido de extremo frontal, compartido que recibe dos o más peticiones de almacenamiento simultáneas desde dos o más clientes.

En una realización adicional del aparato, el dispositivo de almacenamiento HCNV puede ser una unidad de disco duro ("HDD"), una unidad óptica, un almacenamiento de cinta. En otra realización, el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV pueden ser un dispositivo de almacenamiento híbrido. En una realización, el aparato también puede incluir un módulo de emulación del dispositivo normalizado que proporciona acceso al dispositivo de almacenamiento híbrido emulando un dispositivo normalizado conectado a uno o más dispositivos solicitantes antes de cargar el uno o más dispositivos solicitantes con código específico para la operación del dispositivo de almacenamiento híbrido. El dispositivo normalizado usualmente se puede soportar por la normativa BIOS de la industria.

En otra realización, el dispositivo de almacenamiento de estado sólido se puede dividir en dos o más regiones, en el que una o más particiones se pueden usar como un almacenamiento de estado sólido independiente del almacenamiento de estado sólido que funciona como memoria caché para los dispositivos de almacenamiento HCNV. En otra realización más, uno o más clientes envían los mensajes de control de memoria caché al módulo del extremo frontal de la memoria caché y al módulo del extremo posterior de la memoria caché para gestionar el estado de uno o más ficheros u objetos almacenados dentro del dispositivo de almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

En una realización del aparato, los mensajes de control de la memoria caché pueden incluir uno o más mensajes de control. Diversas realizaciones de los mensajes de control pueden incluir un mensaje de control que causa que el módulo del extremo posterior de la memoria caché ancle una porción de un objeto o fichero en el almacenamiento de estado sólido o un mensaje de control que causa que el módulo del extremo posterior de la memoria caché desancle una porción de un objeto o fichero en el almacenamiento de estado sólido. Otras realizaciones de los mensajes de control pueden incluir un mensaje de control que causa que el módulo del extremo posterior de la memoria caché traspase una porción de un objeto o fichero desde el almacenamiento de estado sólido al uno o más dispositivos de almacenamiento HCNV o un mensaje de control causa que el módulo del extremo posterior de la memoria caché precargue una porción de un objeto o fichero al almacenamiento de estado sólido desde el uno o más dispositivos de almacenamiento HCNV. Otra realización más de un mensaje de control puede ser un mensaje de control que causa que el módulo del extremo posterior de la memoria caché descargue una o más porciones de uno o más objetos o ficheros desde el almacenamiento de estado sólido al uno o más dispositivos de almacenamiento HCNV para liberar una cantidad determinada de espacio de almacenamiento en el almacenamiento de estado sólido. En una realización, los mensajes de control de la memoria caché se comunican mediante metadatos ("metadatos de control de la memoria caché") para el objeto o fichero. En una realización adicional, los metadatos de control de la memoria caché pueden ser persistentes. En otra realización, los metadatos de control de la memoria caché se pueden establecer a través de atributos establecidos en el momento de creación del fichero u objeto. En un otra realización más, los metadatos de control de la memoria caché se pueden obtener a partir de un sistema de gestión del fichero u objeto.

En una realización del aparato, el aparato puede incluir un elemento de almacenamiento de la memoria caché volátil en el que el módulo del extremo frontal de la memoria caché y el módulo del extremo posterior de la memoria caché almacenan datos en el elemento de almacenamiento de la memoria caché volátil y gestionan los datos almacenados en el almacenamiento de estado sólido y el elemento de almacenamiento de la memoria caché volátil. El módulo del extremo posterior puede gestionar además transferencias de datos entre el elemento de almacenamiento de la memoria caché volátil, el almacenamiento de estado sólido y los dispositivos de almacenamiento HCNV. En una realización adicional, los metadatos y/o los metadatos de índices para los objetos y ficheros almacenados en los dispositivos de almacenamiento HCNV se pueden mantener dentro del dispositivo de almacenamiento de estado sólido y el elemento de almacenamiento de la memoria caché volátil.

En una realización adicional del aparato, los metadatos y/o los metadatos de índices para objetos y ficheros almacenados en los dispositivos de almacenamiento HCNV se pueden mantener dentro del dispositivo de almacenamiento de estado sólido. En otra realización, el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV pueden incluir un dispositivo de almacenamiento de modo que los dispositivos de almacenamiento HCNV están ocultos de la vista del cliente conectado al dispositivo de almacenamiento.

También se presenta un sistema de la presente invención. El sistema sustancialmente incluye los módulos y realizaciones descritos anteriormente con respecto al aparato. En una realización, el sistema incluye un almacenamiento de estado sólido que incluye una red de elementos de almacenamiento de datos de estado sólido no volátil. El sistema también incluye uno o más dispositivos de almacenamiento HCNV y un controlador de almacenamiento. El controlador de almacenamiento, en una realización, puede incluir un controlador de almacenamiento de estado sólido y un controlador del dispositivo de almacenamiento HCNV. El controlador de almacenamiento puede incluir también un módulo del extremo frontal de la memoria caché y un módulo del extremo posterior de la memoria caché. El módulo del extremo frontal de la memoria caché gestiona las transferencias de datos asociadas con una petición de almacenamiento. Las transferencias de datos son usualmente entre un dispositivo solicitante y el almacenamiento de estado sólido que funciona como una memoria caché para el uno o más dispositivos de almacenamiento HCNV. Las transferencias de datos pueden incluir uno o más de datos, metadatos, e índices de metadatos. El módulo del extremo posterior de la memoria caché gestiona las transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

En una realización, el sistema incluye una interfaz de red conectada al controlador de almacenamiento, en el que la interfaz de red facilita transferencias de datos entre el dispositivo solicitante y el controlador del almacenamiento de estado sólido a través de una red de ordenadores. En otra realización, el sistema incluye un servidor que incluye el almacenamiento de estado sólido, el uno o más dispositivos de almacenamiento HCNV, y el controlador de almacenamiento. En otra realización más, el uno o más dispositivos de almacenamiento HCNV se conectan al controlador de almacenamiento a través de una red del área de almacenamiento ("SAN").

Un método de la presente invención también se presenta para compartir un dispositivo entre múltiples ordenadores. El método en las realizaciones desveladas sustancialmente incluye las etapas necesarias para realizar las funciones presentadas anteriormente con respecto a la operación del aparato y el sistema descritos. En una realización, el método incluye gestionar las transferencias de datos asociadas con una petición de almacenamiento, en el que las transferencias de datos son entre un dispositivo solicitante y el almacenamiento de estado sólido que funciona como una memoria caché para uno o más dispositivos de almacenamiento HCNV. Las transferencias de datos pueden incluir uno o más de datos, metadatos e índices de metadatos. El almacenamiento de estado sólido puede incluir una red de elementos de almacenamiento de datos de estado sólido. El método también puede incluir gestionar transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

Un aspecto de la invención proporciona un aparato para gestionar el almacenamiento de datos sobre uno o más dispositivos de almacenamiento no volátil, de alta capacidad ("HCNV") comprendiendo el aparato: un módulo del extremo frontal de la memoria caché que gestiona las transferencias de datos asociadas con una petición de almacenamiento, las transferencias de datos entre un dispositivo solicitante y un almacenamiento de estado sólido que funciona como una memoria caché para uno o más dispositivos de almacenamiento HCNV, comprendiendo las transferencias de datos uno o más de datos, metadatos e índices de metadatos, comprendiendo el almacenamiento de estado sólido una red de elementos de almacenamiento de datos de estado sólido, no volátil; y un módulo del extremo posterior de la memoria caché que gestiona las transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

El módulo del extremo frontal de la memoria caché y el módulo del extremo posterior de la memoria caché pueden estar localizados conjuntamente con un controlador de almacenamiento de estado sólido que gestiona el almacenamiento de estado sólido.

El módulo del extremo frontal de la memoria caché, el módulo del extremo posterior de la memoria caché y el controlador del almacenamiento de estado sólido pueden operar de forma autónoma del dispositivo solicitante.

El controlador de estado sólido puede comprender además un módulo controlador del almacenamiento de objetos que sirve las peticiones de objetos desde uno o más dispositivos solicitantes y gestiona los objetos de las peticiones de objetos dentro del almacenamiento de estado sólido.

El aparato puede comprender además un módulo RAID de HCNV que almacena datos en memoria caché en el almacenamiento de estado sólido en dos o más dispositivos de almacenamiento HCNV en una red redundante de unidades independientes ("RAID") consistente con un nivel de RAID, en el que los datos aparecen para un dispositivo solicitante como un todo.

El almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV pueden comprender un dispositivo de almacenamiento híbrido dentro de un conjunto de dispositivos de almacenamiento híbrido que está configurado como un grupo de RAID, en el que un segmento de datos en memoria caché en el almacenamiento de estado sólido y almacenado más tarde sobre un dispositivo de HCNV comprende uno de N segmentos de datos de una banda o un segmento de datos de paridad de la banda, en el que el dispositivo de almacenamiento híbrido recibe peticiones de almacenamiento desde uno o más clientes independientes de segmentos de datos de una banda de RAID.

El dispositivo de almacenamiento híbrido puede ser un dispositivo de almacenamiento de un grupo de RAID distribuido de extremos de entrada compartidos que recibe dos o más peticiones de almacenamiento simultáneas desde uno o más clientes.

El dispositivo de almacenamiento HCNV puede ser uno de una unidad de disco duro ("HDD"), una unidad óptica, y un almacenamiento de cinta.

El almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV pueden comprender un dispositivo de almacenamiento híbrido y pueden comprender además un módulo de emulación de dispositivo normalizado que proporciona acceso al dispositivo de almacenamiento híbrido emulando un dispositivo normalizado conectado al uno o más dispositivos solicitantes antes de cargar el uno o más dispositivos solicitantes con código específico para la operación del dispositivo de almacenamiento híbrido, soportándose el dispositivo normalizado por la normativa BIOS de la industria.

El dispositivo de almacenamiento de estado sólido se puede dividir en dos o más regiones, y se pueden usar una o más particiones como almacenamiento de estado sólido independiente del almacenamiento de estado sólido que funciona como una memoria caché para los dispositivos de almacenamiento HCNV.

Uno o más clientes pueden enviar mensajes de control de la memoria caché al módulo del extremo frontal de la memoria caché y el módulo del extremo posterior de la memoria caché para gestionar el estado de uno o más ficheros u objetos almacenados dentro del dispositivo de almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

El mensaje de control de la memoria caché puede comprender uno o más de: un mensaje de control que causa que el módulo del extremo posterior de la memoria caché ancle una porción de un objeto o fichero en el almacenamiento de estado sólido; un mensaje de control que causa que el módulo del extremo posterior de la memoria caché desancle una porción de un objeto o fichero en el almacenamiento de estado sólido; un mensaje de control que causa que el módulo del extremo posterior de la memoria caché traspase una porción de un objeto o fichero desde el almacenamiento de estado sólido al uno o más dispositivos de almacenamiento HCNV; un mensaje de control que causa que el módulo del extremo posterior de la memoria caché precargue una porción de un objeto o fichero al almacenamiento de estado sólido desde el uno o más dispositivos de almacenamiento HCNV; un mensaje de control que causa que el módulo del extremo posterior de la memoria caché descargue una o más porciones de uno o más objetos o ficheros desde el almacenamiento de estado sólido al uno o más dispositivos de almacenamiento HCNV para liberar una cantidad determinada de espacio de almacenamiento en el almacenamiento de estado sólido.

Los mensajes de control de la memoria caché se pueden comunicar mediante metadatos ("metadatos de control de la memoria caché") para el objeto o fichero.

Los metadatos del control de la memoria caché pueden ser persistentes.

Los metadatos del control de la memoria caché se pueden establecer mediante de atributos fijados en el momento de creación del fichero u objeto.

Los metadatos del control de la memoria caché se pueden obtener a partir de un sistema de gestión de ficheros u objetos.

El aparato puede comprender además un elemento de almacenamiento de la memoria caché volátil y el módulo del extremo frontal de la memoria caché y el módulo del extremo posterior de la memoria caché pueden comprender además almacenar datos en el elemento de almacenamiento de la memoria caché volátil y gestionar los datos almacenados en el almacenamiento de estado sólido y el elemento de almacenamiento de la memoria caché volátil, y el módulo del extremo posterior puede gestionar además transferencias de datos entre el elemento de almacenamiento de la memoria caché volátil, el almacenamiento de estado sólido y los dispositivos de almacenamiento HCNV.

Uno o más de los metadatos, los metadatos de índices para objetos y ficheros almacenados en los dispositivos de almacenamiento HCNV se pueden mantener dentro del dispositivo de almacenamiento de estado sólido y el elemento de almacenamiento de la memoria caché volátil.

Uno o más de los metadatos y los metadatos de índices para objetos y ficheros almacenados en los dispositivos de almacenamiento HCNV se pueden mantener dentro del dispositivo de almacenamiento de estado sólido.

El almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV pueden comprender un dispositivo de almacenamiento de modo que los dispositivos de almacenamiento HCNV están ocultos de la vista del cliente conectado al dispositivo de almacenamiento.

Un aspecto de la invención proporciona un sistema para la gestión del almacenamiento de datos sobre uno o más dispositivos de almacenamiento no volátil de alta capacidad ("HCNV"), comprendiendo el sistema: un almacenamiento de estado sólido que comprende una red de elementos de almacenamiento de datos de estado sólido no volátil; uno o más dispositivos de almacenamiento HCNV; y un controlador de almacenamiento que comprende un controlador del almacenamiento de estado sólido; un controlador del dispositivo de almacenamiento HCNV; un módulo del extremo frontal de la memoria caché que gestiona transferencias de datos asociadas con una petición de almacenamiento, siendo las transferencias de datos entre un dispositivo solicitante y el almacenamiento de estado sólido que funciona como memoria caché para el uno o más dispositivos de almacenamiento HCNV, comprendiendo las transferencias de datos uno o más de datos, metadatos e índices de metadatos; y un módulo del extremo posterior de la memoria caché que gestiona las transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

El sistema puede comprender además una interfaz de red conectada al controlador de almacenamiento, facilitando la interfaz de red las transferencias de datos entre el dispositivo solicitante y el controlador de almacenamiento de estado sólido a través de una red de ordenadores.

El sistema puede comprender además un servidor, en el que el servidor incluye el almacenamiento de estado sólido, el uno o más dispositivos de almacenamiento HCNV, y el controlador de almacenamiento.

El uno o más dispositivos de almacenamiento HCNV se pueden conectar a un controlador de almacenamiento a través de una red del área de almacenamiento ("SAN").

Otro aspecto de la invención proporciona un producto de programa de ordenador que comprende un medio legible por ordenador que tiene un código ejecutable de programa utilizable por el ordenador que se puede ejecutar para realizar las operaciones para la gestión del almacenamiento de datos sobre uno o más dispositivos de almacenamiento no volátil de alta capacidad ("HCNV"), comprendiendo las operaciones del producto de programa de ordenador: gestionar las transferencias de datos asociadas con una petición de almacenamiento, las transferencias de datos entre un dispositivo solicitante y un almacenamiento de estado sólido que funciona como una memoria caché para uno o más dispositivos de almacenamiento HCNV, comprendiendo las transferencias de datos uno o más de datos, metadatos e índices de metadatos, comprendiendo el almacenamiento de estado sólido una red de elementos de almacenamiento de datos de estado sólido no volátil; y gestionar las transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

La referencia a lo largo de esta memoria descriptiva a características, ventajas o lenguaje similar no implica que todas las características ventajas que se pueden realizar con la presente invención debieran estar o que estén en cualquier realización única de la invención. Más bien, el lenguaje referente a las características y ventajas se entiende que significa que la propiedad, ventaja o característica específica descrita en conexión con una realización está incluida en al menos una realización de la presente invención. De este modo, la discusión de las características y ventajas, y lenguaje similar, a lo largo de la presente memoria descriptiva se puede referir, pero no necesariamente, a la misma realización.

Además, las propiedades, ventajas y características descritas de la invención se pueden combinar en cualquier forma adecuada en una o más realizaciones. Un experto en la materia relevante reconocerá que la invención se puede poner en práctica sin una o más de las características específicas o ventajas de una realización particular. En otros casos, se pueden reconocer características y ventajas adicionales en ciertas realizaciones que puede que no estén presentes en todas las realizaciones de la invención.

Estas características y ventajas de la presente invención se harán totalmente evidentes a partir de la siguiente descripción y reivindicaciones adjuntas, o se puede aprender por la puesta en práctica de la invención como se muestra en este documento más adelante.

Breve descripción de los dibujos

Para que las ventajas de la invención se entiendan con facilidad, se proporcionará una descripción más particular de la invención descrita brevemente anteriormente por referencia a realizaciones específicas que se ilustran en los dibujos adjuntos. Entendiendo que estos dibujos representan solo realizaciones típicas de la invención y que por lo tanto no se deben considerar como limitativas de su ámbito, la invención se describirá y explicará con especificidad y detalle adicionales mediante el uso de los dibujos adjuntos, en los que:

la Figura 1A es un diagrama de bloques esquemático que ilustra una realización de un sistema para la gestión de datos en un dispositivo de almacenamiento de estado sólido de acuerdo con la presente invención;
la Figura 1B es un diagrama de bloques esquemático que ilustra una realización de un sistema para la gestión de objetos en un dispositivo de almacenamiento de acuerdo con la presente invención;
la Figura 1C es un diagrama de bloques esquemático que ilustra una realización de un sistema para una red del área de almacenamiento en servidor de acuerdo con la presente invención;
la Figura 2A es un diagrama de bloques esquemático que ilustra una realización de un aparato para la gestión de objetos en un dispositivo de almacenamiento de acuerdo con la presente invención;
la Figura 2B es un diagrama de bloques esquemático que ilustra una realización de un controlador del dispositivo de almacenamiento de estado sólido en un dispositivo de almacenamiento de estado sólido de acuerdo con la presente invención;
la Figura 3 es un diagrama de bloques esquemático que ilustra una realización de un controlador del almacenamiento de estado sólido con una conducción de datos de escritura y una conducción de datos de lectura en un dispositivo de almacenamiento de estado sólido de acuerdo con la presente invención;
la Figura 4A es un diagrama de bloques esquemático que ilustra una realización de un controlador de intercalado de bancos en el controlador de almacenamiento de estado sólido de acuerdo con la presente invención;
la Figura 4B es un diagrama de bloques esquemático que ilustra una realización alternativa de un controlador de intercalado de bancos en el controlador del almacenamiento de estado sólido de acuerdo con la presente invención;
la Figura 5A es un diagrama de flujo esquemático que ilustra una realización de un método para la gestión de datos en un dispositivo de almacenamiento de estado sólido que usa una conducción de datos de acuerdo con la presente invención;
la Figura 5B es un diagrama de flujo esquemático que ilustra una realización de un método para una SAN en servidor de acuerdo con la presente invención;
la Figura 6 es un diagrama de flujo esquemático que ilustra otra realización de un método para la gestión de datos en un dispositivo de almacenamiento de estado sólido que usa una conducción de datos de acuerdo con la presente invención;
la Figura 7 es un diagrama de flujo esquemático que ilustra una realización de un método para la gestión de datos en un dispositivo de almacenamiento de estado sólido que usa un intercalado de bancos de acuerdo con la presente invención;
la Figura 8 es un diagrama de bloques esquemático que ilustra una realización de un aparato para la recogida de basura en un dispositivo de almacenamiento de estado sólido de acuerdo con la presente invención;
la Figura 9 es un diagrama de flujo esquemático que ilustra otra realización de un método para la recogida de basura en un dispositivo de almacenamiento de estado sólido de acuerdo con la presente invención;
la Figura 10 es un diagrama de bloques esquemático que ilustra una realización de un sistema para una RAID progresiva, una RAID distribuida del extremo frontal, una RAID distribuida y compartida del extremo frontal de acuerdo con las presentes invenciones;
la Figura 11 es un diagrama de bloques esquemático que ilustra una realización de un aparato para una RAID distribuida del extremo frontal de acuerdo con la presente invención;
la Figura 12 es un diagrama de flujo esquemático que ilustra una realización de un método para una RAID distribuida del extremo frontal de acuerdo con la presente invención;
la Figura 13 es un diagrama de bloques esquemático que ilustra una realización de un aparato para una RAID distribuida compartida del extremo frontal de acuerdo con la presente invención;
la Figura 14 es un diagrama de flujo esquemático que ilustra una realización de un método para una RAID distribuida compartida del extremo frontal de acuerdo con la presente invención;
la Figura 15 es un diagrama de bloques esquemático que ilustra una realización de un sistema con almacenamiento de estado sólido como memoria caché para un dispositivo de almacenamiento no volátil de alta capacidad de acuerdo con la presente invención;
la Figura 16 es un diagrama de bloques esquemático que ilustra una realización de un aparato con almacenamiento de estado sólido como memoria caché para un dispositivo de almacenamiento no volátil de alta capacidad de acuerdo con la presente invención;
la Figura 17 es un diagrama de flujo esquemático que ilustra una realización de un método con almacenamiento de estado sólido como memoria caché para un dispositivo de almacenamiento no volátil de alta capacidad de acuerdo con la presente invención;

Descripción detallada de la invención

Muchas de las unidades funcionales descritas en esta memoria descriptiva se han etiquetado como módulos, para enfatizar más particularmente en su independencia de implementación. Por ejemplo, un módulo se puede implementar como un circuito hardware que comprende circuitos VLSI a medida o redes de puertas, o semiconductores distribuidos tales como chips de lógica, transistores u otros componentes discretos. Un módulo también se puede implementar en dispositivos de hardware programable tales como redes de puertas programables en campo, lógica de red programable, dispositivos de lógica programable o similares.

Los módulos también se pueden implementar en software para su ejecución por diversos tipos de procesadores. Un módulo identificado de código ejecutable puede comprender, por ejemplo, uno o más bloques físicos o lógicos de

instrucciones de ordenador que se pueden organizar, por ejemplo, como un objeto, procedimiento o función. Sin embargo, las instrucciones ejecutables de un módulo identificado no necesariamente están localizadas juntas físicamente, sino que pueden comprender instrucciones dispares almacenadas en diferentes localizaciones que cuando se juntan lógicamente, comprenden el módulo y consiguen el propósito establecido para el módulo.

En efecto, un módulo de código ejecutable puede ser una única instrucción, o muchas instrucciones y puede incluso estar distribuido sobre varios segmentos de código diferentes entre diferentes programas y a través de varios dispositivos de memoria. De forma similar, los datos operativos se pueden identificar e ilustrar en este documento dentro de módulos, y se pueden realizar en cualquier forma adecuada y organizarse dentro de cualquier tipo adecuado de estructura de datos. Los datos operativos se pueden recoger como un conjunto de datos únicos, o pueden estar distribuidos sobre diferentes localizaciones incluyendo sobre dispositivos de almacenamiento diferentes, y puede existir, al menos parcialmente, meramente como señales electrónicas sobre un sistema o red. Donde un módulo o porciones de un módulo se implementan en software, las porciones de software se almacenan sobre uno o más medios legibles por ordenador.

A lo largo de esta memoria descriptiva se hace referencia a "una realización" o lenguaje similar que significa que una propiedad, estructura o característica particular descrita en conexión con la realización está incluida en al menos una realización de la presente invención. De este modo, las apariciones de la frase "en una realización" y lenguaje similar a lo largo de esta memoria descriptiva se puede referir, pero no necesariamente a la misma realización.

La referencia a un medio de transporte de señal puede tomar cualquier forma capaz de generar una señal, que causa que se genere una señal, o que causa la ejecución de un programa de instrucciones legibles por máquina sobre un aparato de procesamiento digital. El medio que transporta una señal se puede realizar por una línea de transmisión, un disco compacto, un disco de video digital, una cinta magnética, una unidad de Bernouilli, un disco magnético, una tarjeta perforada, una memoria flash, circuitos integrados u otro dispositivo de memoria de un aparato de procesamiento digital.

Además, las propiedades, estructuras o características descritas de la invención se pueden combinar en cualquier modo adecuado en una o más realizaciones. En la siguiente descripción, se proporcionan, numerosos detalles específicos, tales como ejemplos de programación, módulos software, selecciones de usuario, transacciones de red, consultas de bases de datos, estructuras de bases de datos, módulos hardware, circuitos hardware, chips hardware, etc. para proporcionar un entendimiento completo de las realizaciones de la invención. Un experto en la materia relevante reconocerá, sin embargo que la invención se puede poner en práctica sin uno o más de los detalles específicos o con otros métodos, componentes, materiales y así sucesivamente. En otros casos, no se muestran o se describen con detalle estructuras, materiales u operaciones bien conocidas para evitar oscurecer aspectos de la invención.

Los diagramas de flujo esquemáticos incluidos en este documento se muestran en general como diagramas de flujos lógicos. Como tal, el orden representado y las etapas etiquetadas son indicativas de una realización del presente método. Otras etapas y métodos se pueden concebir como equivalentes en función, lógica, o efecto a una o más etapas, o porciones de las mismas, del método ilustrado. Adicionalmente, el formato y los símbolos empleados se proporcionan para explicar las etapas lógicas del método y se entiende que no limitan el ámbito del método. Aunque se pueden emplear diversos tipos de flechas y tipos de líneas en los diagramas de flujo, se entiende que no limitan el ámbito del método correspondiente. En efecto, algunas flechas u otras conexiones se pueden usar para indicar solo el flujo lógico del método. Por ejemplo, una flecha puede indicar un periodo de espera o monitorización de una duración no específica entre etapas numeradas del método representado. Adicionalmente, el orden en el que ocurre un método particular se puede adherir o no de forma estricta al orden de las etapas correspondientes mostradas.

Sistema de almacenamiento de estado sólido

La Figura 1A es un diagrama de bloques esquemático que ilustra una realización de un sistema 100 para la gestión de datos en un dispositivo de almacenamiento de estado lógico de acuerdo con la presente invención. El sistema 100 incluye un dispositivo de almacenamiento de estado lógico 102, un controlador del almacenamiento de estado sólido 104, una conducción de datos de escritura 106, una conducción de datos de lectura 108, un almacenamiento de estado sólido 110, un ordenador 112, un cliente 114, y una red de ordenadores 116, que se describen a continuación.

El sistema 100 incluye al menos un dispositivo de almacenamiento de estado sólido 102. En otra realización, el sistema 100 incluye dos o más dispositivos de almacenamiento de estado sólido 102. Cada uno de los dispositivos de almacenamiento de estado sólido 102 puede incluir un almacenamiento de estado sólido no volátil 110, tal como una memoria flash, una nano memoria de acceso aleatorio ("nano RAM o NRAM"), una RAM magneto - resistiva ("MRAM"), una RAM dinámica ("DRAM"), una RAM de cambio de fase ("PRAM"), etc. El dispositivo de almacenamiento de estado sólido 102 se describe con más detalle con respecto a las Figuras 2 y 3. El dispositivo de almacenamiento de estado sólido 102 se representa en un ordenador 112 conectado a un cliente 114 a través de una red de ordenadores 116. En una realización, el dispositivo de almacenamiento de estado sólido 102 es interno al ordenador 112 y está conectado usando un bus del sistema, tal como un bus exprés de interconexión de

componentes periféricos ("PCI-e"), un bus de Conexión de Tecnología Avanzada Serie ("ATA serie"), o similares. En otra realización, el dispositivo de almacenamiento de estado sólido 102 es externo al ordenador 112 y está conectado por una conexión del bus serie universal ("USB"), un bus del Instituto de Ingenieros Eléctricos y Electrónicos ("IEEE") 1394 ("FireWire") o similares. En otras realizaciones, el dispositivo de almacenamiento de estado sólido 102 está conectado al ordenador 112 usando un bus exprés de interconexión de componentes periféricos ("PCI") que usa una extensión del bus óptico o eléctrico externo o la solución de bus de funcionamiento en red tal como Infiniband o la Conmutación Avanzada del PCI Exprés ("PCIe-AS"), o similares.

En diversas realizaciones, el dispositivo de almacenamiento de estado sólido 102 puede ser en la forma de un módulo de memoria dual en línea ("DIMM"), una tarjeta secundaria, o un micro - módulo. En otra realización, el dispositivo de almacenamiento de estado sólido 102 es un elemento dentro de un módulo montado en armazón. En otra realización, el dispositivo de almacenamiento de estado sólido 102 está contenido dentro de un paquete que está integrado directamente sobre un ensamblaje de alto nivel (por ejemplo, una tarjeta base, ordenador portátil, procesador gráfico). En una realización, los componentes individuales que comprenden el dispositivo de almacenamiento de estado sólido 102 están integrados directamente sobre un ensamblaje de nivel superior sin empaquetamiento intermedio.

El dispositivo de almacenamiento de estado sólido 102 incluye uno o más controladores de almacenamiento de estado sólido 104, cada uno de ellos puede incluir una conducción de datos de escritura 106 y una conducción de datos de lectura 108 y cada uno incluye un almacenamiento de estado sólido 110, que se describen con más detalle más adelante con respecto a las Figuras 2 y 3.

El sistema 100 incluye uno o más ordenadores 112 conectados al dispositivo de almacenamiento de estado sólido 102. Un ordenador 112 puede ser un hospedador, un servidor, un controlador de almacenamiento de una red de área de almacenamiento ("SAN"), una estación de trabajo, un ordenador portátil, un ordenador de mano, un superordenador, un agrupamiento de ordenadores, un conmutador de red, un encaminador o dispositivo, una base de datos o dispositivo de almacenamiento, un sistema de adquisición de datos o captura de datos, un sistema de diagnóstico, un sistema de prueba, un robot, un dispositivo electrónico portátil, un dispositivo inalámbrico, o similares. En otra realización, un ordenador 112 puede ser un cliente y el dispositivo de almacenamiento de estado sólido 102 opera de forma autónoma para servir a las peticiones de datos enviadas desde el ordenador 112. En esta realización, el ordenador 112 y el dispositivo de almacenamiento de estado sólido 102 se puede conectar usando una red de ordenadores, un bus de sistema u otro medio de comunicación adecuado para la conexión entre un ordenador 112 y un dispositivo de almacenamiento de estado sólido autónomo 102.

En una realización, el sistema 100 incluye uno o más clientes 114 conectados a uno o más ordenadores 112 a través de una o más redes de ordenadores 116. Un cliente 114 puede ser un hospedador, un servidor, un controlador de almacenamiento de una SAN, una estación de trabajo, un ordenador personal, un ordenador portátil, un ordenador de mano, un superordenador, un agrupamiento de ordenadores, un conmutador de red, un encaminador o dispositivo, una base de datos o dispositivo de almacenamiento, un sistema de adquisición de datos o captura de datos, un sistema de diagnóstico, un sistema de prueba, un robot, un dispositivo electrónico portátil, un dispositivo inalámbrico, o similares. La red de ordenadores 116 puede incluir la Internet, una red de área ancha ("WAN"), una red de área metropolitana ("MAN"), una red de área local ("LAN"), un sistema de token ring, una red inalámbrica, una red de canal de fibra, una SAN, un almacenamiento conectado en red ("NAS"), ESCON o similares o cualquier combinación de las redes. La red de ordenadores 116 puede incluir también una red de la familia 802 del IEEE de las tecnologías de red, como la Ethernet, token ring, WiFi, WiMax, y similares.

La red de ordenadores 116 puede incluir servidores, conmutadores, enrutadores, cableados, radios, y otro equipo usados para facilitar la conexión en red de ordenadores 112 y clientes 114. En una realización, el sistema 100 incluye múltiples ordenadores 112 que comunican como pares sobre una red de ordenadores 116. En otra realización, el sistema 100 incluye múltiples dispositivos de almacenamiento de estado sólido 102 que comunican como pares sobre una red de ordenadores 116. Un experto en la materia reconocerá otras redes de ordenadores 116 que comprenden una o más redes de ordenadores 116 y equipo relacionado con una conexión única o redundante entre uno o más clientes 114 u otro ordenador con uno o más dispositivos de almacenamiento de estado sólido 102 o uno o más dispositivos de almacenamiento de estado sólido 102 conectado a uno o más ordenadores 112. En una realización, el sistema 100 incluye dos o más dispositivos de almacenamiento de estado sólido 102 conectados a través de la red de ordenadores 116 a un cliente 114 sin un ordenador 112.

Controlador de almacenamiento - objetos gestionados

La Figura 1B es un diagrama de bloques esquemático que ilustra una realización de un sistema 101 para la gestión de objetos en un dispositivo de almacenamiento de acuerdo con la presente invención. El sistema 101 incluye uno o más dispositivos de almacenamiento 150, cada uno con un controlador de almacenamiento 152 y uno o más dispositivos de almacenamiento de datos 154, y uno o más dispositivos solicitantes 155. Los dispositivos de almacenamiento 152 están conectados juntos en red y acoplados a uno o más dispositivos solicitantes 155. El dispositivo solicitante 155 envía peticiones de objetos a un dispositivo de almacenamiento 150a. Una petición de objetos puede ser una petición para crear un objeto, una petición para escribir datos en un objeto, una petición para

leer datos desde un objeto, una petición para borrar un objeto, una petición para un punto de control de un objeto, una petición para copiar un objeto y similares. Un experto en la materia reconocerá otras peticiones de objetos.

En una realización, el controlador de almacenamiento 152 y el dispositivo de almacenamiento de datos 154 son dispositivos separados. En otra realización, el controlador de almacenamiento 152 y el dispositivo de almacenamiento de datos 154 están integrados dentro de un dispositivo de almacenamiento 150. En otra realización, un dispositivo de datos 154 es un almacenamiento de estado sólido 110 y el controlador de almacenamiento 152 es un controlador del dispositivo de almacenamiento de estado sólido 202. En otras realizaciones, el dispositivo de almacenamiento de datos 154 puede ser una unidad de disco duro, una unidad óptica, un almacenamiento de cinta o similares. En otra realización, un dispositivo de almacenamiento 150 puede incluir dos o más dispositivos de almacenamiento de datos 154 de diferentes tipos.

En una realización, el dispositivo de almacenamiento de datos 154 es un almacenamiento de estado sólido 110 y está dispuesto como una red de elementos de almacenamiento de estado sólido 216, 218, 220. En otra realización, el almacenamiento de estado sólido 110 está dispuesto en dos o más bancos 214a - n. El almacenamiento de estado sólido 110 se describe con más detalle más adelante con respecto a la Figura 2B.

Los dispositivos de almacenamiento 150a - n pueden estar juntos conectados en red y actuar como un dispositivo de almacenamiento distribuido. El dispositivo de almacenamiento 150a acoplado al dispositivo solicitante 155 controla las peticiones de objetos al dispositivo de almacenamiento distribuido. En una realización, los dispositivos de almacenamiento 150 y los controladores de almacenamiento asociados 152 gestionan objetos y aparecen a los dispositivos solicitantes 155 como un sistema de ficheros de objetos distribuidos. En este contexto, un sistema de ficheros de objetos en paralelo es un ejemplo de un tipo de sistema de ficheros de objetos distribuidos. En otra realización, los dispositivos de almacenamiento 150 y los controladores de almacenamiento asociados 152 gestionan los objetos y aparecen al dispositivo solicitante 155 como servidores de ficheros de objetos distribuidos. En este contexto, un servidor de ficheros de objetos en paralelo es un ejemplo de un tipo de servidor de ficheros de objetos distribuidos. En estas y otras realizaciones el dispositivo solicitante 155 puede gestionar exclusivamente objetos o participar en la gestión de los objetos en conjunción con los dispositivos de almacenamiento 150; esto usualmente no limita la capacidad de los dispositivos de almacenamiento 150 para gestionar completamente objetos para otros clientes 114. En el caso degenerado, cada uno de los dispositivos de almacenamiento distribuido, el sistema de ficheros de objetos distribuidos y el servidor de ficheros de objetos distribuidos pueden operar independientemente como un único dispositivo. Los dispositivos de almacenamiento conectados en red 150a - n pueden operar como dispositivos de almacenamiento distribuidos, sistemas de ficheros de objetos distribuidos, servidores de ficheros de objetos distribuidos, y cualquier combinación de los mismos que tienen imágenes de una o más de estas capacidades configuradas para uno o más dispositivos solicitantes 155. Por ejemplo, los dispositivos de almacenamiento 150 se pueden configurar para operar como dispositivos de almacenamiento distribuidos para un primer dispositivo solicitante 155a, mientras que operan como dispositivos de almacenamiento distribuido y sistemas de ficheros de objetos distribuidos para los dispositivos solicitantes 155b. Donde el sistema 101 incluye un dispositivo de almacenamiento 150a, el controlador de almacenamiento 152a del dispositivo de almacenamiento 150a, gestiona objetos que pueden aparecer a los dispositivos solicitantes 155 como un sistema de ficheros de objetos o un servidor de ficheros de objetos.

En una realización donde los dispositivos de almacenamiento 150 están juntos conectados en red como un dispositivo de almacenamiento distribuido, los dispositivos de almacenamiento 150 sirven como una red redundante de unidades independientes ("RAID") gestionada por uno o más controladores del almacenamiento distribuido 152. Por ejemplo, una petición de escritura de un segmento de datos de un objeto da como resultado en el segmento de datos que se desmonte a través de los dispositivos de almacenamiento de datos 154a - n con una banda de paridad, que depende del nivel de RAID. Un beneficio de tal disposición es que tal sistema de gestión de objetos puede continuar estando disponible cuando el dispositivo de almacenamiento único 150 tiene un fallo, ya sea del controlador de almacenamiento 152, el dispositivo de almacenamiento de datos 154, u otros componentes del dispositivo de almacenamiento 150.

Cuando se usan redes redundantes para interconectar los dispositivos de almacenamiento 150 y los dispositivos solicitantes 155, el sistema de gestión de objetos puede continuar estando disponible en presencia de fallos de red siempre que una de las redes siga estando operativa. Un sistema 101 con un dispositivo de almacenamiento único 150a también puede incluir múltiples dispositivos de almacenamiento de datos 154a y el controlador de almacenamiento 152a del dispositivo de almacenamiento 150a puede actuar como un controlador de RAID y desmontar los segmentos de datos a través de los dispositivos de almacenamiento de datos 154a del dispositivo de almacenamiento 150a y puede incluir una banda de paridad, dependiendo del nivel de RAID.

En una realización, donde el uno o más dispositivos de almacenamiento 150a - n son dispositivos de almacenamiento de estado sólido 102 con un controlador del dispositivo de almacenamiento de estado sólido 202 y el almacenamiento de estado sólido 110, el dispositivo de almacenamiento de estado sólido 102 puede estar configurado como una configuración DIMM, una tarjeta secundaria, un micro - módulo, etc. y residir en un ordenador 112. El ordenador 112 puede ser un servidor o dispositivo similar con dispositivos de almacenamiento de estado sólido 102 conectados en red y actuando como controladores de RAID distribuidos. Ventajosamente, los dispositivos

de almacenamiento 102 pueden estar conectados usando PCI-e, PCIe-AS, Infiniband u otro bus de altas prestaciones, bus conmutado, bus conectado en red, o una red y puede proporcionar un sistema de almacenamiento de RAID muy compacto, de altas prestaciones con un controlador único o controladores de almacenamiento de estado sólido único o distribuido 202 que desmontan de forma autónoma un segmento de datos a través del almacenamiento de estado sólido 110a - n.

En una realización, la misma red usada por el dispositivo solicitante 155 para comunicar con los dispositivos de almacenamiento 150 se puede usar por el dispositivo de almacenamiento par 150a para comunicar con los dispositivos de almacenamiento pares 150b - n para lograr la funcionalidad de RAID. En otra realización, se puede usar una red separada entre los dispositivos de almacenamiento 150 para el fin de estructuración de RAID. En otra realización, los dispositivos solicitantes 155 pueden participar en el proceso de estructuración de RAID enviando peticiones redundantes a los dispositivos de almacenamiento 150. Por ejemplo, el dispositivo solicitante 155 puede enviar una primera petición de escribir un objeto a un primer dispositivo de almacenamiento 150a y una segunda petición de escribir un objeto con el mismo segmento de datos a un segundo dispositivo de almacenamiento 150b para conseguir una simetría simple.

Con la capacidad para el manejo de objetos dentro de los dispositivos de almacenamiento 102, los controladores de almacenamiento 152 tienen únicamente la capacidad de almacenar un segmento de datos u objeto usando un nivel de RAID mientras que otro segmento de datos u objeto se almacena usando un nivel de RAID diferente o sin el desmontaje de RAID. Estos agrupamientos de RAID múltiples se pueden asociar con múltiples particiones dentro de los dispositivos de almacenamiento 150. RAID 0, RAID 1, RAID 5, RAID 6 y tipos compuestos de RAID 10, 50, 60, se pueden soportar simultáneamente a través de una diversidad de grupos de RAID que comprenden dispositivos de almacenamiento de datos 154a - n. Un experto en la materia reconocerá otros tipos y configuraciones de RAID y configuraciones que se pueden soportar simultáneamente.

También, debido a que los controladores de almacenamiento 152 operan de forma autónoma como controladores de RAID, los controladores de RAID pueden realizar estructuraciones de RAID progresivas y pueden transformar objetos o porciones de objetos desmontados a través de los dispositivos de almacenamiento de datos 154 con un nivel de RAID a otro nivel de RAID sin que se afecte al dispositivo solicitante 155, que participa o incluso que detecta el cambio en los niveles de RAID. En la realización preferida, el progreso de la configuración de RAID desde un nivel a otro nivel se puede lograr de forma autónoma sobre un objeto o incluso sobre la base de un paquete e iniciarse por un módulo de control de RAID distribuido que opera en uno de los dispositivos de almacenamiento 150 o los controladores de almacenamiento 152. Usualmente, la progresión de RAID será desde un funcionamiento superior y una configuración de almacenamiento de más baja eficiencia tal como una RAID a un funcionamiento inferior y una configuración de almacenamiento de más alta eficiencia tal como RAID 5 donde la transformación se inicia dinámicamente en base a la frecuencia de acceso. Pero se puede ver que el progreso de la configuración desde RAID 5 a RAID 1 también es posible. Otros procesos para iniciar la progresión de RAID se puede configurar o solicitar desde clientes o agentes externos tal como una petición del servidor de gestión del sistema de almacenamiento. Un experto en la materia reconocerá otras características y beneficios del dispositivo de almacenamiento 102 con un controlador de almacenamiento 152 que gestiona objetos de forma autónoma.

Dispositivo de almacenamiento de estado sólido con una san en servidor

La Figura 1C es un diagrama de bloques esquemático que ilustra una realización de un sistema 103 para una red de área de almacenamiento ("SAN") en servidor de acuerdo con la presente invención. El sistema 103 incluye un ordenador 112 configurado usualmente como un servidor ("servidor 112"). Cada servidor 112 incluye uno o más dispositivos de almacenamiento 150 donde el servidor 112 y los dispositivos de almacenamiento 150 están cada uno conectado a una interfaz de red compartida 156. Cada dispositivo de almacenamiento 150 incluye un controlador de almacenamiento 152 y el dispositivo de almacenamiento de datos correspondiente 154. El sistema 103 incluye los clientes 114, 114a, 114b que son bien internos o externos a los servidores 112. Los clientes 114, 114a, 114b pueden comunicar con cada servidor 112 y cada dispositivo de almacenamiento 150 a través de uno o más redes de ordenadores 116, que son sustancialmente similares a las descritas anteriormente.

El dispositivo de almacenamiento 150 incluye un módulo DAS 158, un módulo NAS 160, un módulo de comunicaciones de almacenamiento 162, un módulo de SAN en servidor 164, un módulo de interfaz común 166, un módulo proxy 170, un módulo de bus virtual 172, un módulo RAID del extremo frontal 174, y un módulo RAID del extremo posterior 176, que se describen más adelante. Aunque los módulos 158 - 176 se muestran en un dispositivo de almacenamiento 150, todos o una porción de cada módulo 158 - 176 puede estar en el dispositivo de almacenamiento 150, el servidor 112, el controlador de almacenamiento 152 u otra localización.

Un servidor 112, usado en conjunción con una SAN en servidor, es un ordenador que funciona como un servidor. El servidor 112 incluye al menos una función de servidor, tal como una función de servidor de ficheros, pero también puede incluir otras funciones de servidor. Los servidores 112 pueden ser parte de una granja de servidores y pueden dar servicio a otros clientes 114. En otras realizaciones, el servidor 112 también puede ser un ordenador personal, una estación de trabajo u otro ordenador que alberga dispositivos de almacenamiento 150. Un servidor 112 puede acceder a uno o más dispositivos de almacenamiento 150 en el servidor 112 como un almacenamiento de conexión

directa ("DAS"), un almacenamiento de conexión SAN o un almacenamiento de conexión en red ("NAS"). Los controladores de almacenamiento 152 que participan en una SAN en servidor o NAS pueden ser internos o externos al servidor 12.

En una realización, el aparato de SAN en servidor incluye un módulo DAS 158 que configura al menos una porción de al menos un dispositivo de almacenamiento de datos 154 controlado por un controlador de almacenamiento 152 en un servidor 112 como un dispositivo DAS conectado al servidor 112 para dar servicio a las peticiones de almacenamiento desde al menos un cliente 114 al servidor 112. En una realización, un primer dispositivo de almacenamiento de datos 154a se configura como un DAS para el primer servidor 112a mientras que también se configura como un dispositivo de almacenamiento SAN en servidor para el primer servidor 112a. En otra realización, el primer dispositivo de almacenamiento de datos 154a se divide de modo que una partición es un DAS y la otra es una SAN en servidor. En otra realización, al menos una porción del espacio de almacenamiento dentro del primer dispositivo de almacenamiento de datos 154a se configura como un DAS para el primer servidor 112a y la misma porción de espacio de almacenamiento sobre el primer dispositivo de almacenamiento de datos 154a se configura como una SAN en servidor para el primer servidor 112a.

En otra realización, el aparato de SAN en servidor incluye un módulo de NAS 160 que configura un controlador de almacenamiento 152 como un dispositivo NAS para al menos un cliente 114 y sirve las peticiones de ficheros desde el cliente 114. El controlador de almacenamiento 152 también se puede configurar como un dispositivo SAN en servidor para el primer servidor 112a. Los dispositivos de almacenamiento 150 pueden conectar directamente con la red de ordenadores 116 a través de la interfaz de red compartida 156 independiente del servidor 112 en el que reside el dispositivo de almacenamiento 150.

En una forma elemental, un aparato en una SAN en servidor incluye un primer controlador de almacenamiento 152a dentro un primer servidor 112a donde el primer controlador de almacenamiento 152a controla al menos un dispositivo de almacenamiento 154a. El primer servidor 112a incluye una interfaz de red 156 compartida por el primer servidor 112a y el primer controlador de almacenamiento 152a. El aparato de SAN en servidor incluye un módulo de comunicaciones de almacenamiento 162 que facilita las comunicaciones entre el primer controlador de almacenamiento 152a y al menos un dispositivo externo al primer servidor 112a de modo que la comunicación entre el primer controlador de almacenamiento 152a y el dispositivo externo es independiente del primer servidor 112a. El módulo de comunicaciones de almacenamiento 162 puede permitir al primer controlador de almacenamiento 152a acceder independientemente a la interfaz de red 156a para la comunicación externa. En una realización, el módulo de comunicaciones de almacenamiento 162 accede a un conmutador en la interfaz de red 156a para el tráfico de red directo entre el primer controlador de almacenamiento 152a y dispositivos externos.

El aparato de SAN en servidor también incluye un módulo de SAN en servidor 164 que sirve a una petición de almacenamiento usando uno o ambos de un protocolo de red y un protocolo de bus. El módulo de SAN en servidor 164 sirve a la petición de almacenamiento independiente desde el primer servidor 112a y la petición de servicio que se recibe desde un cliente interno o externo 114, 114a.

En una realización, el dispositivo externo al primer servidor 112a es un segundo controlador de almacenamiento 152b. El segundo controlador de almacenamiento 152b controla al menos un dispositivo de almacenamiento de datos 154b. El módulo de SAN en servidor 164 sirve la petición de almacenamiento usando la comunicación a través de la interfaz de red 156a y entre el primer y segundo controladores de almacenamiento 152a, 152b independientes del primer servidor 112a. El segundo controlador de almacenamiento 152b puede estar dentro de un segundo servidor 112b o dentro de algún otro dispositivo.

En otra realización, el dispositivo externo al primer servidor 112a es un cliente 114 y la petición de almacenamiento se origina con el cliente externo 114 donde el primer controlador de almacenamiento 152a está configurado como al menos parte de una SAN y el módulo de SAN en servidor 164 sirve la petición de almacenamiento a través de la interfaz de red 156a independiente del primer servidor 112a. El cliente externo 114 puede estar en el segundo servidor 112b o puede ser externo al segundo servidor 112b. En una realización, el módulo de SAN en servidor 164 puede servir las peticiones de almacenamiento desde el cliente externo 114 incluso cuando el primer servidor 112a no está disponible.

En otra realización, el cliente 114a que origina la petición de almacenamiento es interno al primer servidor 112a donde el primer controlador de almacenamiento 152a está configurado como al menos parte de una SAN y el módulo de SAN en servidor 164 sirve la petición de almacenamiento a través de una o más de la interfaz de red 156a y el bus de sistema.

Las configuraciones de SAN tradicionales permiten el acceso a un dispositivo de almacenamiento remoto del servidor 112b como si el dispositivo de almacenamiento residiese dentro del servidor 112 como un almacenamiento de conexión directa ("DAS") de modo que el dispositivo de almacenamiento aparece como un dispositivo de almacenamiento de bloques. Usualmente, un dispositivo de almacenamiento conectado a una SAN requiere un protocolo de SAN, tal como un canal de fibra, una interfaz de Internet de un pequeño sistema de ordenadores ("iSCSI"), HyperSCSI, Conectividad de Fibra ("FICON"), la Conexión de Tecnología Avanzada ("ATA") sobre

Ethernet, etc. La SAN en servidor incluye un controlador de almacenamiento 152 dentro de un servidor 112 mientras que aun permite la conexión de red entre el controlador de almacenamiento 152a y el controlador de almacenamiento remoto 152b o un cliente externo 114 usando un protocolo de red y/o un protocolo de bus.

Usualmente, los protocolos de SAN son una forma de protocolo de red y más protocolos de red que están emergiendo, tal como infiniband que permitiría a un controlador de almacenamiento 152a y los dispositivos de almacenamiento de datos asociados 154a, configurarse como una SAN y comunicar con un cliente externo 114 o un segundo controlador de almacenamiento 152b. En otro ejemplo, un primer controlador de almacenamiento 152a puede comunicar con un cliente externo 114 o un segundo controlador de almacenamiento 152b usando Ethernet.

Un controlador de almacenamiento 152 puede comunicar sobre un bus con controladores de almacenamiento interno 152 o clientes 114a. Por ejemplo, un controlador de almacenamiento 152 puede comunicar sobre un bus usando PCI-e que puede soportar la Virtualización de Entradas / Salidas sobre PCI Exprés ("PCIe-IOV"). Otros protocolos de bus emergentes permiten a un bus de sistema extenderse fuera del ordenador o servidor 112 y permitirían a un controlador de almacenamiento 152a configurarse como una SAN. Uno de tales protocolos de bus es el PCIe-AS. La presente invención no está limitada simplemente a los protocolos de SAN, sino que también se puede aprovechar de los protocolos de red y de bus emergentes para servir las peticiones de almacenamiento. Un dispositivo externo, bien en la forma de un cliente 114 o de controlador de almacenamiento externo 152b, puede comunicar sobre un bus de sistema extendido o una red de ordenadores 116. Una petición de almacenamiento, como se usa en este documento, incluye peticiones para escribir datos, leer datos, borrar datos, consultar datos, etc. y puede incluir datos de objetos, metadatos, y peticiones de gestión así como peticiones de datos de bloques.

Un servidor tradicional 112 usualmente tiene una raíz compleja que controla el acceso a los dispositivos dentro del servidor 112. Usualmente, esta raíz compleja del servidor 112 pertenece a la interfaz de red 156 de modo que cualquier comunicación a través de la interfaz de red 156 está controlada por el servidor 112. Sin embargo, en la realización preferida del aparato de SAN en servidor, el controlador de almacenamiento 152 es capaz de acceder a la interfaz de red 156 independientemente de modo que los clientes 114 pueden comunicar directamente con uno o más de los controladores de almacenamiento 152a en el primer servidor 112a formando una SAN o de modo que uno o más de los primeros controladores de almacenamiento 152a pueden estar conectados en red con un segundo controlador de almacenamiento 152b u otros controladores de almacenamiento remotos 152 para formar una SAN. En la realización preferida, los dispositivos remotos del primer servidor 112a pueden acceder al primer servidor 112a o el primer controlador de almacenamiento 152a a través de una dirección de red única, compartida. En una realización, el aparato de SAN en servidor incluye un módulo de interfaz común 166 que configura la interfaz de red 156, el controlador de almacenamiento 152 y el servidor 112 de modo que el servidor 112 y el controlador de almacenamiento 152 son accesibles usando una dirección de red compartida.

En otras realizaciones, el servidor 112 incluye dos o más interfaces de red 156. Por ejemplo, el servidor 112 puede comunicar sobre una interfaz de red 156 mientras que el dispositivo de almacenamiento 150 puede comunicar sobre otra interfaz. En otro ejemplo, el servidor 112 incluye múltiples dispositivos de almacenamiento 150, cada uno de ellos con una interfaz de red 156. Un experto en la materia reconocerá otras configuraciones de un servidor 112 con uno o más dispositivos de almacenamiento 150 y una o más interfaces de red 156 donde uno o más de los dispositivos de almacenamiento 150 accede a la interfaz de red 156 independiente del servidor 112. Un experto en la materia también reconocerá cómo se pueden extender estas diversas configuraciones para soportar la redundancia de red y mejorar la disponibilidad.

Ventajosamente, el aparato de SAN en servidor elimina mucha de la complejidad y el coste de una SAN tradicional. Por ejemplo, una SAN típica requiere servidores 112 con controladores de almacenamiento externo 152 y dispositivos de almacenamiento de datos asociados 154. Esto ocupa un espacio adicional en un armazón y requiere cableado, conmutadores, etc. El cableado, la conmutación y otro código de control distinto requerido para configurar un SAN tradicional ocupan espacio, degradan el ancho de banda y son caros. El aparato de SAN en servidor permite a los controladores de almacenamiento 152 y el almacenamiento asociado 154 fijar en un servidor 112 un factor de forma, reduciendo de este modo el espacio requerido y con un coste inferior. Una SAN en servidor también permite la conexión usando una comunicación relativamente rápida sobre buses de datos de alta velocidad internos y externos.

En una realización, el dispositivo de almacenamiento 150 es un dispositivo de almacenamiento de estado sólido 102, el controlador de almacenamiento 152 es un controlador de almacenamiento de estado sólido 104, y el dispositivo de almacenamiento de datos 154 es un almacenamiento de estado sólido 110. Esta realización es ventajosa debido a la velocidad del dispositivo de almacenamiento de estado sólido 102 como se describe en este documento. Además, el dispositivo de almacenamiento de estado sólido 102 se puede configurar en un DIMM que puede fijarse convenientemente en un servidor 112 y requiere una pequeña cantidad de espacio.

El uno o más clientes internos 114a, b en el servidor 112 también pueden conectar con la red de ordenadores 116 a través de la interfaz de red 156 del servidor 112 y la conexión del cliente está usualmente controlada por el servidor 112. Esto tiene varias ventajas. Los clientes 114a pueden acceder lógicamente y remotamente a los dispositivos de

almacenamiento 150 directamente y pueden iniciar una transferencia de datos de acceso directo a memoria local o remoto ("DMA", "RDMA") entre la memoria de un cliente 114a y un dispositivo de almacenamiento 150.

En otra realización, los clientes 114, 114a dentro o fuera del servidor 112 pueden actuar como servidores de ficheros para los clientes 114 a través de una o más redes 116 mientras que usan dispositivos de almacenamiento conectados localmente 150 como dispositivos DAS, dispositivos de almacenamiento conectados en red 150, dispositivos de almacenamiento de estado sólido conectados en red 102 que participan como parte de unas SAN en servidor, unas SAN externas, y SAN híbridas. Un dispositivo de almacenamiento 150 puede participar en un DAS, una SAN en servidor, SAN, NAS, etc., simultáneamente y en cualquier combinación. Adicionalmente, cada uno de los dispositivos de almacenamiento 150 se pueden dividir de tal modo que una primera partición hace al dispositivo de almacenamiento 150 disponible como un DAS, una segunda partición hace al dispositivo de almacenamiento 150 disponible como un elemento en una SAN en servidor, una tercera partición hace al dispositivo de almacenamiento 150 disponible como un NAS, una cuarta partición hace al dispositivo de almacenamiento 150 disponible como un elemento en una SAN, etc. De forma similar, el dispositivo de almacenamiento 150 se puede dividir de forma consistente con la seguridad y los requisitos de control de acceso. Un experto en la materia reconocerá que cualquier número de combinaciones y permutaciones de los dispositivos de almacenamiento, dispositivos de almacenamiento virtual, redes de almacenamiento, redes de almacenamiento virtual, almacenamiento privado, almacenamiento compartido, sistemas de ficheros en paralelo, sistemas de ficheros de objetos en paralelo, dispositivos de almacenamiento de bloques, dispositivos de almacenamiento de objetos, dispositivos de almacenamiento, dispositivos de red y similares se pueden construir y soportar.

Además, conectando directamente a la red de ordenadores 116, los dispositivos de almacenamiento 150 pueden comunicar con cada uno de los otros y pueden actuar como una SAN en servidor. Los clientes 114a en los servidores 112 y los clientes 114 conectados a través de la red de ordenadores 116 pueden acceder a los dispositivos de almacenamiento 150 como una SAN. Moviendo los dispositivos de almacenamiento 150 dentro de los servidores 112 y teniendo la capacidad de configurar los dispositivos de almacenamiento 150 como una SAN, la combinación de servidor 112 / dispositivo de almacenamiento 150 elimina la necesidad en las SAN convencionales de controladores de almacenamiento dedicados, redes de canal de fibra, y otros equipos. El sistema de SAN en servidor 103 tiene la ventaja de posibilitar al dispositivo de almacenamiento 150 la compartición de los recursos comunes tales como potencia, enfriamiento, gestión y espacio físico con el cliente 114 y el ordenador 112. Por ejemplo, los dispositivos de almacenamiento 150 pueden rellenar ranuras vacías de los servidores 112 y proporcionar todas las capacidades de funcionamiento, fiabilidad y disponibilidad de una SAN o NAS. Un experto en la materia reconocerá que otras características y beneficios de un sistema de SAN en servidor 103.

En otra configuración, se colocan múltiples dispositivos de almacenamiento de SAN en servidor 150a dentro de una infraestructura de servidor único 112a. En una realización, el servidor único 112a está comprendido de uno o más clientes de servidores de módulo internos 114a interconectados usando una IOV de PCI exprés sin una interfaz de red externa 156, cliente externo 114, 114b o el dispositivo de almacenamiento externo 150b.

Además, el dispositivo de almacenamiento de SAN en servidor 150 puede comunicar a través de una o más redes de ordenadores 116 con dispositivos de almacenamiento pares 150 que están localizados en un ordenador 112 (para la Figura 1A), o se conectan directamente a la red de ordenadores 116 sin un ordenador 112 para formar una SAN híbrida que tiene todas las capacidades de ambas una SAN y una SAN en servidor. Esta flexibilidad tiene el beneficio de simplificar la extensibilidad y la migración entre una diversidad de implementaciones de red de almacenamiento de estado sólido posibles. Un experto en la técnica reconocerá otras combinaciones, configuraciones, implementaciones, y arquitecturas para la localización y la interconexión de controladores de estado sólido 104.

Donde la interfaz de red 156a se puede controlar por solo un agente que opera dentro del servidor 112a, un módulo de establecimiento de enlace 168 que opera dentro de ese agente puede establecer trayectorias de comunicación entre clientes internos 114a y dispositivos de almacenamiento 150a / primeros controladores de almacenamiento 152a a través de la interfaz de red 156a con los dispositivos de almacenamiento externos 150b y los clientes 114, 114b. En una realización preferida, una vez que se ha establecido la trayectoria de comunicación, los dispositivos de almacenamiento internos individuales 150a y los clientes internos 114a son capaces de establecer y gestionar sus propias colas de comandos y transferir tanto comandos como datos a través de la interfaz de red 156a a dispositivos de almacenamiento externos 150b y clientes 114, 114b en cualquier dirección, directamente y a través de un RDMA independiente del proxy o agente que controla la interfaz de red 156a. En una realización, el módulo de establecimiento del enlace 168 establece los enlaces de comunicaciones durante el proceso de inicialización, tal como en el arranque o inicialización del hardware.

En otra realización, un módulo proxy 170 dirige al menos una porción de los comandos usados en servir una petición de almacenamiento a través del primer servidor 112a mientras que al menos los datos, y posiblemente otros comandos, asociados con la petición de almacenamiento se comunican entre el primer controlador de almacenamiento 152a y el dispositivo de almacenamiento externo 150b independiente del primer servidor 112a. En otra realización, el módulo proxy 170 retransmite los comandos o los datos en nombre de los dispositivos de almacenamiento internos 150a y los clientes 114a.

En una realización, el primer servidor 112a incluye uno o más servidores dentro del primer servidor 112a e incluye un módulo de bus virtual 172 que permite al uno o más servidores en el primer servidor 112a acceder de forma independiente a uno o más controladores de almacenamiento 152a a través de buses virtuales separados. Los buses virtuales se pueden establecer usando un protocolo de bus avanzado tal como un PCIe-IOV. Las interfaces de red 156a que soportan la IOV pueden permitir al uno o más servidores y el uno o más controladores de almacenamiento 152a controlar de forma independiente la una o más interfaces de red 156a.

En diversas realizaciones el aparato de SAN en servidor permite a dos o más dispositivos de almacenamiento 150 estar configurados en una RAID. En una realización, el aparato de SAN en servidor incluye un módulo RAID del extremo frontal 174 que configura dos o más controladores de almacenamiento 152 como una RAID. Donde una petición de almacenamiento desde un cliente 114, 114a incluye una petición para almacenar datos, el módulo RAID del extremo frontal 174 sirve la petición de almacenamiento escribiendo los datos en la RAID de forma consistente con el nivel de RAID implementado. Un segundo controlador de almacenamiento 152 puede estar localizado bien en el primer servidor 112a o externo al primer servidor 112a. El módulo RAID del extremo frontal 174 permite el establecimiento de RAID de los controladores de almacenamiento 152 de modo que los controladores de almacenamiento 152 son visibles para el cliente 114, 114a que envía la petición de almacenamiento. Esto permite el desmontaje y la información de paridad a gestionar por el controlador de almacenamiento 152 designado como maestro o por el cliente 114, 114a.

En otra realización, el aparato de SAN en servidor incluye un módulo de RAID del extremo posterior 176 que configura dos o más dispositivos de almacenamiento de datos 154 controlados por un controlador de almacenamiento 152 como una RAID. Donde la petición de almacenamiento desde el cliente 114 comprende una petición de almacenar datos, el módulo RAID del extremo posterior 176 sirve la petición de almacenamiento escribiendo los datos en la RAID de forma consistente con el nivel de RAID implementado de modo que los dispositivos de almacenamiento 154 configurados como una RAID se acceden por el cliente 114, 114a como un dispositivo de almacenamiento de datos único 154 controlado por el primer controlador de almacenamiento 152. Esta implementación de RAID permite la estructuración en RAID de dispositivos de almacenamiento de datos 154 controlados por un controlador de almacenamiento 152 en un modo en el que la estructuración en RAID es transparente para cualquier cliente 114, 114a que accede a los dispositivos de almacenamiento de datos 154. En otra realización, tanto la RAID del extremo frontal como la RAID del extremo posterior se implementan para tener una RAID multinivel. Un experto en la materia reconocerá otros modos de estructurar en RAID los dispositivos de almacenamiento 154 consistentes con el controlador de almacenamiento de estado sólido 104 y al almacenamiento de estado sólido asociado 110 descrito en este documento.

Aparato para el almacenamiento de objetos gestionados por el controlador

La Figura 2A es un diagrama de bloques que ilustra una realización de un aparato 200 para la gestión de objetos en un dispositivo de almacenamiento de acuerdo con la presente invención. El aparato 200 incluye un controlador de almacenamiento 152 con un módulo receptor de peticiones de objetos 260, un módulo de análisis 262, un módulo de ejecución de comandos 264, un módulo de índices de objetos 266, un módulo de poner en cola las peticiones de objetos 268, un empaquetador 302 con un módulo de mensajes 270, y un módulo de reconstrucción de índices de objetos 272, que se describen a continuación.

El controlador de almacenamiento 152 es sustancialmente similar al controlador de almacenamiento 152 descrito en relación con el sistema 101 de la Figura 1B y puede ser el controlador de dispositivos de almacenamiento de estado sólido 202 descrito en relación con la Figura 2. El aparato 200 incluye un módulo receptor de peticiones de objetos 260 que recibe una petición de objetos desde uno o más dispositivos solicitantes 155. Por ejemplo, para una petición de almacenar datos de un objeto, el controlador de almacenamiento 152 almacena el segmento de datos como un paquete de datos en un dispositivo de almacenamiento de datos 154 acoplado al controlador de almacenamiento 152. La petición de objetos está usualmente dirigida a un segmento de datos almacenado o que se va a almacenar en uno o más paquetes de datos de objetos para un objeto gestionado por el controlador de almacenamiento 152. La petición de objetos puede solicitar que el controlador de almacenamiento 152 cree un objeto a rellenar más tarde con datos mediante una petición de objeto posterior que pueden usar una transferencia de acceso directo a memoria local o remota ("DMA", "RDMA").

En una realización, la petición de objetos es una petición de escritura para escribir todo o parte de un objeto para un objeto creado anteriormente. En un ejemplo, la petición de escritura es para un segmento de datos de un objeto. Los otros segmentos de datos del objeto se pueden escribir en el dispositivo de almacenamiento 150 o en los otros dispositivos de almacenamiento. En otro ejemplo, la petición de escritura es para un objeto completo. En otro ejemplo, la petición de objetos es para leer datos desde un segmento de datos gestionado por el controlador de almacenamiento 152. En otra realización más, la petición de objetos es una petición de borrado, para borrar un segmento de datos u objeto.

Ventajosamente, el controlador de almacenamiento 152 puede aceptar peticiones de escritura que hacen más que escribir un nuevo objeto o añadir datos a un objeto existente. Por ejemplo, una petición de escritura recibida por el

módulo receptor de peticiones de objetos 260 puede incluir una petición de añadir datos por delante de los datos almacenados por el controlador de almacenamiento 152, insertar datos dentro de los datos almacenados, o reemplazar un segmento de datos. El índice de objetos mantenido por el controlador de almacenamiento 152 proporciona la flexibilidad requerida para estas operaciones de escritura complejas que no está disponible en otros controladores de almacenamiento, pero actualmente está disponible solo fuera de los controladores de almacenamiento en los sistemas de ficheros de servidores y otros ordenadores.

El aparato 200 incluye un módulo de análisis 262 que analiza la petición de objetos dentro de uno o más comandos. Usualmente, el módulo de análisis 262 analiza la petición de objetos dentro de uno o más memorias intermedias. Por ejemplo, se pueden analizar uno o más comandos en la petición de objetos dentro de una memoria intermedia de comandos. Usualmente el módulo de análisis 262 prepara una petición de objeto de modo que la información en la petición de objeto se puede entender y ejecutar por el controlador de almacenamiento 152. Un experto en la materia reconocerá otras funciones de un módulo de análisis 262 que analiza una petición de objetos dentro de uno o más comandos.

El aparato 200 incluye un módulo de ejecución de comandos 264 que ejecuta los comandos analizados de la petición de objetos. En una realización, el módulo de ejecución de comandos 264 ejecuta un comando. En otra realización, el módulo de ejecución de comandos 264 ejecuta múltiples comandos. Usualmente, el módulo de ejecución de comandos 264 interpreta un comando analizado desde la petición de objetos, tal como un comando de escritura, y crea a continuación, colas y ejecuta subcomandos. Por ejemplo, un comando de escritura analizado a partir de una petición de objetos puede dirigir el controlador de almacenamiento 152 para almacenar múltiples segmentos de datos. La petición de objetos puede incluir también los atributos requeridos tales como el cifrado, la compresión, etc. El módulo de ejecución de comandos 264 puede dirigir el controlador de almacenamiento 152 para comprimir los segmentos de datos, cifrar los segmentos de datos, crear uno más paquetes de datos y las cabeceras asociadas para cada paquete de datos, cifrar los paquetes de datos con una clave de cifrado de medios, añadir el código de corrección de errores, y almacenar los paquetes de datos en una localización específica. Almacenar los paquetes de datos en una localización específica y otros subcomandos también se pueden descomponer en otros subcomandos de nivel inferior. Un experto en la materia reconocerá otros modos en los que el módulo de ejecución de comandos 264 puede ejecutar uno o más comandos analizados a partir de una petición de objetos.

El aparato 200 incluye un módulo de índices de objetos 266 que crea una entrada de objetos en un índice de objetos en respuesta al controlador de almacenamiento 152 creando un objeto o almacenando el segmento de datos del objeto. Usualmente, el controlador de almacenamiento 152 crea un paquete de datos a partir del segmento de datos y la localización de donde se almacena el paquete de datos se asigna en el momento que se almacena el segmento de datos. Los metadatos de objetos recibidos con un segmento de datos o una parte de la petición de objeto se pueden almacenar en un modo similar.

El módulo de índices de objetos 266 crea una entrada de objetos dentro de un índice de objetos en el momento que se almacena el paquete de datos y se asigna la dirección física del paquete de datos. La entrada de objetos incluye un mapeo entre un identificador lógico del objeto y una o más direcciones físicas correspondientes a dónde almacenó el controlador de almacenamiento 152 uno o más paquetes de datos y cualesquiera paquetes de metadatos de objetos. En otra realización, la entrada en el índice de objetos se crea antes de que se almacenen los paquetes de datos del objeto. Por ejemplo, si el controlador de almacenamiento 152 determina una dirección física de dónde están los paquetes de datos que se van a almacenar más pronto, el módulo del índice de objetos 266 puede crear una entrada en el índice de objetos más temprano.

Usualmente, cuando una petición de objetos o grupo de peticiones de objetos da como resultado que se modifica un objeto o segmento de datos, posiblemente durante una operación de leer - modificar - escribir, el módulo de índices de objetos 266 actualiza una entrada en el índice de objetos correspondiente al objeto modificado. En una realización, el índice de objetos crea un nuevo objeto y una nueva entrada en el índice de objetos para el objeto modificado. Usualmente, cuando se modifica solo una porción de un objeto, el objeto incluye los paquetes de datos modificados y algunos paquetes de datos que permanecen sin cambiar. En ese caso, la nueva entrada incluye un mapeo a los paquetes de datos sin cambiar y donde estaban escritos originalmente y a los objetos modificados escritos en una nueva localización.

En otra realización, cuando el módulo receptor de peticiones de objetos 260 recibe una petición de objetos que incluye un comando que borra un bloque de datos u otros elementos de objetos, el controlador de almacenamiento 152 puede almacenar al menos un paquete tal como un paquete de borrado que incluye información que incluye una referencia al objeto, la relación con el objeto, y el tamaño del bloque de datos borrado. Adicionalmente, puede indicar además que los elementos del objeto borrado se rellenen con ceros. De este modo, la petición de borrar el objeto se puede usar para emular la memoria o almacenamiento real que se borra y realmente tiene una porción de la memoria / almacenamiento apropiado realmente almacenado con ceros en las células de la memoria / almacenamiento.

Ventajosamente, la creación de un índice de objetos con las entradas que indican el mapeo entre los segmentos de datos y los metadatos de un objeto permite al controlador de almacenamiento 152 manejarse de forma autónoma y

gestionar los objetos. Esta capacidad permite una gran cantidad de flexibilidad para almacenar datos en el dispositivo de almacenamiento 150. Una vez que se crea la entrada de índices para el objeto, se puede dar servicio a peticiones de objetos posteriores con respecto al objeto de forma eficiente por el controlador de almacenamiento 152.

En una realización, el controlador de almacenamiento 152 incluye un módulo de puesta en cola de las peticiones de objetos 268 que pone en cola una o más peticiones de objetos recibidas por el módulo receptor de peticiones de objetos 260 antes de analizarse por el módulo de análisis 262. El módulo de puesta en cola de las peticiones de objetos 268 permite flexibilidad entre cuándo se recibe la petición del objeto y cuándo se ejecuta.

En otra realización, el controlador de almacenamiento 152 incluye un empaquetador 302 que crea uno o más paquetes de datos a partir del uno o más segmentos de datos donde los paquetes de datos se dimensionan para el almacenamiento en el dispositivo de almacenamiento de datos 154. El empaquetador 302 se describe a continuación con más detalle con respecto a la Figura 3. El empaquetador 302 incluye, en una realización, un módulo de mensajes 270 que crea una cabecera para cada paquete. La cabecera incluye un identificador de paquete y una longitud de paquete. El identificador de paquete relaciona el paquete con el objeto para el que se formó el paquete.

En una realización, cada paquete incluye un identificador de paquete que está auto-contenido ya que el identificador de paquete contiene información adecuada para identificar el objeto y la relación dentro del objeto de los elementos del objeto contenidos dentro del paquete. Sin embargo, una realización más eficiente preferida es almacenar los paquetes en contenedores.

Un contenedor es una construcción de datos que facilita un almacenamiento más eficiente de paquetes y ayuda a establecer la relación entre un objeto y los paquetes de datos, los paquetes de metadatos y otros paquetes relacionados con el objeto que se almacenan dentro del contenedor. Obsérvese que el controlador de almacenamiento 152 usualmente trata los metadatos de objetos recibidos como parte de un objeto y los segmentos de datos en un modo similar. Usualmente "paquete" se puede referir a un paquete de datos que comprende datos, un paquete de metadatos que comprende metadatos, u otro paquete de otro tipo de paquete. Un objeto se puede almacenar en uno o más contenedores y un contenedor usualmente incluye paquetes para no más de un objeto único. Un objeto se puede distribuir entre múltiples contenedores. Usualmente un contenedor se almacena dentro de un bloque de borrado lógico único (división de almacenamiento) y usualmente nunca se divide entre bloques lógicos de borrado.

Un contenedor, en un ejemplo, se puede dividir entre dos o más páginas lógicas / virtuales. Un contenedor se identifica por una etiqueta de contenedor que asocia ese contenedor con un objeto. Un contenedor puede contener cero para muchos paquetes y los paquetes dentro de un contenedor son usualmente de un objeto. Un paquete puede ser de muchos tipos de elementos de objetos, incluyendo elementos de atributos de objetos, elementos de datos de objetos, elementos de índices de objetos y similares. Se pueden crear paquetes híbridos que incluyen más de un tipo de elemento de objeto. Cada uno de los paquetes puede contener de cero para muchos elementos del mismo tipo de elemento. Cada paquete dentro de un contenedor usualmente contiene un identificador único que identifica la relación con el objeto.

Cada paquete está asociado con un contenedor. En una realización preferida, los contenedores están limitados a un bloque de borrado de modo que al comienzo o cerca del comienzo de cada bloque de borrado se puede encontrar un paquete de contenedor. Esto ayuda a limitar las pérdidas de datos para un bloque de borrado con una cabecera de paquete corrompido. En esta realización, si el índice de objeto no está disponible y la cabecera del paquete dentro del bloque de borrado está corrompida, los contenidos desde la cabecera del paquete corrompido hasta el final del bloque de borrado se pueden perder porque posiblemente no hay ningún mecanismo fiable para determinar la localización de los siguientes paquetes. En otra realización, un enfoque más fiable es tener un contenedor limitado a una frontera de página. Esta realización, requiere más código de control para la cabecera. En otra realización, los contenedores pueden fluir a través de la página y las fronteras de bloque de borrado. Esto requiere menos código de control para la cabecera pero se puede perder una mayor porción de los datos si se corrompe una cabecera de paquete. Para estas diversas realizaciones se espera que se use algún tipo de RAID para asegurar adicionalmente la integridad de los datos.

En una realización, el aparato 200 incluye un módulo de reconstrucción de los índices de objetos 272 que reconstruye las entradas en el índice de objetos usando la información procedente de las cabeceras de paquetes almacenadas en el dispositivo de almacenamiento de datos 154. En una realización, el módulo de reconstrucción de índices de objetos 272 reconstruye las entradas del índice de objetos leyendo las cabeceras para determinar el objeto al que pertenece cada paquete y la información de secuencia para determinar a dónde pertenecen en el objeto los datos o metadatos. El módulo de reconstrucción de índices de objetos 272 usa la información de dirección física para cada paquete y la información del sello temporal o de secuencia para crear un mapeo entre las localizaciones físicas de los paquetes y el identificador de objeto y la secuencia de segmento de datos. La información del sello temporal o de secuencia se usa por el módulo de reconstrucción de índices de objetos 272

para repetir la secuencia de cambios realizados para el índice y por lo tanto reestablecer usualmente el estado más reciente.

En otra realización, el módulo de reconstrucción de índices de objetos 272 localiza los paquetes usando la información de la cabecera de los paquetes junto con la información de los paquetes de contenedor para identificar las localizaciones físicas de los paquetes, el identificador de objetos, y el número de secuencia de cada paquete para reconstruir las entradas en el índice de objetos. En una realización, los bloques de borrado están sellados en el tiempo o se les da un número de secuencia cuando se escriben los paquetes y el sello temporal o la información de secuencia de un bloque de borrado se usa junto con la información recogida desde las cabeceras de contenedor y las cabeceras de paquete para reconstruir el índice de objeto. En otra realización, la información del sello temporal o de secuencia se escribe para un bloque de borrado cuando se recupera el bloque de borrado.

Cuando el índice de objeto se almacena en memoria volátil, un error, una pérdida de potencia u otro problema que cause que el controlador de almacenamiento 152 se caiga sin almacenar el índice de objeto podría ser un problema si el índice de objeto no se puede reconstruir. El módulo de reconstrucción del índice de objeto 272 permite almacenar el índice de objeto en una memoria volátil permitiendo las ventajas de la memoria volátil, tales como un rápido acceso. El módulo de reconstrucción del índice de objeto 272 permite la rápida reconstrucción del índice de objeto de forma autónoma sin dependencia de un dispositivo externo al dispositivo de almacenamiento 150.

En una realización, el índice de objeto en la memoria volátil se almacena periódicamente en un dispositivo de almacenamiento de datos 154. En un ejemplo particular, el índice de objeto, o "metadatos de índice" se almacena periódicamente en un almacenamiento de estado sólido 110. En otra realización, los metadatos de índices se almacenan en un almacenamiento de estado sólido 110n separado del almacenamiento de estado sólido 110a-110n-1 que almacena paquetes. Los metadatos de índices se gestionan de forma independiente de los datos y los metadatos de objetos transmitidos desde un dispositivo solicitante 155 y gestionados por el controlador de almacenamiento 152 / controlador del dispositivo de almacenamiento de estado sólido 202. La gestión y el almacenamiento de los metadatos de índices por separado de los otros datos y metadatos de un objeto permite un flujo de datos eficiente sin el controlador de almacenamiento 152 / controlador del dispositivo de almacenamiento de estado sólido 202 que procesa innecesariamente los metadatos de objetos.

En una realización, donde una petición de objeto recibida por el módulo de recepción de peticiones de objetos 260 incluye una petición de escritura, el controlador de almacenamiento 152 recibe uno o más segmentos de datos de un objeto desde la memoria de un dispositivo solicitante 155 como una operación de acceso directo a memoria local o remota ("DMA", "RDMA"). En un ejemplo preferido, el controlador de almacenamiento 152 extrae los datos desde la memoria del dispositivo solicitante 155 en uno o más operaciones de DMA o RDMA. En otro ejemplo, el dispositivo solicitante 155 introduce los segmentos de datos al controlador de almacenamiento 152 en una o más operaciones de DMA o RDMA. En otra realización, donde la petición de objetos incluye una petición de lectura, el controlador de almacenamiento 152 transmite uno o más segmentos de datos de un objeto a la memoria del dispositivo solicitante 155 en una o más operaciones de DMA o RDMA. En un ejemplo preferido, el controlador de almacenamiento 152 introduce datos a la memoria del dispositivo solicitante 155 en una o más operaciones de DMA o RDMA. En otro ejemplo, el dispositivo solicitante 155 introduce datos desde el controlador de almacenamiento 152 en una o más operaciones de DMA o RDMA. En otro ejemplo, el controlador de almacenamiento 152 extrae conjuntos de peticiones de comandos de objetos desde la memoria del dispositivo solicitante 155 en una o más operaciones de DMA o DRMA. En otro ejemplo, el dispositivo solicitante 155 introduce conjuntos de peticiones de comandos de objetos al controlador de almacenamiento 152 en una o más operaciones de DMA o RDMA.

En una realización, el controlador de almacenamiento 152 emula el almacenamiento de bloques y un objeto comunicado entre el dispositivo solicitante 155 y el controlador de almacenamiento 152 comprende uno o más bloques de datos. En una realización, el dispositivo solicitante 155 incluye una unidad de modo que el dispositivo de almacenamiento 150 aparece como un dispositivo de almacenamiento de bloques. Por ejemplo, el dispositivo solicitante 155 puede enviar un bloque de datos de cierto tamaño junto con una dirección física de donde quiere el dispositivo solicitante 155 que se almacene el bloque de datos. El controlador de almacenamiento 152 recibe el bloque de datos y usa la dirección del bloque físico transmitido con el bloque de datos o la transformación de la dirección de bloque físico como un identificador de objeto. El controlador de almacenamiento 152 almacena a continuación el bloque de datos como un objeto o segmento de datos de un objeto empaquetando el bloque de datos y almacenando el bloque de datos a voluntad. El módulo de índices de objetos 266 crea a continuación una entrada en el índice de objeto usando el identificador de objeto basado en bloques físicos y la localización física real donde el controlador de almacenamiento 152 almacenó los paquetes de datos comprendiendo los datos desde el bloque de datos.

En otra realización, el controlador de almacenamiento 152 emula el almacenamiento de bloques aceptando objetos de bloques. Un objeto de bloque puede incluir uno o más bloques de datos en una estructura de bloques. En una realización, el controlador de almacenamiento 152 trata el objeto de bloques como cualquier otro objeto. En otra realización, un objeto puede representar un dispositivo de bloque entero, una partición de un dispositivo de bloques, o algún otro sub-elemento lógico o físico de un dispositivo de bloques incluyendo una pista, sector, canal y similares. Es de notar en particular la capacidad de re-mapear un grupo de RAID de dispositivos de bloque para un objeto que

soporta una diferente construcción de RAID tal como una RAID progresiva. Un experto en la materia reconocerá otros mapeos de dispositivos de bloques tradicionales o futuros para los objetos.

Dispositivo de almacenamiento de estado sólido

La Figura 2B es un diagrama de bloques esquemático que ilustra una realización 201 de un controlador de dispositivo de almacenamiento de estado sólido 202 que incluye una conducción de datos de escritura 106 y una conducción de datos de lectura 108 en un dispositivo de almacenamiento de estado sólido 102 de acuerdo con la presente invención. El controlador del dispositivo de almacenamiento de estado sólido 202 puede incluir varios controladores de almacenamiento de estado sólido 0 - N 104a - n, controlando cada uno el almacenamiento de estado sólido 110. En la realización representada, se muestran dos controladores de estado sólido: el controlador de estado sólido 0 140a y el controlador de almacenamiento de estado sólido N 104n, y cada uno controla el almacenamiento de estado sólido 110a - n. En la realización representada, el controlador de almacenamiento de estado sólido 0 140a controla un canal de datos de modo que el almacenamiento de estado sólido conectado 110a almacena los datos. El controlador de almacenamiento de estado sólido N 104n controla un canal de metadatos de índices asociado con los datos almacenados y el almacenamiento de estado sólido asociado 110n almacena los metadatos de índices. En una realización alternativa, el controlador del dispositivo de almacenamiento de estado sólido 202 incluye un controlador de estado sólido único 104a con un almacenamiento de estado sólido único 110a. En otra realización hay una pluralidad de controladores de almacenamiento de estado sólido 104a - n y el almacenamiento de estado sólido asociado 110a - n. En una realización, uno o más controladores de estado sólido 104a - 104n - 1, acopados a su almacenamiento de estado sólido asociado 110a - 110n - 1 controlan los datos mientras que al menos un controlador de almacenamiento de estado sólido 104n, acoplado a su almacenamiento de estado sólido asociado 110n, controla los metadatos de índices.

En una realización, al menos un controlador de estado sólido 104 es una red de puertas programable en campo ("FPGA") y las funciones de controlador se programan dentro de la FPGA. En una realización particular, la FPGA es una FPGA de Xinlix®. En otra realización, el controlador de almacenamiento de estado sólido 104 comprende componentes específicamente diseñados como un controlador de almacenamiento de estado sólido 104, tal como un circuito integrado de aplicación específica ("ASIC") o una solución de lógica a medida. Cada uno de los controladores de almacenamiento de estado sólido 104, usualmente incluye una conducción de datos de escritura 106 y una conducción de datos de lectura 108, que se describen además con relación a la Figura 3. En otra realización, al menos un controlador de almacenamiento de estado sólido 104 está constituido de una combinación de FPGA, ASIC y componentes de lógica a medida.

Almacenamiento de estado sólido

El almacenamiento de estado sólido 110 es una red de elementos de almacenamiento de estado sólido no volátiles 216, 218, 220, dispuestos en bancos 214 y accedidos en paralelo a través de un bus de entrada / salida ("I/O") de almacenamiento bidireccional 210. El bus de almacenamiento I/O 210, en una realización, es capaz de una comunicación unidireccional en un momento cualquier. Por ejemplo, cuando se están escribiendo datos en el almacenamiento de estado sólido 110, los datos no se pueden leer desde el almacenamiento de estado sólido 110. En otra realización, los datos pueden fluir en ambas direcciones simultáneamente. Sin embargo, el modo bidireccional, como se usa en este documento con respecto a un bus de datos, se refiere a una trayectoria de datos que puede tener el flujo de datos en una única dirección en un momento, pero cuando se para el flujo de datos en una dirección sobre el bus de datos bidireccional, los datos pueden fluir en la dirección opuesta sobre el bus de datos bidireccional.

Un elemento de almacenamiento de estado sólido (por ejemplo, SSS 0.0 216a) está usualmente configurado como un chip (un empaquetamiento de uno o más dados) o un dado sobre el circuito impreso. Como se representa, un elemento de almacenamiento de estado sólido (por ejemplo, 216a) opera independientemente o semi-independientemente de otros elementos de almacenamiento de estado sólido (por ejemplo, 218a) incluso si estos diversos elementos están empaquetados juntos en un empaquetamiento de chip, una pila de empaquetamientos de chips, o algunos otros elementos de empaquetamiento. Como se representa, una columna de elementos de almacenamiento de estado sólido 216, 218, 220 se designa como un banco 214. Como se representa, puede haber "n" bancos 214a - n y "m" elementos de almacenamiento de estado sólido 216a - m, 218a - m, 220a - m por banco en una red de nxm elementos de almacenamiento de estado sólido 216, 218, 220, en un almacenamiento de estado sólido 110. En una realización, un almacenamiento de estado sólido 110a incluye veinte elementos de almacenamiento de estado sólido 216, 218, 220 por banco 214 con ocho bancos 214 y un almacenamiento de estado sólido 110n incluye 2 elementos de almacenamiento de estado sólido 216, 218 por banco 214 con un banco 214. En una realización, cada elemento de almacenamiento de estado sólido 216, 218, 220 está comprendido de dispositivos de célula de un nivel único ("SLC"). En otra realización, cada elemento de almacenamiento de estado sólido 216, 218, 220 está comprendido de dispositivos de célula multi-nivel ("MLC").

En una realización, los elementos de almacenamiento de estado sólido para múltiples bancos que comparten una fila 210a del bus (I/O) de almacenamiento común (216b, 218b, 220b) están empaquetados juntos. En una realización, un elemento de almacenamiento de estado sólido 216, 218, 220 puede tener uno o más dados por chip

con uno o más chips apilados verticalmente y se puede acceder a cada dado independientemente. En otra realización, un elemento de almacenamiento de estado sólido (por ejemplo, SSS 0.0216a) puede tener uno o más dados virtuales por dado y uno o más dados por chip y uno o más chips apilados verticalmente y se puede acceder a cada dado virtual independientemente. En otra realización, un elemento de almacenamiento de estado sólido SSS 0. 216a puede tener uno o más dados virtuales por dado y uno o más dados por chip con algunos o todos de uno o más dados apilados verticalmente y se puede acceder a cada dado virtual independientemente.

En una realización, dos datos se apilan verticalmente con cuatro pilas por grupo para formar ocho elementos de almacenamiento (por ejemplo, SSS 0.0-SSS 0.8) 216a - 220a, cada uno en un banco separado 214a - n. En otra realización, 20 elementos de almacenamiento (por ejemplo, SSS 0.0 - SSS 20.0) 216 forman un banco virtual 214a de modo que cada uno de los ocho bancos virtuales tiene 20 elementos de almacenamiento (por ejemplo, SSS 0.0 - SSS 20.8) 216, 218, 220. Los datos se envían al almacenamiento de estado sólido 110 sobre el bus I/O del almacenamiento 210 a todos los elementos de almacenamiento de un grupo particular de elementos de almacenamiento (SSS 0.0-SSS 0.8) 216a, 218a, 220a. El bus de control de almacenamiento 212a se usa para seleccionar un banco particular (por ejemplo, el banco 0 - 214a) de modo que los datos recibidos sobre el bus I/O de almacenamiento 210 conectado a todos los bancos 214 se escriben solo al banco seleccionado 214a.

En una realización preferida, el bus I/O de almacenamiento 210 está comprendido de uno o más buses I/O independientes ("IIOBa - m" que comprende 210a.a - m, 210n. a -m) en el que los elementos de almacenamiento de estado sólido dentro de cada fila comparten uno de los buses I/O independientes acceden a cada uno de los elementos de almacenamiento de estado sólido 216, 218, 220 en paralelo de modo que todos los bancos 214 se acceden simultáneamente. Por ejemplo, un canal del bus I/O de almacenamiento 210 puede acceder a un primer elemento de almacenamiento de estado sólido 216a, 218a, 220a de cada banco 214a-n simultáneamente. Un segundo canal del bus I/O de almacenamiento 210 puede acceder a un segundo elemento de almacenamiento de estado sólido 216b, 218b, 220b de cada banco 214a-n simultáneamente. Cada fila del elemento de almacenamiento de estado sólido 216, 218, 220 se accede simultáneamente. En una realización, donde los elementos de almacenamiento de estado sólido 216, 218, 220 son multinivel (apilados físicamente), todos los niveles físicos de los elementos de almacenamiento de estado sólido 216, 218, 220 se acceden simultáneamente. Como se usa en este documento "simultáneamente" también incluye un acceso casi simultáneo donde los dispositivos se acceden en intervalos ligeramente diferentes para evitar ruido de conmutación. Simultáneamente como se usa en este contexto se distinguirá de un acceso secuencial o un acceso serie en el que los comandos y/o datos se envían individualmente uno después de otro.

Usualmente, los bancos 214a-n se seleccionan independientemente usando el bus de control de almacenamiento 212. En una realización, se selecciona un banco 214 usando una activación de chip o una selección de chip. Cuando tanto la selección de chip como la activación de chip están disponibles, el bus de control de almacenamiento 212 puede seleccionar un nivel de un elemento de almacenamiento de estado sólido multinivel 216, 218, 220. En otras realizaciones se usan otros comandos por el bus de control de almacenamiento 212 para seleccionar individualmente un nivel de un elemento de almacenamiento de estado sólido multinivel 216, 218, 220. Los elementos de almacenamiento de estado sólido 216, 218, 220 también se pueden seleccionar mediante una combinación de control e información de dirección transmitida sobre el bus I/O de almacenamiento 210 y el bus de control de almacenamiento 212.

En una realización, cada elemento de almacenamiento de estado sólido 216, 218, 220 se divide en bloques de borrado y cada bloque de borrado se divide en páginas. Una página típica es de 2000 bytes ("2 kB"). En un ejemplo, un elemento de almacenamiento de estado sólido (por ejemplo SSS 0.0) incluye dos registros y puede programar dos páginas de modo que un elemento de almacenamiento de estado sólido de dos registros 216, 218, 220 tiene una capacidad de 4kB. Un banco 214 de 20 elementos de almacenamiento de estado sólido 216, 218, 220 tendría entonces una capacidad de 80 kB de páginas accedidas con la misma dirección que salen de los canales del bus I/O de almacenamiento 210.

Este grupo de páginas en un banco 214 de elementos de almacenamiento de estado sólido 216, 218, 220 de 80 kB se puede llamar una página virtual. De forma similar, un bloque de borrado de cada elemento de almacenamiento 216a - m de un banco 214a puede estar agrupado para formar un bloque de borrado virtual. En una realización preferida, un bloque de borrado de páginas dentro de un elemento de almacenamiento de estado sólido 216, 218, 220 se borra cuando se recibe un comando de borrado dentro de un elemento de almacenamiento de estado sólido 216, 218, 220. Mientras que el tamaño y el número de bloques de borrado, páginas, niveles u otras divisiones lógicas y físicas dentro de un elemento de almacenamiento de estado sólido 216, 218, 220 que se espera que cambien con el tiempo con los adelantos en la tecnología, se espera que muchas realizaciones consistentes con las nuevas configuraciones sean posibles y sean consistentes con la descripción general en este documento.

Usualmente, cuando se escribe un paquete a una localización particular dentro de un elemento de almacenamiento de estado sólido 216, 218, 220, en el que el paquete se intenta escribir a una localización dentro de una página particular que es específica de un bloque de borrado particular de un elemento particular de un banco particular, se envía una dirección física sobre el bus I/O de almacenamiento 210 y se sigue por el paquete. La dirección física contiene suficiente información para el elemento de almacenamiento de estado sólido 216, 218, 220 para dirigir el

paquete a la localización designada dentro de la página. Como todos los elementos de almacenamiento en una fila de elementos de almacenamiento (por ejemplo SSS 0.0 - SSS 0.N 216a, 218a, 220a) se acceden simultáneamente por el bus apropiado dentro del bus I/O de almacenamiento 210a.a, para alcanzar la página adecuada y evitar escribir el paquete de datos a páginas direccionadas de forma similar en la fila de elementos de almacenamiento (SSS 0.0 - SSS 0.N 216a, 218a, 220a), el banco 214a que incluye el elemento de almacenamiento de estado sólido SSS 0.0 216a con la página correcta donde se va a escribir el paquete de datos se selecciona simultáneamente por el bus de control de almacenamiento 212.

De forma similar, un comando de lectura que viaja sobre el bus I/O de almacenamiento 210 requiere un comando simultáneo sobre el bus de control de almacenamiento 212 para seleccionar un banco único 214a y la página apropiada dentro del banco 214a. En una realización preferida, un comando de lectura lee una página entera, y debido a que hay múltiples elementos de almacenamiento de estado sólido 216, 218, 220 en paralelo en un banco 214, se lee una página virtual entera con un comando de lectura. Sin embargo, el comando de lectura se puede descomponer en sub-comandos, como se explicará más adelante con respecto a la intercalación de bancos. También se puede acceder una página virtual en una operación de escritura.

Se puede enviar un comando de borrar un bloque de borrado para borrar un bloque de borrado sobre el bus I/O de almacenamiento 210 con una dirección del bloque de borrado particular para borrar el bloque de borrado particular. Usualmente se puede enviar un comando de borrar un bloque de borrado sobre las trayectorias en paralelo del bus I/O de almacenamiento 210 para borrar un bloque de borrado virtual, cada uno con una dirección de bloque de borrado particular para borrar un bloque de borrado particular. Simultáneamente un banco particular (por ejemplo el banco 0 214a) se selecciona sobre el bus de control de almacenamiento 212 para impedir el borrado de bloques de borrado direccionados de forma similar en todos los bancos (bancos 1 - N 214b -n). También se pueden mandar otros comandos a una localización particular usando una combinación del bus I/O de almacenamiento 210 y el bus de control de almacenamiento 212. Un experto en la materia reconocerá otros modos de seleccionar una localización particular de almacenamiento usando el bus I/O de almacenamiento bidireccional 210 y el bus de control de almacenamiento 212.

En una realización. Los paquetes se escriben secuencialmente en el almacenamiento de estado sólido 110. Por ejemplo, los paquetes se hacen fluir hacia las memorias intermedias de escritura de almacenamiento de un banco 214a de los elementos de almacenamiento 216 y cuando las memorias intermedias están llenas, se programan los paquetes en una página virtual designada. Los paquetes rellenan a continuación las memorias intermedias de escritura de almacenamiento y, cuando están llenas, los paquetes se escriben en la siguiente página virtual. La siguiente página virtual puede estar en el mismo banco 214a o en otro banco (por ejemplo 214b). Esto proceso continúa, de página virtual en página virtual, usualmente hasta que el bloque de borrado virtual está relleno. En otra realización, el flujo puede continuar a través de fronteras de bloque de borrado virtuales continuando el proceso, un bloque de borrado virtual tras otro bloque de borrado virtual.

En una operación de lectura, modificación, escritura los paquetes de datos asociados con el objeto se localizan y se leen en una operación de lectura. Los segmentos de datos del objeto modificado que se ha modificado no se escriben en la localización de la que se leen. En cambio, los segmentos de datos modificados se convierten de nuevo a paquetes de datos y a continuación se escriben en la siguiente localización disponible en la página virtual que se está escribiendo actualmente. Las entradas al índice de objetos para los paquetes de datos respectivos se modifican para apuntar a los paquetes que contienen los segmentos de datos modificados. La entrada o entradas en el índice de objetos para los paquetes de datos asociados con el mismo objeto que no se ha modificado incluirán punteros a la localización original de los paquetes de datos sin modificar. De este modo, si se mantiene el objeto original, por ejemplo para mantener una versión anterior del objeto, el objeto original tendrá punteros en el índice de objetos para todos los paquetes de datos como se escribieron originalmente. El nuevo objeto tendrá punteros en el índice de objeto a algunos paquetes de datos originales y punteros a los paquetes de datos modificados en la página virtual que se ha escrito actualmente.

En una operación de copia, el índice de objetos incluye una entrada para el objeto original mapeado a varios paquetes almacenados en el almacenamiento de estado sólido 110. Cuando se realiza una copia, se crea un nuevo objeto y se crea una nueva entrada en el índice de objetos mapeando el nuevo objeto a los paquetes originales. El nuevo objeto también se escribe al almacenamiento de estado sólido 110 con su localización mapeada a la nueva entrada en el índice de objetos. Los paquetes del nuevo objeto se pueden usar para identificar los paquetes dentro del objeto original al que se refiere en caso de que se hayan realizado cambios en el objeto original que no se han propagado a la copia y el índice de objeto se pierde o se corrompe.

Ventajosamente, la escritura secuencial de paquetes facilita un uso más frecuente del almacenamiento de estado sólido 110 y permite al controlador del dispositivo de almacenamiento sólido 202 monitorizar los puntos calientes del almacenamiento y el nivel de uso de las diversas páginas virtuales en el almacenamiento de estado sólido 110. La escritura secuencial de paquetes también facilita un sistema potente, de recogida de basura eficiente, que se describe con detalle más adelante. Un experto en la materia reconocerá otros beneficios del almacenamiento secuencial de los paquetes de datos.

Controlador del dispositivo de almacenamiento de estado sólido

En diversas realizaciones, el controlador del dispositivo de almacenamiento de estado sólido 202 también incluye un bus de datos 204, un bus local 206, un controlador de la memoria intermedia 208, las memorias intermedias 0 - N 222a - n, un controlador maestro 224, un controlador de acceso directo a memoria ("DMA") 226, un controlador de memoria 228, una red de memoria dinámica 230, una red de memoria aleatoria estática 232, un controlador de gestión 234, un bus de gestión 236, un puente 238 para un bus del sistema 240 y lógica diversa 242, que se describe más adelante. En otras realizaciones, el bus del sistema 240 está acoplado a una o más tarjetas de interfaz de red ("NIC") 244, algunas de las cuales pueden incluir controladores de DMA remoto ("RDMA") 246, una o más unidades de procesamiento central ("CPU") 248, uno o más controladores de memoria externos 250 y redes de memorias externas asociadas 252, uno o más controladores de almacenamiento 254, controladores de pares 256, y procesadores específicos de la aplicación 258 que se describen más adelante. Los componentes 244 - 258 conectados al bus del sistema 240 pueden estar localizados en el ordenador 112 o pueden estar en otros dispositivos.

Usualmente, el controlador del almacenamiento de estado sólido 104 comunica los datos al almacenamiento de estado sólido 110 sobre un bus I/O de almacenamiento 210. En una realización típica donde el almacenamiento de estado sólido está dispuesto en bancos 214 y cada banco 214 incluye múltiples elementos de almacenamiento 216, 218, 220 accedidos en paralelo, el bus I/O de almacenamiento 210 es una red de buses, uno para cada fila de elementos de almacenamiento 216, 218, 220 que abarca los bancos 214. Como se usa en este documento, el término "bus I/O de almacenamiento" se puede referir a un bus I/O de almacenamiento 210 o una red de buses de datos independientes 204. En una red preferida, cada bus I/O de almacenamiento 210 que accede a una fila de elementos de almacenamiento (por ejemplo, 216a, 218a, 220a) puede incluir un mapeo de lógico a físico para las divisiones del almacenamiento (por ejemplo, los bloques de borrado) accedidos en una fila de elementos de almacenamiento 216a, 218a, 220a. Este mapeo permite re-mapear una dirección lógica mapeada a una dirección física de una división del almacenamiento a una división de almacenamiento diferente si falla la primera división de almacenamiento, falla parcialmente, es inaccesible o tiene algún otro problema. El re-mapeo se explica adicionalmente en relación con el módulo de re-mapeo 430 en las Figuras 4A y 4B.

Los datos también se pueden comunicar a los controladores de almacenamiento de estado sólido 104 desde un dispositivo solicitante 155 a través del bus del sistema 240, el puente 238, el bus local 206, las memorias intermedias 222, y finalmente sobre un bus de datos 204. El bus de datos 204 usualmente está conectado a una o más memorias intermedias 222a - n controladas con un controlador de memoria intermedia 208. El controlador de memorias intermedias 208 usualmente controla la transferencia de datos desde el bus local 206 a las memorias intermedias 222 y a través del bus de datos 204 a la memoria intermedia de entrada de la conducción 306 y la memoria intermedia de salida 330. El controlador de la memoria intermedia 208 usualmente controla cómo se pueden almacenar temporalmente en una memoria intermedia 222 los datos que llegan desde un dispositivo solicitante 155 y a continuación transmitirlos sobre un bus de datos 204, o viceversa, para tener en cuenta los diferentes dominios de reloj, para impedir colisiones de datos, etc. El controlador de la memoria intermedia 208 usualmente funciona en conjunción con el controlador maestro 224 para coordinar los flujos de datos. A medida que llegan los datos, los datos que llegan sobre el bus del sistema 240, se transfieren al bus local 206 a través de un puente 238.

Usualmente, los datos se transfieren desde el bus local 206 a una o más memorias intermedias de datos 222 según se dirigen por el controlador maestro 224 y el controlador de la memoria intermedia 208. Los datos fluyen a continuación desde las memorias intermedias 222 al bus de datos 204, a través de un controlador de estado sólido 104 y sobre el almacenamiento de estado sólido 110 tal como una memoria flash NAND u otro medio de almacenamiento. En una realización preferida, los datos y los metadatos asociados fuera de banda ("metadatos del objeto") que llegan con los datos se comunican usando uno o más canales de datos que comprenden uno o más controladores de almacenamiento de estado sólido 104a - 104n-1 y almacenamiento de estado sólido asociado 110a - 110n-1 mientras que el menos un canal (controlador de almacenamiento de estado sólido 104n, almacenamiento de estado sólido 110n) se dedica a los metadatos en banda, tales como la información de índices y otros metadatos generados internamente para el dispositivo de almacenamiento de estado sólido 102.

El bus local 206 es usualmente un bus bidireccional o un conjunto de buses que permiten la comunicación de datos y comandos entre dispositivos internos al controlador del dispositivo de almacenamiento de estado sólido 202 y entre los dispositivos internos al dispositivo de almacenamiento de estado sólido 102 y los dispositivos 244 - 258 conectados al bus del sistema 240. El puente 238 facilita la comunicación entre el bus local 206 y el bus del sistema 240. Un experto en la materia reconocerá otras realizaciones tales como las estructuras de anillo o configuraciones de estrella conmutadas y las funciones de los buses 240, 206, 204, 210 y los puentes 238.

El bus del sistema 240 es usualmente un bus de un ordenador 112 u otro dispositivo en el que el dispositivo de almacenamiento de estado sólido 102 está instalado o conectado. En una realización, el bus del sistema 240 puede ser un bus PCI- e, un bus de Conexión de Tecnología Avanzada Serie ("ATA serie"), un bus ATA en paralelo, o similares. En otra realización, el bus del sistema 240 es un bus externo tal como una interfaz de un pequeño sistema de ordenadores ("SCSI"), FireWire, Fiber Channel, USB, PCIe-AS o similares. El dispositivo de almacenamiento de

estado sólido 102 puede estar empaquetado para fijarse internamente a un dispositivo o como un dispositivo conectado externamente.

El controlador del dispositivo de almacenamiento de estado sólido 202 incluye un controlador maestro 224 que controla las funciones de más alto nivel dentro del dispositivo de almacenamiento de estado sólido 102. El controlador maestro 224, en diversas realizaciones, controla el flujo de datos interpretando las peticiones de objetos y otras peticiones, dirige la creación de índices para mapear los identificadores de objetos asociados con datos a localizaciones físicas de datos asociados, coordinando peticiones de DMA, etc. Muchas de las funciones descritas en este documento se controlan completamente o en parte por el controlador maestro 224.

En una realización, el controlador maestro 224 usa controladores incorporados. En otra realización, el controlador maestro 224 usa la memoria local tal como una red de memoria dinámica 230 (memoria de acceso aleatorio dinámica "DRAM"), una red de memoria estática 232 (memoria de acceso aleatorio estática "SRAM"), etc. En una realización, la memoria local se controla usando el controlador maestro 224. En otra realización, el controlador maestro 224 accede a la memoria local a través de un controlador de memoria 228. En otra realización, el controlador maestro 224 se ejecuta en un servidor Linux y puede soportar diversas interfaces de servidor común, tales como el World Wide Web, el lenguaje de marcación de hipertexto ("HTML"), etc. En otra realización, el controlador maestro 224 usa un nano-procesador. El controlador maestro 224 se puede construir usando lógica programable o estándar, o cualquier combinación de tipos de los controladores listados anteriormente. Un experto en la materia reconocerá muchas realizaciones para el controlador maestro 224.

En una realización, donde el controlador de almacenamiento 152 / controlador de dispositivo de almacenamiento de estado sólido 202 gestiona múltiples dispositivos de almacenamiento de datos / almacenamiento de estado sólido 110a-n, el controlador maestro 224 divide la carga de trabajo entre controladores internos, tales como los controladores de almacenamiento de estado sólido 104a - n. Por ejemplo, el controlador maestro 224 puede dividir un objeto a escribir en los dispositivos de almacenamiento de datos (por ejemplo el almacenamiento de estado sólido 110a - n) de modo que una porción del objeto se almacena sobre cada uno de los dispositivos de almacenamiento de datos conectados. Esta característica es una mejora de funcionamiento que permite un almacenamiento y acceso más rápidos a un objeto. En una realización, el controlador maestro 224 se implementa usando una FPGA. En otra realización, el firmware dentro del controlador maestro 224 se puede actualizar a través del bus de gestión 236, el bus del sistema 240 sobre una red conectada a una NIC 244 u otro dispositivo conectado al bus del sistema 240.

En una realización, el controlador maestro 224, que gestiona objetos, emula el almacenamiento de bloques de modo que un ordenador 112 u otro dispositivo conectado al dispositivo de almacenamiento 152 / dispositivo de almacenamiento de estado sólido 102 ve el dispositivo de almacenamiento de estado sólido 102 como un dispositivo de almacenamiento de bloques y envía datos a una dirección física específica en el dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102. A continuación el controlador maestro 224 divide los bloques y almacena los bloques de datos como si fuesen objetos. El controlador maestro 224 mapea a continuación los bloques y las direcciones físicas enviadas con el bloque a las localizaciones reales determinadas por el controlador maestro 224. El mapeo se almacena en el índice de objetos. Usualmente, para la emulación de bloques, una interfaz de programa de aplicación ("API") de dispositivos de bloques se proporciona en una unidad en el ordenador 112, en el cliente 114, o en otro dispositivo que desee usar el dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102 como un dispositivo de almacenamiento de bloques.

En otra realización, el controlador maestro 224 coordina con los controladores de NIC 244 y los controladores de RDMA incorporados 246 para entregar las transferencias de datos de RDMA justo a tiempo y los conjuntos de comandos. El controlador de NIC 244 puede estar oculto detrás de un puerto no transparente para posibilitar el uso de las unidades del cliente. También, una unidad sobre un cliente 114 puede tener acceso a la red de ordenadores 116 a través de una unidad de memoria de I/O usando una API de pila normalizada y operando en conjunción con las NIC 244.

En una realización, el controlador maestro 224 es también un controlador de la red redundante de unidades independientes ("RAID"). Donde el dispositivo de almacenamiento de datos / dispositivo de almacenamiento de estado sólido 102 está conectado en red con uno o más de otros dispositivos de almacenamiento de datos / dispositivos de almacenamiento de estado sólido 102, el controlador maestro 224 puede ser un controlador de RAID para una RAID de nivel único, una RAID multinivel, una RAID progresiva, etc. El controlador maestro 224 también permite almacenar algunos objetos en una red de RAID y almacenar otros objetos sin RAID. En otra realización, el controlador maestro 224 puede ser un elemento de controlador de RAID distribuido. En otra realización, el controlador maestro 224 puede comprender muchas RAID, RAID distribuidas y otras funciones como se describe en otras partes.

En una realización, el controlador maestro 224 coordina con gestores de red únicos o redundantes (por ejemplo, conmutadores) para establecer un enrutamiento, para equilibrar el uso de ancho de banda, conmutación por error, etc. En otra realización, el controlador maestro 224 coordina con lógica específica de la aplicación integrada (a través del bus local 206) y software de la unidad asociada. En otra realización, el controlador maestro 224 coordina

con procesadores específicos de la aplicación conectados 258 o lógica (a través del bus del sistema externo 240) y el software de la unidad asociada. En otra realización, el controlador maestro 224 coordina con la lógica específica de la aplicación remota (a través de la red de ordenadores 116) y el software de la unidad asociada. En otra realización, el controlador maestro 224 coordina con el bus local 206 o el bus externo conectado al controlador de almacenamiento de la unidad de disco duro ("HDD").

En una realización, el controlador maestro 224 comunica con uno o más controladores de almacenamiento 254 donde el dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102 puede aparecer como un dispositivo de almacenamiento conectado a través de un bus SCSI, SCSI de Internet, ("iSCSI"), canal de fibra, etc. Entre tanto el dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102 puede gestionar objetos de forma autónoma y puede aparecer como un sistema de ficheros de objetos o un sistema de ficheros de objetos distribuido. El controlador maestro 224 también se puede acceder por los controladores de pares 256 y/o procesadores específicos de la aplicación 258.

En otra realización, el controlador maestro 224 coordina con un controlador de gestión autónomo integrado para validad periódicamente el código de FPGA y/o el software del controlador, validar el código de FPGA mientras que se ejecuta (reinicio) y/o validar el software del controlador durante el encendido (reinicio), soporta las peticiones de reinicio externas, soporta las peticiones de inicio debidas a excesos de tiempo de ejecución (watchdog timeout), y soporta las mediciones de voltaje, corriente, potencia, temperatura, y otras mediciones del entorno y el establecimiento de interrupciones por umbral. En otra realización, el controlador maestro 224 gestiona la recogida de basura para liberar a los bloques de borrado para su reutilización. En otra realización, el controlador maestro 224 gestiona la nivelación de uso. En otra realización, el controlador maestro 224 permite al dispositivo de almacenamiento de datos / dispositivo de almacenamiento de estado lógico 102 dividirse en múltiples dispositivos virtuales y permite el cifrado de medios con base en la partición. En otra realización más, el controlador maestro 224 soporta un controlador de almacenamiento de estado sólido 104 con una avanzada corrección de ECC multi-bit. Un experto en la materia reconocerá otras características y funciones de un controlador maestro 224 en un controlador de almacenamiento 152, o más específicamente en un dispositivo de almacenamiento de estado sólido 102.

En una realización, el controlador del dispositivo de almacenamiento de estado sólido 202 incluye un controlador de memoria 228 que controla una red de memoria aleatoria dinámica 230 y/o una red de memoria aleatoria estática 232. Como se ha establecido anteriormente, el controlador de memoria 228 puede ser independiente o estar integrado con el controlador maestro 224. El controlador de memoria 228 usualmente controla la memoria volátil de algún tipo tal como una DRAM (red de memoria aleatoria dinámica 230) y SRAM (red de memoria aleatoria estática 232). En otros ejemplos, el controlador de memoria 228 también controla otros tipos de memoria tales como la memoria de solo lectura programable y borrable eléctricamente ("EEPROM"), etc. En otras realizaciones, el controlador de memoria 228 controla dos o más tipos de memoria y el controlador de memoria 228 puede incluir más de un controlador. Usualmente, el controlador de memoria 228 controla tanta SRAM 232 como sea factible y la DRAM 230 para suplementar la SRAM 232.

En una realización, el índice de objeto se almacena en una memoria 230, 232 y a continuación se descarga periódicamente a un canal del almacenamiento de estado sólido 110n u otra memoria no volátil. Un experto en la materia reconocerá otros usos y configuraciones del controlador de memoria 228, la red de memoria dinámica 230 y la red de memoria estática 232.

En una realización, el controlador del dispositivo de almacenamiento de estado sólido 202 incluye un controlador de DMA 226 que controla las operaciones de DMA entre el dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102 y uno o más controladores de memoria externos 250 y redes de memoria externas asociadas 252 y las CPU 248. Obsérvese que los controladores de memoria externos 250 y las redes de memoria externas 252 se llaman externas porque son externas para el dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102. Además el controlador de DMA 226 también puede controlar las operaciones de RDMA con los dispositivos solicitantes a través de una NIC 244 y el controlador de RDMA asociado 246. El DMA y el RDMA se explican con más detalle más adelante.

En una realización, el controlador del dispositivo de almacenamiento de estado sólido 202 incluye un controlador de gestión 234 conectado a un bus de gestión 236. Usualmente el controlador de gestión 234 gestiona las métricas del entorno y el estado del dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102. El controlador de gestión 234 puede monitorizar la temperatura del dispositivo, la velocidad del ventilador, los parámetros de la fuente de alimentación, etc. sobre el bus de gestión 236. El controlador de gestión 234 puede soportar la lectura y programación de la memoria de solo lectura programable y borrable ("EEPROM") para el almacenamiento del código de la FPGA y el software del controlador. Usualmente el bus de gestión 236 está conectado a los diversos componentes dentro del dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102. El controlador de gestión 234 puede comunicar alertas, interrupciones, etc. sobre el bus local 206 o puede incluir una conexión separada a un bus de sistema 240 u otro bus. En una realización el bus de gestión 236 es un bus Entre Circuitos Integrados ("I^2C"). Un experto en la materia reconocerá otras funciones y usos relacionados con un controlador de gestión 234 conectado a los componentes del dispositivo de almacenamiento / dispositivo de almacenamiento de estado sólido 102 por un bus de gestión 236.

En una realización, el controlador del dispositivo de almacenamiento de estado sólido 202 incluye lógica diversa 242 que se puede adaptar para una aplicación específica. Usualmente cuando el controlador del dispositivo de estado sólido 202 o el controlador maestro 224 esta / están configurados usando una FPGA u otro controlador configurable, se puede incluir lógica a medida en base a una aplicación particular, requisitos del cliente, requisitos de almacenamiento, etc.

Conducción de datos

La Figura 3 es un diagrama de bloques esquemático que ilustra una realización 300 de un controlador de almacenamiento de estado sólido 104 con una conducción de datos de escritura 106 y una conducción de datos de lectura 108 en un dispositivo de almacenamiento de estado sólido 102 de acuerdo con la presente invención. La realización 300 incluye un bus de datos 204, un bus local 206, y un control de la memoria intermedia 208, que son sustancialmente similares a los descritos en relación con el controlador del dispositivo de almacenamiento de estado sólido 202 de la Figura 2. La conducción de datos de escritura 106 incluye un empaquetador 302 y un generador del código de corrección de errores ("ECC") 304. En otras realizaciones, la conducción de datos de escritura 106 incluye una memoria intermedia de entrada 306, una memoria intermedia de sincronización de la escritura 308, un módulo del programa de escritura 310, un módulo de compresión 312, un módulo de cifrado 314, una derivación del recogedor de basura 316 (con una porción dentro de la conducción de datos de lectura 108), un módulo de cifrado de medios 318, y una memoria intermedia de escritura 320. La conducción de datos de lectura 108 incluye una memoria intermedia de sincronización de lectura 328, un módulo de corrección de ECC 322, un des-empaquetador 324, un módulo de alineamiento 326 y una memoria intermedia de salida 330. En otras realizaciones, la conducción de datos de lectura 108 puede incluir un módulo de descifrado de medios 332, una porción de la derivación del recogedor de basuras 316, un módulo de descifrado 334, un módulo de descompresión 336, y un módulo del programa de lectura 338. El controlador del almacenamiento de estado sólido 104 también puede incluir los registros de control y estado 340 y las colas de control 342, el controlador de intercalado de bancos 344, una memoria intermedia de sincronización 346, un controlador del bus de almacenamiento 348, y un multiplexor ("MUX") 350. Los componentes del controlador de estado sólido 104 y la conducción de datos de escritura asociada 106 y la conducción de datos de lectura 108 se describen más adelante. En otras realizaciones se puede usar el almacenamiento de estado sólido síncrono 110 y se pueden eliminar las memorias intermedias de sincronización 308, 328.

Conducción de datos de escritura

La conducción de datos de escritura 106 incluye un empaquetador 302 que recibe un segmento de datos o metadatos a escribir al almacenamiento de estado sólido, bien directamente o indirectamente a través de otra etapa de la conducción de datos de escritura 106, y crea uno o más paquetes dimensionados para el almacenamiento de estado sólido 110. El segmento de datos o metadatos es usualmente parte de un objeto, pero también puede incluir un objeto entero. En otra realización, el segmento de datos es parte de un bloque de datos, pero puede incluir también un objeto entero de datos. Usualmente se recibe un objeto desde un ordenador 112, el cliente 114, u otro ordenador o dispositivo y se transmite al dispositivo de almacenamiento de estado sólido 102 en segmentos de datos transmitidos en flujo continuo al dispositivo de almacenamiento de estado sólido 102 o el ordenador 112. Un segmento de datos, también puede ser conocido por otro nombre, tal como una parcela de datos pero como se denomina en este documento incluye todo o una porción de un objeto o bloque de datos.

Cada objeto se almacena como uno o más paquetes. Cada objeto puede tener uno o más paquetes de contenedor. Cada paquete contiene una cabecera. La cabecera puede incluir un campo del tipo de cabecera. Los campos de tipo pueden incluir datos, atributos del objeto, metadatos, delimitadores de segmentos de datos (multi-paquete), estructuras de objetos, enlaces de objetos, y similares. La cabecera también puede incluir información con respecto al tamaño del paquete, tal como el número de bytes de datos incluidos en el paquete. La longitud del paquete se puede establecer por el tipo de paquete. La cabecera puede incluir información que establece la relación del paquete con el objeto. Un ejemplo podría ser el uso de un desplazamiento en una cabecera de paquete de datos para identificar la localización del segmento de datos dentro del objeto. Un experto en la técnica reconocerá otra información que se puede incluir en la cabecera añadida a los datos por el empaquetador 302 y otra información que se puede añadir a un paquete de datos.

Cada paquete incluye una cabecera y posiblemente datos del segmento de datos o metadatos. La cabecera de cada paquete incluye la información pertinente para relacionar el paquete con el objeto al que pertenece el paquete. Por ejemplo, la cabecera puede incluir un identificador de objetos y un desplazamiento que indica el segmento de datos, objeto, o bloque de datos a partir del cual se forma el paquete de datos. La cabecera también puede incluir una dirección lógica usada por el controlador del bus de almacenamiento 348 para almacenar el paquete. La cabecera también puede incluir información con respecto al tamaño del paquete, tal como el número de bytes incluidos en el paquete. La cabecera también puede incluir un número de secuencia que identifica a dónde pertenece el segmento de datos con respecto a otros paquetes dentro del objeto cuando se reconstruye el segmento de datos o el objeto. La cabecera puede incluir un campo de tipo de cabecera. Los campos de tipo pueden incluir datos, atributos de objetos, metadatos, delimitadores del segmento de datos (multi-paquete), estructuras de objetos, enlaces de objetos,

y similares. Un experto en la materia reconocerá otra información que se puede incluir en una cabecera añadida a los datos o metadatos por un empaquetador 302 y otra información que se puede añadir a un paquete.

La conducción de datos de escritura 106 incluye un generador de ECC 304 que genera uno o más códigos de corrección de errores ("ECC") para el uno o más paquetes recibidos desde el empaquetador 302. El generador de ECC 304 usualmente usa un algoritmo de corrección de errores para generar el ECC que se almacena con el paquete. El ECC almacenado con el paquete se usa usualmente para detectar y corregir errores introducidos dentro de los datos a través de la transmisión y almacenamiento. En una realización, los paquetes se transmiten de forma continua dentro del generador de ECC 304 como bloques no codificados de longitud N. Se calcula un síndrome de longitud S, se añade y se saca como un bloque codificado de longitud N + S. El valor de N y S son dependientes de las características del algoritmo que se selecciona para conseguir el funcionamiento específico, eficacia, y métricas de robustez. En la realización preferida no hay ninguna relación fija entre los bloques de ECC y los paquetes, el paquete puede comprender más de un bloque de ECC; el bloque de ECC puede comprender más de un paquete; y un primer paquete puede terminar en cualquier punto dentro del bloque de ECC y un segundo paquete puede comenzar después del fin del primer paquete dentro del mismo bloque de ECC. En la realización preferida, los algoritmos de ECC no se modifican dinámicamente. En una realización preferida, el ECC almacenado con los paquetes de datos es suficientemente robusto para corregir errores en más de dos bits.

Ventajosamente, el uso de un algoritmo robusto de ECC que permita la corrección de más de un único bit o incluso la corrección de un doble bit permite que la vida del almacenamiento de estado sólido 110 se prolongue. Por ejemplo si se usa memoria flash como medio de almacenamiento en el almacenamiento de estado sólido 110, la memoria flash se puede escribir aproximadamente 100.000 veces sin error por ciclo de escritura. Esta limitación de uso se puede extender usando un algoritmo de ECC robusto. Teniendo el generador de ECC 304 y el módulo de corrección de ECC correspondiente 322 incorporado en el dispositivo de almacenamiento de estado sólido 102, el dispositivo de almacenamiento de estado sólido 102 puede corregir errores internamente y tener una vida útil mayor que si se usa un algoritmo de ECC menos robusto, tal como la corrección de un único bit. Sin embargo, en otras realizaciones el generador de ECC 304 puede usar un algoritmo menos robusto y puede corregir errores de un único bit o de doble bit. En otra realización el dispositivo de almacenamiento de estado sólido 110 puede comprender un almacenamiento menos fiable tal como una memoria flash de célula multi-nivel ("MLC") para aumentar la capacidad, cuyo almacenamiento puede no ser suficientemente fiable sin algoritmos de ECC más robustos.

En una realización, la conducción de escritura 106 incluye una memoria intermedia de entrada 306 que recibe un segmento de datos a escribir al almacenamiento de estado sólido 110 y almacena los segmentos de datos entrantes hasta la siguiente etapa de la conducción de datos de escritura 106, tal como el empaquetador 302 (u otra etapa para una conducción de datos de escritura más compleja 106) es fácil procesar el siguiente segmento de datos. La memoria intermedia de entrada 306 usualmente permite discrepancias entre la tasa a la que se reciben los segmentos de datos y la tasa a la que se procesan por la conducción de datos de escritura 106 usando una memoria intermedia de datos dimensionada apropiadamente. La memoria intermedia de entrada 306 también permite que el bus de datos 204 transfiera datos a la conducción de datos de escritura 106 a tasas mayores que la que se puede sostener por la conducción de datos de escritura 106 para mejorar la eficacia de la operación del bus de datos 204. Usualmente, cuando la conducción de datos de escritura 106 no incluye una memoria intermedia de entrada 306, se realiza una función de almacenamiento temporal en otra parte, tal como en el dispositivo de almacenamiento de estado sólido 102 pero fuera de la conducción de datos de escritura 106, en el ordenador 112, tal como dentro de una tarjeta de interfaz de red ("NIC") u otro dispositivo, por ejemplo cuando se usa el acceso de memoria directo remoto ("RDMA").

En otra realización, la conducción de datos de escritura 106 también incluye una memoria intermedia de sincronización de escritura 308 que almacena temporalmente los paquetes recibidos desde el generador de ECC 304 antes de escribir los paquetes al almacenamiento de estado sólido 110. La memoria intermedia de sincronización de escritura 308 está localizada en una frontera entre un domino de reloj local y un dominio de reloj del almacenamiento de estado sólido y proporciona el almacenamiento temporal tener en cuenta las diferencias en los dominios de reloj. En otras realizaciones se puede usar el almacenamiento de estado sólido síncrono 110 y se pueden eliminar las memorias intermedias de sincronización 308, 328.

En una realización, la conducción de datos de escritura 106 también incluye un módulo de cifrado de medios 318 que recibe el uno o más paquetes desde el empaquetador 302, bien directamente o indirectamente, y cifra el uno o más paquetes usando una clave de cifrado única para el dispositivo de almacenamiento de estado sólido 102 antes de enviar los paquetes al generador de ECC 304. Usualmente, se cifra todo el paquete, incluyendo las cabeceras. En otra realización las cabeceras no se cifran. En este documento, la clave de cifrado se entiende que significa una clave de cifrado secreta que se gestiona externamente desde una realización que integra el almacenamiento de estado sólido 110 y donde la realización requiere la protección de cifrado. El módulo de cifrado de medios 318 y el módulo de descifrado de medios correspondiente 332 proporcionan un nivel de seguridad para los datos almacenados en el almacenamiento de estado sólido 110. Por ejemplo, cuando los datos se cifran con el módulo de cifrado de medios 318, si el almacenamiento de estado sólido 110 se conecta a un controlador de almacenamiento de estado sólido diferente 104, el dispositivo de almacenamiento de estado sólido 102 o el ordenador 112, los

contenidos del almacenamiento de estado lógico 110 usualmente no se podrían leer sin el uso de la misma clave de cifrado usada durante la escritura de datos al almacenamiento de estado sólido 110 sin un esfuerzo significativo.
En una realización típica, el dispositivo de almacenamiento de estado sólido 102 no almacena la clave de cifrado en el almacenamiento no volátil y no permite ningún acceso externo a la clave de cifrado. La clave de cifrado se proporciona al controlador de almacenamiento de estado sólido 104 durante la inicialización. El dispositivo de almacenamiento de estado sólido 102 puede usar y almacenar un número arbitrario criptográfico no secreto que se usa en conjunción con una clave de cifrado. Un número arbitrario diferente se puede almacenar con cada paquete. Los segmentos de datos se pueden dividir entre múltiples paquetes con números arbitrarios únicos para el propósito de mejorar la protección por el algoritmo de cifrado. La clave de cifrado se puede recibir desde un cliente 114, un ordenador 112, un gestor de claves u otro dispositivo que gestiona la clave de cifrado a usar por el controlador de almacenamiento de estado sólido 104. En otra realización, el almacenamiento de estado sólido 110 puede tener dos o más particiones y el controlador de almacenamiento de estado sólido 104 se comporta entonces como si fuesen dos o más controladores de almacenamiento de estado sólido 104, operando cada uno sobre una partición única dentro del almacenamiento de estado sólido 110. En esta realización, se puede usar una clave de cifrado de medios única con cada partición.

En otra realización, la conducción de datos de escritura 106 también incluye un módulo de cifrado 314 que cifra un segmento de datos o metadatos recibido desde la memoria intermedia de entrada 306, bien directamente o indirectamente, antes de enviar el segmento de datos al empaquetador 302, usando el segmento de datos cifrado una clave de cifrado recibida en conjunción con el segmento de datos. El módulo de cifrado 314 difiere del módulo de cifrado de medios 318 en que las claves de cifrado usadas por el módulo de cifrado 314 para cifrar los datos no pueden ser comunes para todos los datos almacenados dentro del dispositivo de almacenamiento de estado sólido 102 sino que puede variar sobre la base de un objeto y recibirse en conjunción con los segmentos de datos de recepción como se describe más adelante. Por ejemplo, una clave de cifrado para un segmento de datos a cifrar por el módulo de cifrado 314 se puede recibir con el segmento de datos o se puede recibir como parte de un comando para escribir un objeto al que pertenece el segmento de datos. El dispositivo de almacenamiento de estado sólido 102 puede usar y almacenar un número aleatorio criptográfico no secreto en cada paquete de objeto que se usa en conjunción con la clave de cifrado. Se puede almacenar un número aleatorio diferente con cada paquete. Los segmentos de datos se pueden dividir entre múltiples paquetes con números aleatorios únicos para el propósito de mejorar la protección por el algoritmo de cifrado. En una realización, el número aleatorio usado por el módulo de cifrado de medios 318 es el mismo que se usó por el módulo de cifrado 314.

La clave de cifrado se puede recibir desde un cliente 114, un ordenador 112, un gestor de claves u otro dispositivo que mantiene la clave de cifrado a usar para el cifrado del segmento de datos. En una realización, las claves de cifrado se transfieren al controlador de almacenamiento de estado sólido 104 desde uno de, un dispositivo de almacenamiento de estado sólido 102, un ordenador 112, un cliente 114 y otro agente externo que tiene la capacidad de ejecutar métodos normalizados de la industria para transferir de forma segura y proteger las claves privadas y públicas.

En una realización, el módulo de cifrado 314 cifra un primer paquete con una primera clave de cifrado recibida en conjunción con el paquete y cifra un segundo paquete con una segunda clave de cifrado recibida en conjunción con el segundo paquete. En otra realización, el módulo de cifrado 314 cifra un primer paquete con una primera clave de cifrado recibida en conjunción con el paquete y pasa un segundo paquete de datos sobre la siguiente etapa sin cifrado. Ventajosamente, el módulo de cifrado 314 incluido en la conducción de datos de escritura 106 del dispositivo de almacenamiento de estado sólido 102 permite el cifrado de datos objeto por objeto o segmento por segmento sin un sistema de ficheros único u otro sistema externo para mantener el seguimiento de las diferentes claves de cifrado usadas para almacenar los objetos o segmentos de datos correspondientes. Cada dispositivo solicitante 155 o gestor de claves relacionado gestiona independientemente las claves de cifrado usadas para cifrar solo los objetos o segmentos de datos enviados por el dispositivo solicitante 155.

En otra realización, la conducción de datos de escritura 106 incluye un módulo de compresión 312 que comprime los datos para el segmento de metadatos antes de enviar el segmento de datos al empaquetador 302. El módulo de compresión 312 usualmente comprime un segmento de datos o metadatos usando una rutina de compresión conocida por los expertos en la materia para reducir el tamaño de almacenamiento del segmento. Por ejemplo, si un segmento de datos incluye una cadena de caracteres de 512 ceros, el módulo de compresión 312 puede reemplazar los 512 ceros con un código o clave que indica los 512 ceros donde el código es mucho más compacto que el espacio ocupado por los 512 ceros.

En una realización, el módulo de compresión 312 comprime un primer segmento con una primera rutina de compresión y pasa junto a un segundo segmento sin compresión. En otra realización, el módulo de compresión 312 comprende un primer segmento con una primera rutina de compresión y comprime el segundo segmento con una segunda rutina de compresión. Tener esta flexibilidad dentro del dispositivo de almacenamiento de estado sólido 102 es beneficioso de modo que los clientes 114 u otros dispositivos que escriben datos al dispositivo de almacenamiento de estado sólido 102 pueden especificar cada uno una rutina de compresión o de modo que uno puede especificar una rutina de compresión mientras que otro especifica sin compresión. La selección de rutinas de compresión también se puede seleccionar de acuerdo con configuraciones por defecto sobre cada tipo de objeto o

en base a la clase de objeto. Por ejemplo, un primer objeto de un objeto especificado puede anular el establecimiento de la rutina de compresión por defecto y un segundo objeto de la misma clase de objetos y tipo de objetos puede usar la rutina de compresión por defecto y un tercer objeto de la misma clase de objetos y tipo de objeto puede no usar ninguna compresión.

En una realización, la conducción de datos de escritura 106 incluye una derivación del recogedor de basura 316 que recibe segmentos de datos desde la conducción de datos de lectura 108 como parte de una derivación de datos en un sistema de recogida de basura. Un sistema de recogida de basura usualmente marca los paquetes que ya no son válidos, usualmente porque el paquete se marca para su borrado o porque se han modificado y almacenado los datos modificados en una localización diferente. Al mismo tiempo, el sistema de recogida de basura determina que una sección particular de almacenamiento se puede recuperar. Esta determinación puede ser debida a una falta de capacidad de almacenamiento disponible, un porcentaje de datos marcados como inválidos que alcanza un umbral, una consolidación de datos válidos, una tasa de detección de errores para esa sección de almacenamiento que alcanza un umbral o un mejor funcionamiento basado en la distribución de datos, etc. Se pueden considerar numerosos factores por el algoritmo de recogida de basura para determinar cuándo se recupera una sección de almacenamiento.

Una vez que una sección de almacenamiento se ha marcado para la recuperación, usualmente se deben relocalizar los paquetes válidos en la sección. La derivación del recogedor de basura 316 permite leer los paquetes dentro de la conducción de datos de lectura 108 y transferirlos a continuación directamente a la conducción de datos de escritura 106 sin encaminarse por el almacenamiento de estado sólido 104. En una realización preferida, la derivación de recogida de basura 316 es parte de un sistema autónomo de recogida de basura que opera dentro del dispositivo de almacenamiento de estado sólido 102. Esto permite al dispositivo de almacenamiento de estado sólido 102 gestionar los datos de modo que los datos se difunden sistemáticamente a través del almacenamiento de estado sólido 110 para mejorar el funcionamiento, la fiabilidad de los datos y para evitar la sobreutilización y la subutilización de cualquier localización o área del almacenamiento de estado sólido 110 y alargar la vida útil del almacenamiento de estado sólido 110.

La derivación del recogedor de basura 316 coordina la inserción de segmentos dentro de la conducción de datos de escritura 106 con otros segmentos distintos de los escritos por los clientes 114 u otros dispositivos. En la realización representada, la derivación de recogedor de basura 316 está antes del empaquetador 302 en la conducción de datos de escritura 106 y después del des-empaquetador 324, en la conducción de datos de lectura 108, pero también se puede localizar en otra parte en las conducciones de datos de lectura y escritura 106, 108. La derivación del recogedor de basura 316 se puede usar durante un traspaso de la conducción de escritura 106 para rellenar el resto de la página virtual para mejorar la eficacia de almacenamiento dentro del almacenamiento de estado sólido 110 y por lo tanto reducir la frecuencia de la recogida de basura.

En una realización, la conducción de datos de escritura 106 incluye una memoria intermedia de escritura 320 que almacena temporalmente datos para unas operaciones de escritura eficientes. Usualmente, la memoria intermedia de escritura 320 incluye suficiente capacidad para paquetes que rellenan al menos una página virtual en el almacenamiento de estado sólido 110. Esto permite a una operación de escritura enviar una página entera de datos al almacenamiento de estado sólido 110 sin interrupción. El dimensionamiento de la memoria intermedia de escritura 320 de la conducción de datos de escritura 106 y las memorias intermedias dentro de la conducción de datos de lectura 108 para que sean de la misma capacidad o mayor que la de la memoria intermedia de escritura de almacenamiento dentro del almacenamiento de estado sólido 110, permite que la escritura y lectura de datos sea más eficiente ya que se puede crear un único comando de escritura para enviar una página virtual completa de datos al almacenamiento de estado sólido 110 en lugar de múltiples comandos.

Mientras que la memoria intermedia de escritura 320 se está rellenando, el almacenamiento de estado sólido 110 se puede usar para otras operaciones de lectura. Esto es ventajoso porque otros dispositivos de estado sólido con menor memoria intermedia de escritura o sin memoria intermedia de escritura pueden ocupar el almacenamiento de estado sólido cuando se escriben los datos a una memoria intermedia de escritura de almacenamiento y los datos que fluyen dentro de la memoria intermedia de escritura de almacenamiento se paran. Las operaciones de lectura se bloquearán hasta que toda la memoria intermedia de escritura del almacenamiento se rellene y se programe. Otro enfoque para los sistemas sin una memoria intermedia de escritura o una memoria intermedia de escritura pequeña es traspasar la memoria intermedia de escritura de almacenamiento que no está llena para posibilitar las lecturas. De nuevo esto es ineficiente porque se requieren múltiples ciclos de escritura / programa para rellenar una página.

Para la realización representada con una memoria intermedia de escritura 320 dimensionada mayor que una página virtual, un único comando de escritura que incluye numerosos subcomandos, se puede seguir a continuación por un único comando de programa para transferir la página de datos desde la memoria intermedia de escritura de almacenamiento en cada uno de los elementos de almacenamiento estado sólido 216, 218, 220 a la página designada dentro de cada elemento de almacenamiento de estado sólido 216, 218, 220. Esta técnica tiene los beneficios de eliminar la programación parcial de la página, que es conocido que reduce la fiabilidad y durabilidad de los datos y liberar el banco de destino para lecturas y otros comandos mientras que se rellena la memoria intermedia.

En una realización, la memoria intermedia de escritura 320 es una memoria intermedia ping-pong donde un lado de la memoria intermedia se rellena y a continuación se designa para la transferencia en un momento apropiado mientras que el otro lado de la memoria intermedia de ping-pong se está rellenando. En otra realización, la memoria intermedia de escritura 320 incluye un primer registro del tipo primero en entrar - primero en salir ("FIFO") con una capacidad de más de una página virtual de segmentos de datos. Un experto en la materia reconocerá otras configuraciones de memoria intermedia de escritura 320 que permiten que una página virtual de datos se almacene antes de escribir los datos en el almacenamiento de estado sólido 110.

En otra realización, la memoria intermedia de escritura 320 se dimensiona más pequeña que una página virtual de modo que menos de una página de información se podría escribir a una memoria intermedia de escritura de almacenamiento en el almacenamiento de estado lógico 110. En la realización, para impedir una parada en la conducción de datos de escritura 106 a partir del mantenimiento de las operaciones de lectura, los datos se ponen en cola usando el sistema de recogida de basura que necesita moverse de una localización a otra como parte del proceso de recogida de basura. En el caso de una parada de datos en la conducción de datos de escritura 106, los datos se pueden alimentar a través de la derivación del recogedor de basura 316 a la memoria de escritura 320 y a continuación sobre la memoria intermedia de escritura de almacenamiento en el almacenamiento de estado sólido 110 para rellenar las páginas de una página virtual ante de programar los datos. De este modo una parada de datos en la conducción de datos de escritura 106 no pararía la lectura del dispositivo de almacenamiento de estado sólido 102.

En otra realización, la conducción de datos de escritura 106 incluye un módulo de programa de escritura 310 con una o más funciones definibles por el usuario dentro de la conducción de datos de escritura 106. El módulo de programa de escritura 310 permite a un usuario modelar la conducción de datos de escritura 106. Un usuario puede modelar la conducción de datos de escritura 106 en base al requisito o aplicación de datos particular. Donde el controlador de almacenamiento de estado sólido 104 es una FPGA, el usuario puede programar la conducción de datos de escritura 106 con comandos a medida y funciones de forma relativamente fácil. Un usuario también puede usar el módulo del programa de escritura 310 para incluir funciones a medida con un ASIC, sin embargo, la adaptación con un ASIC puede ser más difícil que con una FPGA. El módulo de programa de escritura 310 puede incluir memorias intermedias y mecanismos de deriva para permitir a un primer segmento de datos ejecutarse en el módulo del programa de escritura 310 mientras que el segundo segmento de datos puede continuar a través de la conducción de datos de escritura 106. En otra realización, el módulo de programa de escritura 310 puede incluir un núcleo de procesador que se puede programar mediante software.

Obsérvese que el módulo del programa de escritura 310 se muestra entre la memoria intermedia de entrada 306 y el módulo de compresión 312, sin embargo, el módulo del programa de escritura 310 podría estar en cualquier parte en la conducción de datos de escritura 106 y se puede distribuir entre las diversas etapas 302 - 320. Además, puede haber múltiples módulos de programa de escritura 310 distribuidos entre las diversas etapas 302 - 320 que se programan y operan independientemente. Además, el orden de las etapas 302 - 320 se puede alterar. Un experto en la materia reconocerá alteraciones realizables en el orden de las etapas 302 - 320 en base a requisitos de usuario particulares.

Conducción de datos de lectura

La conducción de datos de lectura 108 incluye un módulo de corrección de ECC 322 que determina si existe un error de datos en los bloques de ECC de un paquete solicitado recibido desde el almacenamiento de estado sólido 110 usando el ECC almacenado con cada uno de los bloques de ECC del paquete solicitado. El módulo de corrección de ECC 322 corrige a continuación cualesquiera errores en el paquete solicitado si existe cualquier error y los errores son corregibles usando el ECC. Por ejemplo, si el ECC puede detectar un error en seis bits pero solo puede corregir tres errores de bit, el módulo de corrección de ECC 322 corrige los bloques de ECC del paquete solicitado con hasta tres bits en error. El módulo de corrección de ECC 322 corrige los bits en error cambiando los bits en error al estado correcto de uno o cero de modo que el paquete solicitado es idéntico a cuando se escribió en el almacenamiento de estado sólido 110 y se generó el ECC para el paquete.

Si el módulo de corrección de ECC 322 determina que los paquetes solicitados contienen más bits en error de los que puede corregir el ECC, el módulo de corrección de ECC 322 no puede corregir los errores en los bloques de ECC corrompidos del paquete solicitado y envía una interrupción. En una realización, el módulo de corrección de ECC 322 envía una interrupción con un mensaje indicando que el paquete solicitado está en error. El mensaje puede incluir información de que el módulo de corrección de ECC no puede corregir los errores o la incapacidad del módulo de corrección de ECCC 322 para corregir los errores puede estar implicada. En otra realización, el módulo de corrección de ECC 322 envía los bloques de ECC corrompidos del paquete solicitado con la interrupción y/o el mensaje.

En la realización preferida, un bloque de ECC corrompido o porción de un bloque de ECC corrompido del paquete solicitado que no se puede corregir por el módulo de corrección de ECC 322 se lee por el controlador maestro 224, se corrige y se devuelve al módulo de corrección de ECC 322 para un procesamiento adicional por la conducción de datos de lectura 108. En una realización, un bloque de ECC corrompido o porción de un bloque de ECC corrompido

del paquete solicitado se envía al dispositivo solicitante de los datos. El dispositivo solicitante 155 puede corregir el bloque de ECC o reemplazar los datos usando otra copia, tal como una copia de respaldo o espejo, y a continuación puede usar los datos de reemplazo del paquete de datos solicitado o devolverlo a la conducción de datos de lectura 108. El dispositivo solicitante 155 puede usar la información de cabecera en el paquete solicitado en error para identificar los datos requeridos para reemplazar el paquete solicitado corrompido o reemplazar el objeto al que pertenece el paquete. En otra realización preferida, el controlador de almacenamiento de estado sólido 104 almacena los datos usando algún tipo de RAID y es capaz de recuperar los datos corrompidos. En otra realización, el módulo de corrección de ECC 322 envía una interrupción y/o mensaje y el dispositivo receptor falla en la operación de lectura asociada con el paquete de datos solicitado. Un experto en la materia reconocerá otras opciones y acciones a tomar como resultado de que el módulo de corrección de ECC 322 determine que uno o más bloques de ECC del paquete solicitado están corrompidos y que el módulo de corrección de ECC 322 no puede corregir los errores.

La conducción de datos de lectura 108 incluye un des-empaquetador 324 que recibe los bloques de ECC del paquete solicitado desde el módulo de corrección de ECC 322, directamente o indirectamente y comprueba y elimina una o más cabeceras del paquete. El des-empaquetador 324 puede validar las cabeceras de paquete comprobando los identificadores de paquete, la longitud de los datos, la localización de los datos, etc. dentro de las cabeceras. En una realización, la cabecera incluye un código de huella digital que se puede usar para validar que el paquete entregado a la conducción de datos de lectura 108 es el paquete solicitado. El des-empaquetador 324 también elimina las cabeceras del paquete solicitado añadidas por el empaquetador 302. El des-empaquetador 324 se puede dirigir para que no opere en ciertos paquetes sino que pase estos hacia delante sin modificación. Un ejemplo podría ser una etiqueta de un contenedor que se solicita durante el curso de un proceso de reconstrucción donde se requiere la información de la cabecera por el módulo de reconstrucción de índices de objetos 272. Ejemplos adicionales incluyen la transferencia de paquetes de diversos tipos destinados para uso dentro del dispositivo de almacenamiento de estado sólido 102. En otra realización, la operación del des-empaquetador 324 puede ser dependiente del tipo de paquete.

La conducción de datos de lectura 108 incluye un módulo de alineamiento 326 que recibe datos desde el des-empaquetador 324 y elimina los datos no deseados. En una realización, un comendo de lectura enviado al almacenamiento de estado sólido 110 recupera un paquete de datos. Un dispositivo que solicita datos puede no requerir todos los datos dentro del paquete recuperado y el módulo de alineamiento 326 elimina los datos no deseados. Si se solicitan todos los datos dentro de una página recuperada, el módulo de alineamiento 326 no elimina ningún dato.

El módulo de alineamiento 326 re-formatea los datos o segmentos de datos de un objeto en una forma compatible con el dispositivo que solicita el segmento de datos antes de redirigir el segmento de datos a la siguiente etapa. Usualmente, como los datos se procesan por la conducción de datos de lectura 108, el tamaño de los segmentos de datos o paquetes cambia en las diversas etapas. El módulo de alineamiento 326 usa los datos recibidos para formatear los datos dentro de los segmentos de datos adecuados para enviar al dispositivo solicitante 155 y se unen para formar una respuesta. Por ejemplo, los datos desde una porción de un primer paquete de datos se pueden combinar con los datos desde una porción de un segundo paquete de datos. Si un segmento de datos es mayor que un dato solicitado por el dispositivo solicitante 155, el módulo de alineamiento 326 puede descartar los datos no deseados.

En una realización, la conducción de datos de lectura 108 incluye una memoria intermedia de sincronización de lectura 328 que almacena temporalmente uno o más paquetes solicitados leídos desde el almacenamiento de estado sólido 110 antes del procesamiento por la conducción de datos de lectura 108. La memoria intermedia de sincronización de lectura 328 está en el límite entre el dominio de reloj del almacenamiento de estado sólido y el dominio de reloj del bus local y proporciona el almacenamiento temporal para dar cuenta de las diferencias de los dominios de reloj.

En otra realización, la conducción de datos de lectura 108 incluye una memoria intermedia de salida 330 que recibe los paquetes solicitados desde el módulo de alineamiento 326 y almacena los paquetes antes de su transmisión al dispositivo solicitante. La memoria intermedia de salida 330 da cuenta de las diferencias entre cuando se reciben los segmentos de datos desde las etapas de la conducción de datos de lectura 108 y cuando se transmiten los segmentos de datos a otras partes del controlador de almacenamiento de estado sólido 104 o al dispositivo solicitante 155. La memoria intermedia de salida 330 también permite que el bus de datos 204 reciba datos desde la conducción de datos de lectura 108 a tasas mayores que las que se pueden sostener por la conducción de datos de lectura 108 para mejorar la eficacia de operación del bus de datos 204.

En una realización, la conducción de datos de lectura 108 incluye un módulo de descifrado de medios 332 que recibe uno o más paquetes solicitados cifrados desde el módulo de corrección de ECC 322 y descifra el uno o más paquetes solicitados usando una clave de cifrado única para el dispositivo de almacenamiento de estado sólido 102 antes de enviar el uno o más paquetes solicitados al des-empaquetador 324. Usualmente la clave de cifrado usada para descifrar los datos por el módulo de descifrado de medios 332 es idéntica a la clave de cifrado usada por el módulo de cifrado de medios 318. En otra realización, el almacenamiento de estado sólido 110 puede tener dos o

más particiones y el controlador de almacenamiento de estado sólido 104 se comporta como si fuesen dos o más controladores de almacenamiento de estado sólido 104 operando cada uno sobre una partición única dentro del almacenamiento de estado sólido 110. En esta realización, se puede usar una clave de cifrado de medios única con cada partición.

En otra realización, la conducción de datos de lectura 108 incluye un módulo de descifrado 334 que descifra un segmento de datos formateado por el des-empaquetador 324 antes de enviar el segmento de datos a la memoria intermedia de salida 330, usando el segmento de datos descifrado una clave de cifrado recibida en conjunción con la petición de lectura que inicia la recuperación del paquete solicitado recibido por la memoria intermedia de sincronización de lectura 328. El módulo de descifrado 334 puede descifrar un primer paquete con una clave de cifrado recibida en conjunción con la petición de lectura para el primer paquete y a continuación puede descifrar un segundo paquete con una clave de cifrado diferente o puede pasar el segundo paquete sobre la siguiente etapa de la conducción de datos de lectura 108 sin descifrado. Usualmente, el módulo de descifrado 334 usa una clave de cifrado diferente para descifrar un segmento de datos que la que usa el módulo de descifrado de medios 332 para descifrar los paquetes solicitados. Cuando se almacenó el paquete con el número aleatorio de cifrado no secreto, el número aleatorio se usa en conjunción con una clave de cifrado para descifrar el paquete de datos. La clave de cifrado se puede recibir desde un cliente 114, un ordenador 112, un gestor de claves, u otro dispositivo que gestione la clave de cifrado a usar por el controlador de almacenamiento de estado sólido 104.

En otra realización la conducción de datos de lectura 108 incluye un módulo de descompresión 336 que descomprime un segmento de datos formateado por el des-empaquetador 324. En la realización preferida, el módulo de descompresión 336 usa la información de compresión almacenada en uno o ambos de la cabecera del paquete y la etiqueta del contenedor para seleccionar una rutina complementaria que se usó para comprimir los datos por el módulo de compresión 312. En otra realización, la rutina de descompresión usada por el dispositivo de descompresión 336 se dicta por el dispositivo que solicita el segmento de datos que se va a descomprimir. En otra realización, el módulo de descompresión 336 selecciona una rutina de descompresión de acuerdo con la configuración por defecto en base al tipo de objeto o a la clase del objeto. Un primer paquete de un primer objeto puede ser capaz de anular una rutina de descompresión por defecto y un segundo paquete de un segundo objeto de la misma clase de objeto y el mismo tipo de objeto puede usar la rutina de descompresión por defecto y un tercer paquete de un tercer objeto de la misma clase de objeto y tipo de objeto puede no usar ninguna descompresión.

En otra realización, la conducción de datos de lectura 108 incluye un módulo de programa de lectura 338 que incluye una o más funciones definibles por el usuario dentro de la conducción de datos de lectura 108. El módulo de programa de lectura 338 tiene similares características al módulo de programa de escritura 310 y permite a un usuario proporcionar funciones a medida para la conducción de datos de lectura 108. El módulo de programa de lectura 338 puede estar localizado como se muestra en la Figura 3, puede estar localizado en otra posición dentro de la conducción de datos de lectura 108, o puede incluir múltiples partes en múltiples localizaciones dentro de la conducción de datos de lectura 108. Adicionalmente, puede haber múltiples módulos de programa de lectura 338 dentro de múltiples localizaciones dentro de la conducción de datos de lectura 108 que operan independientemente. Un experto en la materia reconocerá otras formas de un módulo de programa de lectura 338 dentro de una conducción de datos de lectura 108. Como con la conducción de datos de escritura 106, las etapas de la conducción de datos de lectura 108 se pueden re-disponer y un experto en la materia reconocerá otros órdenes de las etapas dentro de la conducción de datos de lectura 108.

El controlador de almacenamiento de estado sólido 104 incluye registros de control y estado 340 y las colas de control correspondientes 342. Los registros de control y estatus 340 y las colas de control 342 facilitan el control y los comandos de secuenciación y subcomandos asociados con los datos procesados en las conducciones de datos de escritura y lectura 106, 108. Por ejemplo, un segmento de datos en el empaquetador 302 puede tener uno o más comandos de control correspondientes o instrucciones en una cola de control 342 asociadas con el generador de ECC 304. A medida que el segmento de datos se empaqueta se pueden ejecutar algunas de las instrucciones o comandos dentro del empaquetador 302. Otros comandos o instrucciones se pueden pasar a la siguiente cola de control 342 a través de registros de control y estado 340 a medida que los paquetes de datos formados de nuevo creados desde el segmento de datos se pasan a la siguiente etapa.

Los comandos e instrucciones se pueden cargar simultáneamente dentro de las colas de control 342 para un paquete que se redirige a la conducción de datos de escritura 106 extrayendo cada etapa de la conducción el comando o instrucción apropiada a medida que el paquete respectivo se ejecuta por esa etapa. De forma similar, los comandos y las instrucciones se pueden cargar simultáneamente dentro de las colas de control 342 para un paquete que se está contestando desde la conducción de datos de lectura 108 extrayendo cada etapa de la conducción el comando o instrucción apropiada a medida que se ejecuta el paquete respectivo por esa etapa. Un experto en la materia reconocerá otras características y funciones de los registros de control y estado 340 y las colas de control 342.

El controlador de almacenamiento de estado sólido 104 y/o el dispositivo de almacenamiento de estado sólido 102 pueden incluir también un controlador de intercalado de bancos 344, una memoria intermedia de sincronización 346,

un controlador del bus de almacenamiento 348, un multiplexor ("MUX") 350, que se describen en relación con las Figuras 4A y 4B.

Intercalado de bancos

La Figura 4A es un diagrama de bloques esquemático que ilustra una realización 400 de un controlador de intercalado de bancos 344 en el controlador de almacenamiento de estado sólido 104 de acuerdo con la presente invención. El controlador de intercalado de bancos 344 se conecta a los registros de control y estado 340 y al bus I/O de almacenamiento 210 y el bus de control de almacenamiento 212 a través del MUX 350, el controlador del bus de almacenamiento 348 y la memoria intermedia de sincronización 346 que se describen a continuación. El controlador de intercalado de bancos 344 incluye un agente de lectura 402, un agente de escritura 404, un agente de borrado 406 un agente de gestión 408, colas de lectura 410a - n, colas de escritura 412a - n, colas de borrado 414a - n y colas de gestión 416a - n para los bancos 214 en el almacenamiento de estado sólido 110, los controladores de bancos 418a - n, un árbitro de bus 420, y un MUX de estados 422, que se describen a continuación. El controlador del bus de almacenamiento 348 incluye un módulo de mapeo 424 con un módulo de re-mapeo 430, un módulo de captura de estatus 426, y un controlador de bus NAND 428, que se describen a continuación.

El controlador de intercalado de bancos 344 dirige uno o más comandos a una o más colas en el controlador de intercalado de bancos 344 y coordina entre los bancos 214 del almacenamiento de estado sólido 110, la ejecución de los comandos almacenados en las colas, de modo que un comando de un primer tipo se ejecuta sobre un banco 214a mientras que un comando de un segundo tipo se ejecuta sobre un segundo banco 214b. El uno o más comandos están separados por tipos de comandos dentro de las colas. Cada banco 214 del almacenamiento de estado sólido 110 tiene un conjunto correspondiente de colas dentro del controlador de intercalado de bancos 344 y cada conjunto de colas incluye una cola para cada tipo de comando.

El controlador de intercalado de bancos 344 coordina entre los bancos 214 del almacenamiento de estado sólido 110 la ejecución de los comandos almacenados en las colas. Por ejemplo, un comando de un primer tipo se ejecuta sobre un banco 214a mientras que un comando de un segundo tipo se ejecuta sobre un segundo banco 214b. Usualmente, los tipos de comandos y los tipos de colas incluyen comandos de lectura y escritura y colas 410, 412, pero también pueden incluir otros comandos y colas que son específicas del medio de almacenamiento. Por ejemplo, en la realización representada en la Figura 4A las colas de borrado y gestión 414 y 416 están incluidas y serían apropiadas para memoria flash, NRAM, MRAM, DRAM, PRAM, etc.

Para otros tipos de almacenamiento de estado sólido 110, se pueden incluir otros tipos de comandos y las colas correspondientes sin desviarse del ámbito de la invención. La naturaleza flexible del controlador de almacenamiento de estado sólido de FPGA 104 permite la flexibilidad en los medios de almacenamiento. Si se cambiase la memoria flash a otro tipo de almacenamiento de estado sólido, el controlador de intercalado de bancos 344, el controlador del bus de almacenamiento 348 y el MUX 350 se podrían alterar para acomodarse al tipo de medios sin afectar significativamente a las conducciones de datos 106, 108 y otras funciones del controlador de almacenamiento de estado sólido 104.

En la realización representada en la Figura 4A, el controlador de intercalado de bancos 344 incluye, para cada banco 214; una cola de lectura 410 para la lectura de datos desde el almacenamiento de estado sólido 110, una cola de escritura 412 para escritura de comandos al almacenamiento de estado sólido 110, una cola de borrado 414 para borrar un bloque de borrado en el almacenamiento de estado sólido, una cola de gestión 416 para los comandos de gestión. El controlador de intercalado de bancos 344 también incluye los agentes de lectura, escritura, borrado y gestión correspondientes 402, 404, 406, 408. En otra realización, los registros de control y estado 340 y las colas de control 342 o componentes similares ponen en cola comandos para los datos enviados a los bancos 214 del almacenamiento de estado sólido 110 sin un controlador de intercalado de bancos 344.

Los agentes 402, 404, 406, 408, en una realización, dirigen comandos del tipo apropiado destinados a un banco particular 214a a la cola correcta para el banco 214a. Por ejemplo, el agente de lectura 402 puede recibir un comando de lectura para el banco 1 214b y dirigir un comando de lectura a la cola de lectura del al banco 1 410b. El agente de escritura 404 puede recibir un comando de escritura para escribir datos en una localización en el banco 0 214a del almacenamiento de estado sólido 110 y a continuación enviar el comando de escritura a la cola de escritura del banco 0 412a. De forma similar, el agente de borrado 406 puede recibir un comando de borrado para borrar un bloque de borrado en el banco 1 214b y a continuación pasar el comando de borrado a la cola de borrado del banco 1 414b. El agente de gestión 408 usualmente recibe comandos de gestión, peticiones de estado, y similares, tales como un comando de reinicio, o una petición de lectura de un registro de configuración de un banco 214, tal como un banco 0 214a. El agente de gestión 408 envía el comando de gestión a la cola de gestión del banco 0 416a.

Los agentes 402, 404, 406, 408 usualmente también monitorizan el estado de las colas 410, 412, 414, 416 y envían el estado, interrupciones y otros mensajes cuando las colas 410, 412, 414, 416 están llenas, casi llenas, no funcionales, etc. En una realización, los agentes 402, 404, 406, 408 reciben comandos y generan los subcomandos correspondientes. En una realización, los agentes 402, 404, 406, 408 reciben comandos a través de los registros de

control y estado 340 y generan los subcomandos correspondientes que se redirigen a las colas 410, 412, 414, 416. Un experto en la materia reconocerá otras funciones de los agentes 402, 404, 406, 408.

Las colas 410, 412, 414, 416 usualmente reciben comandos y almacenan los comandos hasta que se requieren para enviarlos a los bancos de almacenamiento de estado sólido 214. En una realización típica, las colas 410, 412, 414, 416 son registros del tipo primero en entrar primero en salir ("FIFO") o un componente similar que opera como un FIFO. En otra realización las colas 410, 412, 414, 416 almacenan comandos en un orden que compagina datos, orden de importancia, u otros criterios.

Los controladores de bancos 418 usualmente reciben comandos desde las colas 410, 412, 414, 416 y generan los subcomandos apropiados. Por ejemplo, la cola de escritura del banco 0 412a puede recibir un comando para escribir una página de paquetes de datos al banco 0 214a. El controlador del banco 0 418a puede recibir el comando de escritura en un momento apropiado y puede generar uno o más subcomandos de escritura para cada paquete de datos almacenado en la memoria intermedia de escritura 320 a escribir a la página en el banco 0 214a. Por ejemplo, el controlador del banco 0 418a puede generar comandos para validar el estatus del banco 0 214a y la red de almacenamiento de estado sólido 216, seleccionar la localización apropiada para escribir uno o más paquetes de datos, borrar las memorias intermedias de entrada dentro de la red de memoria de almacenamiento de estado sólido 216, transferir el uno o más paquetes de datos a las memorias intermedias de entrada, programar las memorias intermedias de entrada dentro de la localización seleccionada, verificar que los datos se programaron correctamente, y si ocurren fallos del programa hacer una o más interrupciones del controlador maestro 224 reintentando la escritura a la misma localización física y reintentando la escritura a una localización física diferente. Adicionalmente, en conjunción con el comando de escritura de ejemplo, el controlador del bus de almacenamiento 348 causará que el uno o más comandos para multiplicar a cada uno de los buses I/O de almacenamiento 210a - n con la dirección lógica del comando mapeado a la primera dirección física para el bus I/O de almacenamiento 210a, y mapeado a la segunda dirección física para el bus I/O de almacenamiento 210b, y así sucesivamente como se describe adicionalmente más adelante.

Usualmente, el árbitro de bus 420 selecciona de entre los controladores de bancos 418 y extrae los subcomandos desde las colas de salida dentro de los controladores de bancos 418 y los redirige al Controlador del Bus de Almacenamiento 348 en una secuencia que optimiza el funcionamiento de los bancos 214. En otra realización, el árbitro de bus 420 puede responder a una interrupción de nivel alto y modificar el criterio de selección normal. En otra realización, el controlador maestro 224 puede controlar el árbitro de bus 420 a través de los registros de control y estado 340. Un experto en la técnica reconocerá otros medios por los que el árbitro de bus 420 puede controlar e intercalar la secuencia de comandos desde los controladores de bancos 418 al almacenamiento de estado sólido 110.

El árbitro de bus 420 usualmente coordina la selección de comandos apropiados, y los datos correspondientes cuando se requiere por el tipo de comando, desde los controladores de bancos 418 y envía los comandos y los datos al controlador del bus de almacenamiento 348. El árbitro de bus 420 usualmente también envía comandos al bus de control de almacenamiento 212 para seleccionar el banco apropiado 214. Para el caso de memoria flash u otro almacenamiento de estado sólido 110 con un bus I/O de almacenamiento serie bidireccional asíncrono 210, solo un comando (información de control) o conjunto de datos se puede transmitir cada vez. Por ejemplo, cuando se está transmitiendo un comando de escritura o datos al almacenamiento de estado sólido 110 sobre el bus I/O de almacenamiento 210, los comandos de lectura, la lectura de datos, los comandos de borrado, los comandos de gestión u otros comandos de estado no se pueden transmitir sobre el bus I/O de almacenamiento 210. Por ejemplo, cuando se están leyendo los datos desde el bus I/O de almacenamiento 210, no se pueden escribir datos al almacenamiento de estado sólido 110.

Por ejemplo, durante una operación de escritura sobre el banco 0, el árbitro de bus 420 selecciona el controlador del banco 0 418a que puede tener un comando de escritura o una serie de subcomandos de escritura en la parte superior de su cola lo que causa que el controlador del bus de almacenamiento 348 ejecute la siguiente secuencia. El árbitro de bus 420 redirige el comando de escritura al controlador de bus del almacenamiento 348, que establece un comando de escritura seleccionando el banco 0 214a a través del bus de control de almacenamiento 212, enviando un comando para borrar las memorias intermedias de entrada de los elementos de almacenamiento de estado sólido 110 asociadas con el banco 0 214a, y enviando un comando para validar el estado de los elementos de almacenamiento de estado sólido 216, 218, 220 asociados con el banco 0 214a. El controlador del bus de almacenamiento 348 transmite a continuación un subcomando de escritura sobre el bus I/O de almacenamiento 210, que contiene las direcciones físicas incluyendo la dirección lógica del bloque de borrado para cada elemento de almacenamiento de estado sólido de borrado físico individual 216a - m ya que se mapean desde la dirección lógica del bloque de borrado. El controlador del bus de almacenamiento 348 multiplexa a continuación la memoria intermedia de escritura 320 a través de la memoria intermedia síncrona de escritura 308 para el bus I/O de almacenamiento 210 a través del MUX 350 y hace fluir los datos de escritura a la página apropiada. A continuación, cuando la página está llena el controlador del bus de almacenamiento 348 causa que los elementos de almacenamiento de estado sólido 216a - m asociados con el banco 0 214a programen la memoria intermedia de entrada a las células de memoria dentro de los elementos de almacenamiento de estado sólido 216a - m.

Finalmente, el controlador del bus de almacenamiento 348 valida el estado para asegurar que la página se programó correctamente.
Una operación de lectura es similar al ejemplo de escritura anterior. Durante una operación de lectura, usualmente el árbitro del bus 420, u otro componente del controlador de intercalado de bancos 344, recibe los datos y la información de estado correspondiente y envía los datos a la conducción de datos de lectura 108 mientras que envía la información de estado sobre los registros de control y estado 340. Usualmente, un comando de lectura de datos redirigido desde el árbitro de bus 420 al controlador de bus de almacenamiento 348 causará que el MUX 350 controle la entrada de los datos de lectura sobre el bus I/O de almacenamiento 210 a la conducción de datos de lectura 108 y envíe información de estado a los registros apropiados de control y estado 340 a través del MUX de estados 422.

El árbitro de bus 420 coordina los diversos tipos de comandos y modos de acceso de datos de modo que solo un tipo de comando apropiado o los datos correspondientes están sobre el bus en un momento determinado. Si el árbitro de bus 420 ha seleccionado un comando de escritura, los subcomandos de escritura y los datos correspondientes se escribirán al almacenamiento de estado sólido 110, el árbitro de bus 420 no permitirá otros tipos de comandos sobre el bus I/O de almacenamiento 210. Ventajosamente, el árbitro de bus 420 usa la información de temporización, tal como los tiempos de ejecución previsibles de los comandos, junto con la información de estado recibida concerniente al estado del banco 214 para coordinar la ejecución de diversos comandos sobre el bus con el objetivo de minimizar o eliminar el tiempo de reposo de los buses.

El controlador maestro 224 a través del árbitro de bus 420 usualmente usa los tiempos de terminación esperados de los comandos almacenados en las colas 410, 42, 414, 416, junto con la información de estado, de modo que cuando los subcomandos asociados con un comando se ejecutan sobre un banco 214a. otros subcomandos de otros comandos se ejecutan sobre otros bancos 214b - n. Cuando un comando se ejecuta completamente sobre el banco 214a, el árbitro de bus 420 dirige otro comando al banco 214a. El árbitro de bus 420 puede coordinar también los comandos almacenados en las colas 410, 412, 414, 416 con otros comandos que no están almacenados en las colas 410, 412, 414, 416.

Por ejemplo, se puede enviar un comando de borrado para borrar un grupo de bloques de borrado dentro del almacenamiento de estado sólido 110. Un comando de borrado puede tardar de 10 a 1000 veces más de tiempo en ejecutarse que un comando de escritura o un comando de lectura o de 10 a 100 veces más de tiempo para ejecutar un comando de programa. Para N bancos 214, el controlador de intercalado de bancos 344 puede dividir el comando de borrado en N comandos, cada uno para borrar un bloque de borrado virtual de un banco 214a. Mientras que el banco 0 214a está ejecutando un comando de borrado, el árbitro de bus 420 puede seleccionar otros comandos para su ejecución sobre los otros bancos 214b - n. El árbitro de bus 420 también puede funcionar con otros componentes, tales como el controlador de bus de almacenamiento 348, el controlador maestro 224, etc. para coordinar la ejecución de comandos entre los buses. La coordinación en la ejecución de los comandos usando el árbitro de bus 420, los controladores de bancos 418, las colas 410, 412, 414, 416 y los agentes 402, 404, 406, 408 del controlador de intercalado de bancos 344 pueden aumentar drásticamente el rendimiento sobre otros sistemas de almacenamiento de estado sólido sin una función de intercalado de bancos.

En una realización, el controlador de estado sólido 104 incluye un controlador de intercalado de bancos 344 que sirve a todos los elementos de almacenamiento 216, 218, 220 del almacenamiento de estado sólido 110. En otra realización, el controlador de estado sólido 104 incluye un controlador de intercalado de bancos 344 para cada fila de elementos de almacenamiento 216a - m, 218a - m, 220a - m. Por ejemplo, un controlador de intercalado de bancos 344 sirve a una fila de elementos de almacenamiento SSS 0.0 - SSS 0.N 216a, 218a, 220a, un segundo controlador de intercalado de bancos 344 sirve a una segunda fila de elementos de almacenamiento SSS 1.0 - SSS 1.N 216b, 218b, 220b etc.

La Figura 4B es un diagrama de bloques esquemático que ilustra una realización alternativa 401 de un controlador de intercalado de bancos 344 en el controlador de almacenamiento de estado sólido 104 de acuerdo con la presente invención. Los componentes 210, 212, 340, 346, 348, 350, 402 - 430 representados en la realización mostrada en la Figura 4B son sustancialmente similares al aparato de intercalado de bancos 400 descrito en relación con la Figura 4A excepto que el cada banco 214 incluye una cola única 432a - n y los comandos de lectura, los comandos de escritura, los comandos de borrado, los comandos de gestión, etc. para un banco (por ejemplo el banco 0 214a) se dirigen a una cola única.432a para el banco 214a. Las colas 432, en una realización son FIFO. En otra realización, las colas 432 pueden tener comandos extraídos desde las colas 432 en un orden distinto que el orden en el que se almacenaron. En otra realización alternativa (no mostrada), el agente de lectura 402, el agente de escritura 404, el agente de borrado 406, y el agente de gestión 408 se pueden combinar en un único agente que asigna comandos a las colas apropiadas 432a - n.

En otra realización alternativa (no mostrada) los comandos se almacenan en una cola única donde los comandos se pueden extraer de la cola en un orden distinto a como se almacenaron de modo que el controlador de intercalado de bancos 344 puede ejecutar un comando sobre un banco 214a mientras que otros comandos se ejecutan en los bancos restantes 214b - n. Un experto en la materia reconocerá fácilmente otras configuraciones de colas y tipos

para posibilitar la ejecución de un comando sobre un banco 214a mientras que otros comandos se ejecutan sobre otros bancos 214b - n.

Componentes específicos de almacenamiento

El controlador de almacenamiento de estado sólido 104 incluye una memoria intermedia de sincronización 346 que memoriza temporalmente los comandos y los mensajes de estado enviados hacia y recibidos desde el almacenamiento de estado sólido 110. La memoria intermedia de sincronización 346 se localiza en una frontera entre el domino del reloj del almacenamiento de estado sólido y el domino del reloj del bus local y proporciona el almacenamiento temporal para dar cuenta de las diferencias en el dominio del reloj. La memoria intermedia de sincronización 346, la memoria intermedia de sincronización de escritura 308, y la memoria intermedia de sincronización de lectura 328 pueden ser independientes o pueden actuar juntas para almacenar temporalmente los datos, comandos, mensajes de estado, etc. En la realización preferida, la memoria intermedia de sincronización 346 está localizada donde hay el menor número de señales que cruzan los dominios de reloj. Un experto en la materia reconocerá que la sincronización entre los dominios de reloj se puede mover arbitrariamente a otras localizaciones dentro del dispositivo de almacenamiento de estado sólido 102 para optimizar algún aspecto de la implementación del diseño.

El controlador de almacenamiento de estado sólido 104 incluye un controlador del bus de almacenamiento 348 que interpreta y traduce los comandos para los datos enviados y leídos desde el almacenamiento de estado sólido 110 y los mensajes de estado recibidos desde el almacenamiento de estado sólido 110 en base al tipo de almacenamiento de estado sólido 110. Por ejemplo, el controlador del bus de almacenamiento 348 puede tener diferentes requisitos de temporización para los diferentes tipos de almacenamiento, almacenamientos con diferentes características de funcionamiento, almacenamientos de diferentes fabricantes, etc. El controlador del bus de almacenamiento 348 también envía comandos de control al bus de control de almacenamiento 212.

En la realización preferida, el controlador de almacenamiento de estado sólido 104 incluye un MUX 350 que comprende una red de multiplexores 350a - n donde cada multiplexor está dedicado a una fila en la red de almacenamiento de estado sólido 110. Por ejemplo, un multiplexor 350a está asociado con los elementos de almacenamiento de estado sólido 216a, 218a, 220a. El MUX 350 encamina los datos desde la conducción de datos de escritura 106 y los comandos desde el controlador del bus de almacenamiento 348 al almacenamiento de estado sólido 110 a través del bus I/O de almacenamiento 210 y encamina los datos y los mensajes de estado desde el almacenamiento de estado sólido 110 a través del bus I/O de almacenamiento 210 a la conducción de datos de lectura 108 y los registros de control y estado 340 a través del controlador del bus de almacenamiento 348, la memoria intermedia de sincronización 346, y el controlador de intercalado de bancos 344.

En la realización preferida, el controlador de almacenamiento de estado sólido 104 incluye un MUX 350 para cada fila de elementos de almacenamiento de estado sólido (por ejemplo SSS 0.1 216a, SSS 0.2 218a, SSS 0.N 220a). Un MUX 350 combina los datos desde la conducción de datos de escritura 106 y los comandos enviados al almacenamiento de estado sólido 110 a través del bus I/O de almacenamiento 210 y separa los datos a procesar por la conducción de datos de lectura 108 de los comandos. Los paquetes almacenados en la memoria intermedia de escritura 320 se dirigen sobre buses de la memoria intermedia de escritura 320 a través de la memoria intermedia de sincronización de escritura 308 para cada fila de elementos de almacenamiento de estado sólido (SSS x.0 a SSS x.N 216, 218, 220) al MUX 350 para cada fila de elementos de almacenamiento de estado sólido (SSS x.0 a SSS x.N 216, 218, 220). Los comandos y los datos de lectura se reciben por los MUX 350 desde el bus I/O de almacenamiento 210. Los MUX 350 también dirigen mensajes de estado al controlador del bus de almacenamiento 348.

El controlador del bus de almacenamiento 348 incluye un módulo de mapeo 424. El módulo de mapeo 424 mapea una dirección lógica de un bloque de borrado a una o más direcciones físicas de un bloque de borrado. Por ejemplo, un almacenamiento de estado sólido 110 con una red de veinte elementos de almacenamiento (por ejemplo SSS 0.0 a SSS M.0 216) por bloque 214a pueden tener una dirección lógica para un bloque de borrado particular mapeado a veinte direcciones físicas por elemento de borrado. Debido a que los elementos de almacenamiento se acceden en paralelo, los bloques de borrado en la misma posición en cada elemento de almacenamiento en una fila de elementos de almacenamiento 216a, 218a, 220a compartirán una dirección física. Para seleccionar un bloque de borrado (por ejemplo en el elemento de almacenamiento SSS 0.0 216a) en lugar de todos los bloques de borrado en la fila (por ejemplo en los elementos de almacenamiento SSS 0.0, 0,1, … 0..N 216a, 218a, 220a), se selecciona un banco (en este caso el banco 0 214a).

Este mapeo de lógico a físico para los bloques de borrado para borrar bloques es beneficioso debido a que si un bloque de borrado resulta dañado o inaccesible, se puede cambiar el mapeo para mapear a otro bloque de borrado. Esto mitiga el perjuicio de perder un bloque de borrado virtual entero cuando un bloque de borrado del elemento falla. El módulo de re-mapeo 430 cambia un mapeo de una dirección lógica de un bloque de borrado a uno o más direcciones físicas de un bloque de borrado virtual (difundido sobre la red de elementos de almacenamiento). Por ejemplo, el bloque de borrado virtual 1 se puede mapear para borrar el bloque 1 del elemento de almacenamiento SSS 0.0 216a, para borrar el bloque 1 del elemento de almacenamiento SSS 1.0 216b …, y el elemento de

almacenamiento M.0 216m, el bloque de borrado virtual 2 se puede mapear al bloque de borrado 2 del elemento de almacenamiento SSS 0.1 218a, para borrar el bloque 2 del elemento de almacenamiento SSS 1.1 218b … y para el elemento de almacenamiento M.1 218m, etc.

Si el bloque de borrado 1 de un elemento de almacenamiento SSS 0.0 216a está dañado, experimentando errores debido al uso, etc. o no se puede usar por alguna razón, el módulo de re-mapeo 430 podría cambiar el mapeo de lógico a físico para la dirección lógica que apunta al bloque de borrado 1 del bloque de borrado virtual 1. Si un bloque de borrado de repuesto (sea el bloque de borrado 221) del elemento de almacenamiento SSS 0.0 216a está disponible y actualmente no mapeado, el módulo de re-mapeo 430 podría cambiar el mapeo del bloque de borrado virtual 1 para apuntar al bloque de borrado 221 del elemento de almacenamiento SSS 0.0 216a, mientras que continúa apuntado al bloque de borrado 1 del elemento de almacenamiento SSS 1.0 216b, el bloque de borrado 1 del elemento de almacenamiento SSS 2.0 (no mostrado)…, y al elemento de almacenamiento M. 0.216m. El módulo de mapeo 424 o el módulo de re-mapeo 430 podrían mapear bloques de borrado en un orden prescrito (el bloque de borrado virtual 1 al bloque de borrado 1 de los elementos de almacenamiento, el bloque de borrado virtual 2 al bloque de borrado 2 del elemento de almacenamiento, etc.) o puede mapear bloques de borrado de los elementos de almacenamiento 216, 218, 220 en otro orden en base a algunos otros criterios.

En una realización, los bloques de borrado se podrían agrupar por tiempo de acceso. Significando el agrupamiento por tiempo de acceso que el tiempo de ejecutar un comando, tal como la programación de datos (escritura) en páginas de bloques de borrado específicas, pueden nivelar la terminación de comandos de modo que un comando ejecutado a través de bloques de borrado de un bloque de borrado virtual no esté limitado por el bloque de borrado más lento. En otras realizaciones, los bloques de borrado se pueden agrupar por el nivel de uso, salud, etc. Un experto en la técnica reconocerá otros factores a considerar con el mapeo o re-mapeo de los bloques de borrado.

En una realización, el controlador del bus de almacenamiento 348 incluye un módulo de captura de estados 426 que recibe mensajes de estado desde el almacenamiento de estado sólido 110 y envía los mensajes de estado al MUX de estados 422. En otra realización, cuando el almacenamiento de estado sólido 110 es una memoria flash, el controlador del bus de almacenamiento 348 incluye un controlador de bus NAND 428. El controlador de bus NAND 428 dirige los comandos desde las conducciones de datos de lectura y escritura 106, 108 a la localización correcta en el almacenamiento de estado sólido 110, coordina la temporización de la ejecución de comandos en base a las características de la memoria flash, etc. si el almacenamiento de estado sólido 110 es otro tipo de almacenamiento de estado sólido, el controlador de bus NAND 428 se reemplazaría por un controlador de bus específico para el tipo de almacenamiento. Un experto en la materia reconocerá otras funciones de un controlador de bus NAND 428.

Diagramas de flujo

La Figura 5A es un diagrama de flujo esquemático que ilustra una realización de un método 500 para gestionar los datos en un dispositivo de almacenamiento de estado sólido 102 que usa una conducción de datos de acuerdo con la presente invención. El método 500 comienza en 502 y la memoria intermedia de entrada 306 recibe 504 uno o más segmentos de datos a escribir al almacenamiento de estado sólido 110. El uno o más segmentos de datos usualmente incluyen al menos una porción de un objeto pero puede ser un objeto entero. El empaquetador 302 puede crear uno o más paquetes específicos del objeto en conjunción con un objeto. El empaquetador 302 añade una cabecera a cada paquete que usualmente incluye la longitud del paquete y un número de secuencia para el paquete dentro del objeto. El empaquetador 302 recibe 504 el uno o más segmentos de datos o metadatos que se almacenaron en la memoria intermedia de entrada 306 y empaqueta 506 el uno o más segmentos de datos o metadatos creando uno o más paquetes dimensionados para el almacenamiento de estado sólido 110 donde cada paquete incluye una cabecera y datos desde el uno o más segmentos.

Usualmente, un primer paquete incluye un identificador de objetos que identifica el objeto para el cual se creó el paquete. Un segundo paquete puede incluir una cabecera con información usada por el dispositivo de almacenamiento sólido 102 para asociar el segundo paquete al objeto identificado en el primer paquete y la información de desplazamiento localizando el segundo paquete dentro del objeto y los datos. El controlador del dispositivo de almacenamiento de estado sólido 202 gestiona el banco 214 y el área física al que fluyen los paquetes.

El generador de ECC 304 recibe un paquete desde el empaquetador 302 y genera 508 el ECC para los paquetes de datos. Usualmente, no hay una relación fija entre los paquetes y los bloques de ECC. Un bloque de ECC puede comprender uno o más paquetes. Un paquete puede comprender uno o más bloques de ECC. Un paquete puede comenzar y terminar en cualquier parte dentro de un bloque de ECC. Un paquete puede comenzar en cualquier parte en un primer bloque de ECC y terminar en cualquier parte de un bloque de ECC posterior.

La memoria intermedia de sincronización de escritura 308 memoriza temporalmente 510 los paquetes según se distribuyen dentro de los bloques de ECC correspondientes antes de escribir los bloques de ECC en el almacenamiento de estado sólido 110 y a continuación el controlador de almacenamiento de estado sólido 104 escribe 512 los datos en el momento apropiado considerando las diferencias de los dominios de reloj, y el método 500 termina 514. La memoria intermedia de sincronización de escritura 308 está localizada en la frontera entre el dominio de reloj local y el domino de reloj del almacenamiento de estado sólido 110. Obsérvese que el método 500

describe recibir uno o más segmentos de datos y escribir uno o más paquetes de datos por conveniencia, pero usualmente se recibe un flujo de segmentos de datos y un grupo. Usualmente varios bloques de ECC que comprenden una página virtual completa del almacenamiento de estado sólido 110 se escriben en el almacenamiento de estado sólido 110. Usualmente el empaquetador 302 recibe segmentos de datos de un tamaño y genera paquetes de otro tamaño. Esto requiere necesariamente segmentos de datos o metadatos o partes de segmentos de datos o metadatos a combinar para formar paquetes de datos para capturar todos los datos de los segmentos dentro de los paquetes.

La Figura 5B es un diagrama de flujo esquemático que ilustra una realización de un método para una SAN en servidor de acuerdo con la presente invención. El método 501 comienza 552 y el módulo de comunicación de almacenamiento 162 facilita 554 la comunicación entre un primer controlador de almacenamiento 152a y al menos un dispositivo externo al primer servidor 112a. La comunicación entre el primer controlador de almacenamiento 152a y el dispositivo externo es independiente del primer servidor 112a. El primer controlador de almacenamiento 152a está dentro del primer servidor 112a y el primer controlador de almacenamiento 152a controla al menos un dispositivo de almacenamiento 154a. El primer servidor 112a incluye una interfaz de red 156a colocada con el primer servidor 112a y el primer controlador de almacenamiento 152a. El módulo de SAN en servidor 164 sirve 556 la petición de almacenamiento y el método 501 termina 558. El módulo de SAN en servidor 164 sirve 556 la petición de almacenamiento usando un protocolo de red y/o un protocolo de bus. El módulo de SAN en servidor 164 sirve 556 la petición de almacenamiento independiente del primer servidor 112a y la petición de servicio se recibe desde un cliente 114, 114a.

La Figura 6 es un diagrama de flujo esquemático que ilustra otra realización de un método 600 para gestionar los datos en un dispositivo de almacenamiento de estado sólido 102 usando una conducción de datos de acuerdo con la presente invención. El método 600 comienza 602 y la memoria intermedia de entrada 306 recibe 604 uno o más segmentos de datos o metadatos a escribir en el almacenamiento de estado sólido 110. El empaquetador 302 añade una cabecera a cada paquete que típicamente incluye la longitud del paquete dentro del objeto. El empaquetador 302 recibe 604 el uno o más segmentos que se almacenan en la memoria intermedia de entrada 306 y empaqueta 606 el uno o más segmentos creando uno o más paquetes dimensionados para el almacenamiento de estado sólido 110 donde cada paquete incluye una cabecera y datos del uno o más segmentos.

El generador de ECC 304 recibe un paquete desde el empaquetador 302 y genera 608 uno o más bloques de ECC para los paquetes. La memoria intermedia de sincronización de escritura 308 almacena temporalmente 610 los paquetes según se distribuyen dentro de los bloques de ECC correspondientes antes de escribir los bloques de ECC en el almacenamiento de estado sólido 110 y a continuación el controlador del almacenamiento de estado sólido 104 escribe 612 los datos en un momento apropiado considerando las diferencias de los dominios de reloj. Cuando se solicitan datos desde el almacenamiento de estado sólido 110, los bloques de ECC que comprenden uno o más paquetes de datos se leen dentro de la memoria intermedia de sincronización de lectura 328 y se almacenan temporalmente 614. Los bloques de ECC del paquete se reciben sobre el bus I/O de almacenamiento 210. Como el bus I/O de almacenamiento 210 es bidireccional, cuando se leen datos, se paran las operaciones de escritura, las operaciones de comandos, etc.

El módulo de corrección de ECC 322 recibe los bloques de ECC de los paquetes solicitados mantenidos en la memoria intermedia de sincronización de lectura 328 y corrige 616 los errores dentro de cada bloque de ECC cuando sea necesario. Si el módulo de corrección de ECC 322 determina que existen uno o más errores en un bloque de ECC y los errores son corregibles usando el síndrome de ECC, el módulo de corrección de ECC 322 corrige 616 los errores en el bloque de ECC. Si el módulo de corrección de ECC 322 determina que un error detectado no es corregible usando el ECC, el módulo de corrección de ECC 322 envía una interrupción.

El des-empaquetador 324 recibe 618 el paquete solicitado después de que el módulo de corrección de ECC 322 corrige cualesquiera errores y desempaqueta 618 los paquetes comprobando y eliminando la cabecera de paquete de cada paquete. El módulo de alineamiento 326 recibe los paquetes después de desempaquetar, elimina los datos no deseados, y reformatea 620 los paquetes de datos como segmentos de datos o metadatos de un objeto en una forma compatible con el dispositivo solicitante del segmento u objeto. La memoria intermedia de salida 330 recibe los paquetes solicitados después de desempaquetar y almacena temporalmente 622 los paquetes antes de su transmisión al dispositivo solicitante 155, y el método 600 termina 624.

La Figura 7 es un diagrama de flujo esquemático que ilustra una realización de un método 700 para la gestión de datos en un dispositivo de almacenamiento de estado sólido 102 usando un intercalado de bancos de acuerdo con la presente invención. El método 700 comienza 702 y el controlador de intercalado de bancos 344 dirige 604 uno o más comandos a dos o más colas 410, 42, 414, 416. Usualmente los agentes 402, 404, 406, 408 dirigen 704 los comandos a las colas 410, 42, 414, 416 por tipo de comando. Cada conjunto de colas 410, 42, 414, 416 incluye una cola para cada tipo de comando. El controlador de intercalado de bancos 344 coordina 706 entre los bancos 214 la ejecución de los comandos almacenados en las colas 410, 42, 414, 416 de modo que un comando de un primer tipo se ejecuta en un banco 214a mientras que un comando de un segundo tipo se ejecuta en un segundo banco 214b, y el método 700 termina 708.

Recuperación del espación de almacenamiiento

La Figura 8 es un diagrama de bloques esquemático que ilustra una realización de un aparato 800 para la recogida de basura en un dispositivo de almacenamiento de estado sólido 102 de acuerdo con la presente invención. El aparato 800 incluye un módulo de almacenamiento secuencial 802, un módulo de selección de división de almacenamiento 804, un módulo de recuperación de datos 806, y un módulo de recuperación de la división de almacenamiento 808, que se describen a continuación. En otras realizaciones, el aparato 800 incluye un módulo de marcación de basura 810 y un módulo de borrado 812.

El aparato 800 incluye un módulo de almacenamiento secuencial 802 que escribe secuencialmente paquetes de datos en una página dentro de una división de almacenamiento. Los paquetes se almacenan secuencialmente según sean paquetes nuevos o paquetes modificados. Los paquetes modificados en esta realización típicamente no se escriben de nuevo en la localización donde estaban almacenados anteriormente. En una realización, el módulo de almacenamiento secuencial 802 escribe un paquete a la primera localización en una página de una división de almacenamiento, a continuación a la siguiente localización de la página, y a la siguiente, y la siguiente, hasta que se rellena la página. El módulo de almacenamiento secuencial 802 comienza a continuación a rellenar la siguiente página en la división de almacenamiento. Esto continúa hasta que se rellena la división de almacenamiento.

En una realización preferida, el módulo de almacenamiento secuencial 802 comienza a escribir paquetes a las memorias intermedias de escritura de almacenamiento en los elementos de almacenamiento (por ejemplo SSS 0.0 a SSS M.0 216) de un banco (banco 0 214a). Cuando las memorias intermedias de escritura de almacenamiento están llenas, el controlador de almacenamiento de estado sólido 104 causa que los datos en las memorias intermedias de escritura de almacenamiento se programen en las páginas designadas dentro de los elementos de almacenamiento 216 del banco 214a. A continuación se selecciona otro banco (por ejemplo, el banco 1 214b) y el módulo de almacenamiento secuencial 802 comienza a escribir los paquetes a las memorias intermedias de escritura de almacenamiento de los elementos de almacenamiento 218 del banco 214b mientras que el primer banco 0 214a está programando las páginas designadas. Cuando las memorias intermedias de escritura de almacenamiento de este banco 214b están llenas, los contenidos de las memorias intermedias de escritura de almacenamiento se programan dentro de otra página designada en cada elemento de almacenamiento 218. Este proceso es eficiente porque mientras que un banco 214a está programando una página, las memorias intermedias de escritura de almacenamiento de otro banco 214b se pueden estar rellenando.

La división de almacenamiento incluye una porción de un almacenamiento de estado sólido 110 en un dispositivo de almacenamiento de estado sólido 102. Usualmente la división de almacenamiento es un bloque de borrado. Para memorias flash, una operación de borrado sobre un bloque de borrado escribe unos a cada bit en el bloque de borrado cargando cada célula. Este es un proceso largo en comparación con una operación de programa que comienza con una localización en la que son todo unos, y según se escriben los datos, se cambian algunos bits a cero descargando las células escritas con un cero. Sin embargo, cuando el almacenamiento de estado sólido 110 no es una memoria flash o tiene memoria flash donde el ciclo de borrado tarda una cantidad de tiempo similar que otras operaciones, tales como una operación de lectura o de programa, puede que no se requiera borrar la división de almacenamiento.

Como se usa en este documento, una división de almacenamiento es equivalente en área a un bloque de borrado pero puede borrarse o no. Cuando se usa un bloque de borrado en este documento, un bloque de borrado se puede referir a un área particular de un tamaño designado dentro de un elemento de almacenamiento (por ejemplo SS 0.0 216a) y usualmente incluye una cierta cantidad de páginas. Donde se usa "bloque de borrado" en conjunción con memoria flash, usualmente es una división de almacenamiento que se borra antes de escribirse. Donde se usa "bloque de borrado" con "almacenamiento de estado sólido" se puede borrar o no. Como se usa en este documento, un bloque de borrado puede incluir un bloque de borrado o un grupo de bloques de borrado con un bloque de borrado en cada una de las filas de elementos de almacenamiento (por ejemplo, SSS 0.0 a SSS M.0 216a - n) que también se puede referir en este documento como un bloque de borrado virtual. Cuando se refiere a la construcción lógica asociada con el bloque de borrado virtual, los bloques de borrado se pueden referir en este documento como un bloque de borrado lógico ("LEB").

Usualmente, los paquetes se almacenan secuencialmente por orden de procesamiento. En una realización, donde se usa una conducción de datos de escritura 106, el módulo de almacenamiento secuencial 802 almacena paquetes en el orden en el que salen de la conducción de datos de escritura 106. Este orden puede ser el resultado de segmentos de datos que llegan desde un dispositivo solicitante 155 mezclados con paquetes de datos válidos que se están leyendo desde otra división de almacenamiento como datos válidos que se están recuperando desde una división de almacenamiento durante una operación de recuperación como se explica más adelante. El re-enrutamiento de paquetes de datos válidos recuperados para la conducción de datos de escritura 106 puede incluir la derivación del recogedor de basuras 316 como se ha descrito anteriormente en relación con el controlador del almacenamiento de estado sólido 104 de la Figura 3.

El aparato 800 incluye un módulo de selección de la división de almacenamiento 804 que selecciona una división de almacenamiento para la recuperación. La selección de una división de almacenamiento para la recuperación puede

ser para la reutilización de la división de almacenamiento por el módulo de almacenamiento secuencial 802 para la escritura de datos, añadiendo de este modo la división de almacenamiento recuperado para la pila de almacenamiento, o para recuperar datos válidos desde la división de almacenamiento después de la determinación de que está fallando la división de almacenamiento, que no es fiable, que se debería refrescar, o por otra razón para tomar la división de almacenamiento temporalmente o permanentemente fuera de la pila de almacenamiento. En otra realización, el módulo de la sección de división de almacenamiento 804 selecciona una división de almacenamiento para la recuperación identificando una división de almacenamiento o bloque de borrado con una cantidad elevada de datos inválidos.

En otra realización, el módulo de selección de la división de almacenamiento 804 selecciona una división de almacenamiento para la recuperación identificando una división de almacenamiento o bloque de borrado con una baja cantidad de uso. Por ejemplo, identificar una división de almacenamiento o bloque de borrado con una baja cantidad de uso puede incluir identificar una división de almacenamiento con una baja cantidad de datos inválidos, un bajo número de ciclos de borrado, una baja tasa de errores de bits o una baja cuenta de programa (bajo número de veces que se escribe una página de datos en una memoria intermedia para una página en la división de almacenamiento; la cuenta de programa se puede medir desde cuando se fabricó el dispositivo, desde cuando se borró la última división de almacenamiento, desde otros eventos arbitrarios y desde combinaciones de estos). El módulo de selección de la división de almacenamiento 804 también puede usar cualquier combinación de los anteriores u otros parámetros para determinar una división de almacenamiento con una baja cantidad de uso. La selección de una división de almacenamiento para la recuperación determinando una división de almacenamiento con una baja cantidad de uso puede ser deseable para encontrar divisiones de almacenamiento que están infrautilizadas, se puede recuperar la nivelación de uso, etc.

En otra realización el módulo de selección de la división de almacenamiento 804 selecciona una división de almacenamiento para la recuperación identificando una división de almacenamiento o bloque de borrado con una alta cantidad de uso. Por ejemplo, la identificación de una división de almacenamiento o bloque de borrado con una cantidad elevada de uso puede incluir identificar una división de almacenamiento con un número elevado de ciclos de borrado, una elevada tasa de errores de bits, una división de almacenamiento con un bloque de ECC no recuperable o una elevada cuenta de programa. El módulo de selección de la división de almacenamiento 804 puede usar también cualquier combinación de los anteriores u otros parámetros para determinar una división de almacenamiento con una elevada cantidad de uso. La selección de una división de almacenamiento para recuperación determinando una división de almacenamiento con una elevada cantidad de uso puede ser deseable encontrar divisiones de almacenamiento que están sobre utilizadas, se puede recuperar refrescando la división de almacenamiento usando un ciclo de borrado, etc. o retirar la división de almacenamiento del servicio como no utilizable.

El aparato 800 incluye un módulo de recuperación de datos 806 que lee paquetes de datos válidos desde la división de almacenamiento seleccionada para recuperación, pone en cola los paquetes de datos válidos con otros paquetes de datos a escribir secuencialmente por el módulo de almacenamiento secuencial 802 y actualiza un índice con una nueva dirección física de los datos válidos escritos por el módulo de almacenamiento secuencial 802. Usualmente el índice es el índice de objeto que mapea los identificadores de objetos de datos de objetos a direcciones físicas de donde se derivan los paquetes a partir de los objetos de datos que están almacenados en el almacenamiento de estado sólido 110.

En una realización el aparato 800 incluye un módulo de recuperación de la división de almacenamiento 808 que prepara la división de almacenamiento para utilizar o reutilizar y marca la división de almacenamiento como disponible al módulo de almacenamiento secuencial 802 para escribir secuencialmente paquetes de datos después de que el módulo de recuperación de datos 806 ha completado la copia de datos válidos desde la división de almacenamiento. En otra realización, el aparato 800 incluye un módulo de recuperación de la división de almacenamiento 808 que marca la división de almacenamiento seleccionada para la recuperación como no disponible para el almacenamiento de los datos. Usualmente esto es debido a que el módulo de selección de la división de almacenamiento 804 que identifica una división de almacenamiento o bloque de borrado con una elevada cantidad de uso de modo que la división de almacenamiento o bloque de borrado no está en condición de usarse para almacenamiento de datos fiable

En una realización, el aparato 800 es un controlador del dispositivo de almacenamiento de estado sólido 202 de un dispositivo de almacenamiento de estado sólido 102. En otra realización, el aparato 800 controla un controlador del dispositivo de almacenamiento de estado sólido 202. En otra realización, una porción del aparato 800 está en un controlador del dispositivo de almacenamiento de estado sólido 202. En otra realización, el índice de objetos actualizado por el módulo de recuperación de datos 806 está también localizado en el controlador del dispositivo de almacenamiento de estado sólido 202.

En una realización, la división de almacenamiento es un bloque de borrado y el aparato 800 incluye un módulo de borrado 810 que borra un bloque de borrado seleccionado para la recuperación después de que el módulo de recuperación de datos 806 haya copiado los paquetes de datos válidos desde el bloque de borrado seleccionado y antes de que el módulo de recuperación de la división de almacenamiento 808 marque el bloque de borrado como

disponible. Para la memoria flash y otro almacenamiento de estado sólido con una operación de borrado que tarda mucho más tiempo que las operaciones de lectura o escritura, es deseable borrar un bloque de datos antes de que esté disponible para la escritura de nuevos datos para una operación eficiente. Cuando el almacenamiento de estado sólido 110 está dispuesto en bancos 214, la operación de borrado por el módulo de borrado 810 se puede ejecutar sobre un banco mientras que en otros bancos se ejecutan lecturas, escrituras u otras operaciones.

En una realización, el aparato 800 incluye un módulo de marcación de basuras 812 que identifica un paquete de datos en una división de almacenamiento como inválido en respuesta a una operación que indica que el paquete de datos ya no es válido. Por ejemplo, si se borra un paquete de datos, el módulo de marcación de basuras 812 puede identificar el paquete de datos como inválido. Una operación de leer - modificar - escribir es otro modo para identificar un paquete de datos como inválido. En una realización el módulo de marcación de basuras 812 puede identificar el paquete de datos como inválido actualizando un índice. En otra realización, el módulo de marcación de basuras 812 puede identificar el paquete de datos como inválido almacenando otro paquete de datos que indica que el paquete de datos inválido se ha borrado. Esto es ventajoso porque el almacenamiento de la información de que se ha borrado el paquete, en el almacenamiento de estado sólido 110, permite al módulo de reconstrucción de índices de objetos 272 o módulo similar reconstruir el índice del objeto con una entrada que indica que se ha borrado el paquete de datos inválidos.

En una realización, el aparato 800 se puede usar para rellenar el resto de una página virtual de datos siguiendo un comando de traspaso para mejorar el funcionamiento global, donde el comando de traspaso para el flujo de datos, dentro de la conducción de escritura 106 hasta que la conducción de escritura 106 se vacía y todos los paquetes se han escrito permanentemente dentro del almacenamiento de estado sólido no volátil 110. Esto tiene el beneficio de reducir la cantidad requerida de recogida de basura, la cantidad de tiempo usado para borrar las divisiones de almacenamiento y la cantidad de tiempo requerido para programar páginas virtuales. Por ejemplo, se puede recibir un comando de traspaso cuando solo un pequeño paquete está preparado para escribirse dentro de la página virtual del almacenamiento de estado sólido 110. La programación de esta página virtual casi vacía podría dar como resultado la necesidad de recuperar inmediatamente el espacio gastado, causando que los datos válidos dentro de la división de almacenamiento se recojan como basura innecesariamente y la división de almacenamiento borrada, recuperada y devuelta a la pila de espacio disponible para escritura por el módulo de almacenamiento secuencial 802.

La marcación de los paquetes de datos como inválidos en lugar de borrar realmente un paquete de datos inválido, es eficiente porque, como se ha mencionado anteriormente, para la memoria flash y para otros almacenamientos similares una operación de borrado tarda una cantidad significativa de tiempo. Permitir a un sistema de recogida de basura como se ha descrito en el aparato 800 operar de forma autónoma dentro del almacenamiento de estado sólido 110 proporciona un modo de separar las operaciones de borrado de las operaciones de lectura, escritura y otras operaciones más rápidas de modo que el dispositivo de almacenamiento de estado sólido 102 puede operar más rápido que muchos otros sistemas de almacenamiento de estado sólido o dispositivos de almacenamiento de datos.

La Figura 9 es un diagrama de flujo esquemático que ilustra una realización de un método 900 para la recuperación de almacenamiento de acuerdo con la presente invención. El método 900 comienza 902 y el módulo de almacenamiento secuencial 802 escribe secuencialmente 904 paquetes de datos en una división de almacenamiento. La división de almacenamiento es una porción de un almacenamiento de estado sólido 110 en un dispositivo de almacenamiento de estado sólido 102. Usualmente una división de almacenamiento es un bloque de borrado. Los paquetes de datos se derivan de un objeto y los paquetes de datos se almacenan secuencialmente por orden de procesamiento.

El módulo de selección de la división de almacenamiento 804 selecciona 906 una división de almacenamiento para recuperar y el módulo de recuperación de datos 806 lee 908 los paquetes de datos válidos desde la división de almacenamiento seleccionada para la recuperación. Típicamente los paquetes de datos válidos son paquetes de datos que no se han marcado para el borrado o eliminación o alguna otra marcación de datos inválidos y se consideran datos válidos o "buenos". El módulo de recuperación de datos 806 pone en cola 910 los paquetes de datos válidos con otros paquetes de datos programados para escribir secuencialmente por el módulo de almacenamiento secuencial 802. El módulo de recuperación de datos 806 actualiza 912 un índice con una nueva dirección física de datos válidos escritos por el módulo de almacenamiento secuencial 802. El índice incluye un mapeo de direcciones físicas de paquetes de datos a identificadores de objetos. Los paquetes de datos son los almacenados en el almacenamiento de estado sólido 110 y los identificadores de objetos corresponden a los paquetes de datos.

Después de que el módulo de recuperación de datos 806 completa la copia de datos válidos desde la división de almacenamiento, el módulo de recuperación de la división de almacenamiento 808 marca 914 la división de almacenamiento seleccionada para recuperación como disponible para el módulo de almacenamiento secuencial 802 para la escritura secuencial de paquetes de datos y el método 900 termina 916.

Raid distribuida del extremo frontal

Los sistemas de RAID tradicionales están configurados con un controlador de RAID que funciona para recibir datos, calcular patrones de desmontaje para los datos, dividir los datos en segmentos de datos, calcular una banda de paridad, almacenar los datos sobre dispositivos de almacenamiento, actualizar los segmentos de datos, etc. Aunque algunos controladores de RAID permiten que algunas funciones estén distribuidas, los dispositivos de almacenamiento gestionados por el controlador de RAID no comunican con los clientes 114 directamente para el almacenamiento de los datos desmontados en una RAID. En efecto las solicitudes de almacenamiento y los datos para una estructuración en RAID pasan a través del controlador de RAID.

Requerir al controlador de RAID para que toque todos los datos a almacenar en una RAID es ineficiente porque crea un cuello de botella en el flujo de datos. Esto es especialmente cierto durante un proceso de lectura - modificación - escritura donde se consume ancho de banda y tiempo de funcionamiento de todas las unidades en el grupo de RAID mientras que solo se actualiza realmente un subconjunto. Además, la región del dispositivo de almacenamiento designado para los datos gestionados por el controlador de RAID se dedica usualmente al grupo de RAID y no se puede acceder de forma independiente. El acceso a un dispositivo de almacenamiento 150 por un cliente usualmente debe lograrse por partición del dispositivo de almacenamiento 150. Cuando se usa la partición, las particiones accesibles para el almacenamiento general no se usan para la RAID y las particiones asignadas al grupo de RAID no son accesibles para el almacenamiento de datos general. Los esquemas que sobre-suscriben las particiones para una utilización de optimización global son complejos y más difíciles de gestionar. Además, el espacio de almacenamiento asignando para un grupo de RAID no se puede acceder por más de un controlador RAID a menos que se designe uno como maestro y otros controladores de RAID actúen como esclavos salvo que el controlador de RAID maestro esté inactivo, no funcional, etc.

Los controladores de RAID típicos también generan segmentos de datos de paridad fuera de los dispositivos de almacenamiento 150 del grupo de RAID. Esto puede ser ineficaz porque los segmentos de los datos de paridad se generan usualmente y se envían a continuación a un dispositivo de almacenamiento 150 para su almacenamiento, lo que requiere capacidad de cálculo del controlador de RAID. El seguimiento de la localización de segmentos de datos de paridad y las actualizaciones también se debe hacer en el controlador de RAID en vez de hacerse autónomamente en el dispositivo de almacenamiento 150.

Cuando sea necesario asegurar que los datos permanecen disponibles si el controlador de RAID separado está fuera de línea, los controladores de RAID típicamente están en conexión cruzada con las unidades y entre sí, y/o en espejo como conjuntos completos, haciendo la disponibilidad de datos cara y difícil de gestionar y reduciendo drásticamente la fiabilidad del subsistema de almacenamiento

Lo que se necesita es un sistema, un aparato y un método para una RAID distribuida del extremo frontal que permita la estructuración en RAID por segmento de datos, por objeto, por fichero o base similar y que elimine la necesidad de controladores RAID y los pares de controladores RAID situados entre el cliente y los dispositivos de almacenamiento. En tal sistema, aparato y método se puede crear un grupo de RAID para un segmento de datos objeto o fichero y gestionarse dentro de un grupo de dispositivos de almacenamiento por un controlador de RAID mientras que un segundo grupo de RAID se puede crear para otro segmento de datos, objeto, fichero que abarca algunos de los mismos dispositivos de almacenamiento del primer grupo de RAID. Las funciones de control de RAID se pueden distribuir entre los clientes 114, un dispositivo de gestión de RAID de tercera parte o entre los dispositivos de almacenamiento 150. El sistema de RAID distribuido del extremo frontal, el aparato y el método también pueden enviar comandos a los dispositivos de almacenamiento 150 de un grupo de RAID y pueden permitir a los dispositivos de almacenamiento 150 acceder directamente y copiar los datos a través de un acceso directo a memoria ("DMA"), o DMA remoto ("RDMA").

La Figura 10 es un diagrama de bloques esquemático que ilustra una realización de un sistema 1600 que se puede acceder para una RAID distribuida del extremo frontal de acuerdo con las presentes invenciones. Las descripciones anteriores para las componentes representadas en la Figura 10 relacionadas con una RAID progresiva también son aplicables a una RAID distribuida del extremo frontal. Con respecto a una RAID distribuida del extremo frontal, el conjunto de dispositivos de almacenamiento 1604 forma un grupo de RAID e incluye dispositivos de almacenamiento 150 que son autónomos y son capaces de recibir independientemente y dar servicio a las peticiones de almacenamiento desde un cliente 114 sobre una red 116 o una o más redes redundantes 116.

Entre los dispositivos de almacenamiento 150 dentro del conjunto de dispositivos de almacenamiento 1604, uno o más se designan como dispositivos de almacenamiento de paridad-espejo 1602 para una banda. Típicamente, el uno o más dispositivos de almacenamiento de paridad-espejo 1602 funcionan de forma sustancialmente similar a los otros dispositivos de almacenamiento 150. En configuraciones típicas donde los dispositivos de almacenamiento designados de paridad-espejo 1602 alternan entre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604, los dispositivos de almacenamiento de paridad-espejo 1602 tienen esencialmente las mismas características que los otros dispositivos de almacenamiento 150 debido a que también deben operar como dispositivos de almacenamiento no de paridad-espejo. Las características similares son con respecto a la operación dentro de un grupo de RAID y la operación autónoma para una comunicación con un cliente

independiente 114 como se ha descrito anteriormente. En diversas realizaciones, los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 pueden diferir en otros aspectos no relacionados con el funcionamiento dentro del entorno de RAID descrito.

Los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 pueden ser independientes, agrupados dentro de uno o más servidores 112, puede residir cada uno en un servidor 112, se pueden acceder a través de uno o más servidores 112, etc. Uno o más clientes 114 pueden residir en servidores 112 que incluyen uno o más dispositivos de almacenamiento 150, pueden residir en servidores separados 112, pueden residir en ordenadores, estaciones de trabajo, ordenadores portátiles, etc. que acceden a los dispositivos de almacenamiento 150 a través de una o más redes de ordenadores 116 o similares.

En una realización, la red 116 incluye un bus de sistema y uno o más de los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604 comunican usando el bus del sistema. Por ejemplo, el bus del sistema puede ser un bus PCI-e, un bus de Conexión de Tecnología Avanzada Serie ("ATA serie"), un bus ATA en paralelo o similares. En otra realización, el bus del sistema es un bus externo tal como una interfaz de un pequeño sistema de ordenadores ("SCSI"), FireWire, Fiber Channel, USB, PCIe-AS, Infiniband o similares. Un experto en la materia apreciará otras configuraciones de sistema 1600 con dispositivos de almacenamiento 150 que son autónomos y son capaces de recibir de forma independiente y servir peticiones de almacenamiento desde un cliente 114 sobre una o más redes 116.

La Figura 11 es un diagrama de bloque esquemático que ilustra una realización de un aparato 2100 para una RAID distribuida del extremo frontal de acuerdo con la presente invención. El aparato 2100, en diversas realizaciones, incluye un módulo receptor de peticiones de almacenamiento 2102, un módulo de asociación de desmontaje 2104, un módulo de asociación de paridad-espejo 2106, un módulo transmisor de peticiones de almacenamiento 2108, un módulo de generación de paridad del extremo frontal 2110, u módulo de alternancia de paridad 2118, un módulo de recuperación de segmentos de datos 2112, un módulo de reconstrucción de datos 2114, un módulo de reconstrucción de paridad 2116 y un módulo de comunicaciones de entre pares 2120, que se describen a continuación. En diversas realizaciones, el aparato 2100 se puede incluir en un dispositivo de almacenamiento 150 tal como un dispositivo de almacenamiento de estado sólido 102, un controlador del dispositivo de almacenamiento 152 tal como un controlador de almacenamiento de estado sólido 104, un servidor 112, un dispositivo de gestión de RAID de terceras partes, etc. o se puede distribuir entre más de un componente.

El aparato 2100 incluye un módulo receptor de peticiones de almacenamiento 2102 que recibe una petición de almacenamiento para almacenar datos en un conjunto de dispositivos de almacenamiento 1604. Los datos pueden ser una porción de un fichero o un objeto o pueden ser un fichero entero u objeto. Un fichero puede incluir un bloque de información arbitraria o recurso para almacenar información, que está disponible para un programa de ordenador. Un fichero puede incluir cualquier estructura de datos accedida por un procesador. Un fichero puede incluir una base de datos, una cadena de texto, un código de ordenador, etc. Un objeto es usualmente una estructura de datos para la programación orientada a objetos y puede incluir una estructura con o sin datos. En una realización, un objeto es un subconjunto de un fichero. En otra realización un objeto es independiente de un fichero. En cualquier caso; un objeto y un fichero se definen en este documento para incluir el conjunto entero de datos, estructuras de datos, código de ordenador y otra información que se puede almacenar sobre un dispositivo de almacenamiento.

El conjunto de dispositivos de almacenamiento 1604 incluye dispositivos de almacenamiento autónomos 150 que forman un grupo de RAID que reciben independientemente peticiones de almacenamiento desde un cliente 114 sobre una o más redes 116. Uno o más de los dispositivos de almacenamiento autónomos 150 dentro del conjunto de dispositivos de almacenamiento 1604 se designan como dispositivos de almacenamiento de paridad-espejo 1602 para una banda. Otras peticiones de almacenamiento desde otro cliente 114 se pueden almacenar sobre un segundo conjunto de dispositivos de almacenamiento donde el segundo conjunto de dispositivos de almacenamiento puede incluir uno o más de los mismos dispositivos de almacenamiento 150 (y dispositivos de almacenamiento de paridad-espejo 1602) como el primer conjunto de dispositivos de almacenamiento 1604. Los dispositivos de almacenamiento 150 comunes a ambos conjuntos de dispositivos de almacenamiento 1604 pueden tener espacio de almacenamiento asignado solapante dentro de los dispositivos de almacenamiento común 150.

El aparato 2100 incluye un módulo de asociación de desmontaje 2104 que calcula un patrón de bandas para los datos. El patrón de bandas incluye una o más bandas. Cada banda incluye un conjunto de N segmentos de datos. Los N segmentos de datos de un banda pueden incluir también uno o más segmentos de datos vacíos. El módulo de asociación de desmontaje 2104 asocia cada uno de los N segmentos de datos con uno de los N dispositivos de almacenamiento 150a - n en el conjunto de dispositivos de almacenamiento 1604 asignados a la banda. En una realización el módulo de asociación de desmontaje 2104 asocia un segmento de datos con un dispositivo de almacenamiento 150 con una petición de almacenamiento a enviar al dispositivo de almacenamiento 150 que dirige al dispositivo de almacenamiento 150 para obtener los datos correspondientes al segmento de datos desde el cliente 114 que envía la petición de almacenamiento.

En otra realización, la petición de almacenamiento está sustancialmente libre de datos de los segmentos de datos. Sustancialmente libre de datos significa que la petición de almacenamiento generalmente no incluye los datos que

son el sujeto de la petición de almacenamiento pero pueden incluir caracteres, cadenas de caracteres, etc. que pueden ser parte de los datos. Por ejemplo, si los datos comprenden una serie de caracteres repetidos, idénticos, tal como una serie de ceros, la petición de almacenamiento puede incluir una indicación de que los datos incluyen una serie de ceros, sin incluir todos los ceros contenidos en los datos. Un experto en la materia reconocerá otros modos para enviar una petición de almacenamiento sin enviar el volumen de datos mientras que se permite una pequeña cantidad o una instancia única de ciertos caracteres o cadenas de caracteres en la petición de almacenamiento. La petición de almacenamiento incluye comandos que permiten N dispositivos de almacenamiento 150a - n para recuperar los datos usando una operación de DMA o RDMA o similares.

En otra realización, el módulo de asociación de desmontaje 2104 asocia un segmento de datos con un dispositivo de almacenamiento 150 identificando en una petición de almacenamiento a enviar al dispositivo de almacenamiento 150 los datos del segmento de datos. La Identificación de datos del segmento de datos puede incluir un identificador del segmento de datos, una localización o dirección del segmento de datos, una longitud del segmento de datos u otra información que permitirá al dispositivo de almacenamiento 150 reconocer qué datos comprende el segmento de datos.

En una realización, el módulo de asociación de desmontaje 2104 asocia un segmento de datos con un dispositivo de almacenamiento 150 en una petición de almacenamiento de modo que un cliente 114 puede enviar los datos que comprenden los segmentos de datos en una difusión de modo que cada dispositivo de almacenamiento 150 es capaz de almacenar los segmentos de datos asociados y descartar los datos correspondientes a los segmentos de datos no asignados al dispositivo de almacenamiento 150. En otra realización el módulo de asociación de desmontaje 2104 asocia un segmento de datos con un dispositivo de almacenamiento 150 en una petición de almacenamiento, posiblemente direccionando cada segmento de datos, de modo que un cliente 114 puede enviar datos que comprenden los segmentos de datos en una transmisión multi-destino de modo que cada dispositivo de almacenamiento 150 es capaz de almacenar los segmentos de datos asociados y descartar los datos correspondientes a segmentos de datos no asignados al dispositivo de almacenamiento 150. Un experto en la materia reconocerá otros modos para que el módulo de asociación de desmontaje 2104 asocie un segmento de datos con un dispositivo de almacenamiento 150 para su difusión, transmisión multi-destino, transmisión de origen y destino únicos, difusión limitada, etc. uno o más segmentos de datos a uno o más dispositivos de almacenamiento.

En una realización relacionada, el módulo de asociación de desmontaje 2104 asocia un segmento de datos con un dispositivo de almacenamiento 150 en una petición de almacenamiento de modo que un cliente 114 puede difundir, transmitir multi-destino, transmitir con destino único, etc. la petición de almacenamiento y cada dispositivo de almacenamiento 150 es capaz de recibir una porción de la petición de almacenamiento desde el cliente 114 que pertenece al segmento de datos asociado con el dispositivo de almacenamiento 150 y puede descartar porciones de la petición de almacenamiento que no pertenecen al uno o más segmentos de datos no asociados con el dispositivo de almacenamiento 150.

En otra realización, la petición de almacenamiento recibida por el módulo receptor de peticiones de almacenamiento 2102 incluye los datos que son el sujeto de la petición de almacenamiento y el módulo de asociación de desmontaje 2104 asocia un segmento de datos con un dispositivo de almacenamiento 150 preparando una petición de almacenamiento para el dispositivo de almacenamiento 150 que incluye el segmento de datos. El módulo de asociación de desmontaje 2104 puede operar dentro de un cliente 114, un dispositivo de gestión de terceras partes RAID, un dispositivo de almacenamiento 150, 1602, etc.

El aparato 2100 incluye un módulo de asociación de paridad-espejo 2106 que asocia un conjunto de N segmentos de datos con uno o más dispositivos de almacenamiento de paridad-espejo 1602 en el conjunto de dispositivos de almacenamiento 1604. El uno o más dispositivos de almacenamiento de paridad-espejo 1602 están en adición a los N dispositivos de almacenamiento 150a - n. En una realización el módulo de asociación de paridad-espejo 2106 asocia un conjunto de N segmentos de datos a los dispositivos de almacenamiento de paridad-espejo 1602 de modo que cada dispositivo de almacenamiento de paridad-espejo 1602 puede recibir y almacenar los N segmentos de datos de una banda para la generación de un segmento de datos de paridad. En otra realización, el módulo de asociación de paridad-espejo 2106 asocia un segmento de datos de la banda con cada dispositivo de almacenamiento de paridad-espejo 1602 de modo que los dispositivos de almacenamiento 1602a - m actúan como un espejo para los N segmentos de datos almacenados sobre los N dispositivos de almacenamiento 150a - n.

En diversas realizaciones, el módulo de asociación de paridad- espejo 2106 asocia el conjunto de N segmentos de datos con el uno o dos dispositivos de almacenamiento de paridad-espejo 1602 usando una única petición de almacenamiento, múltiples peticiones de almacenamiento, u otra técnica de asociación descrita anteriormente en relación con el módulo de asociación de desmontaje 2104, tal como las peticiones de almacenamiento que configuran el dispositivo de almacenamiento de paridad-espejo 1602 para DMA, RDMA, difusión, transmisión multi-destino, o incluyendo los N segmentos de datos en las peticiones de almacenamiento. El módulo de asociación de paridad-espejo 2106 puede operar dentro de un cliente 114, un dispositivo de gestión de RAID de tercera parte, un dispositivo de almacenamiento 150, 1602, etc.

El aparato 2100 incluye un módulo transmisor de peticiones de almacenamiento 2108 que transmite una o más peticiones de almacenamiento a cada uno de los dispositivos de almacenamiento 150, 1602 en el conjunto de dispositivos de almacenamiento 1604, siendo cada una de las peticiones de almacenamiento suficiente para almacenar sobre el dispositivo de almacenamiento 150, 1602 el uno o más segmentos de datos asociados con el dispositivo de almacenamiento 150, 1602 que recibe la petición de almacenamiento. En una realización, cada una de las peticiones de almacenamiento no incluye los datos que son el sujeto de la petición de almacenamiento. En una realización adicional, cada petición de almacenamiento posibilita a los N dispositivos de almacenamiento 150 y los dispositivos de almacenamiento de paridad-espejo 1602 del conjunto de dispositivos de almacenamiento 1604 para descargar datos de un segmento de datos asociado usando DMA o RDMA. En otra realización, una petición de almacenamiento contiene suficiente información para recoger las peticiones de almacenamiento relevantes o datos relevantes para los segmentos de datos asociados a partir de una difusión desde el cliente 114. En otra realización, una petición de almacenamiento incluye los datos de un segmento de datos asociado.

En una realización, cada petición de almacenamiento identifica los dispositivos de almacenamiento 150, 1602 que son parte del conjunto de dispositivos de almacenamiento 1604 de una banda. Incluyendo una identificación de los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604. En el caso de un fallo de un dispositivo de almacenamiento 150 que actúa como maestro, otro dispositivo de almacenamiento 150 puede tomar el control como maestro para gestionar los datos con estructuración en RAID. En otra realización, la identificación del conjunto de dispositivos de almacenamiento 1604 posibilita a los dispositivos de almacenamiento autónomos 150, 1602 recuperar los datos cuando un dispositivo de almacenamiento está fuera de línea, y reconstruir los datos cuando se añade un dispositivo de almacenamiento de sustitución dentro del conjunto de dispositivos de almacenamiento 1604 independientes del cliente. En otra realización, la identificación de los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604 representa un grupo de transmisión multi-destino para la transmisión de segmentos de datos o peticiones de almacenamiento. La identificación se puede almacenar junto con los metadatos para el objeto o fichero almacenado sobre los dispositivos de almacenamiento 150, 1602 dentro del conjunto de dispositivos de almacenamiento 1604.

En una realización, cuando el módulo de asociación de paridad-espejo 2106 asocia un conjunto de N segmentos de datos con cada uno del uno o más dispositivos de almacenamiento de paridad-espejo 1602, el aparato 2100 incluye un módulo de generación de paridad del extremo frontal 2110 que calcula, independientemente del cliente 114, un segmento de datos de paridad para la banda y almacena el segmento de datos de paridad sobre el dispositivo de almacenamiento de paridad-espejo 1602. El segmento de datos de paridad se calcula a partir del conjunto de N segmentos de datos proporcionados al dispositivo de almacenamiento de paridad-espejo 1602. Cuando más de un dispositivo de almacenamiento de paridad-espejo 1602 está incluido en el conjunto de dispositivos de almacenamiento 1604, el módulo de generación de paridad del extremo frontal 2110 usualmente genera los diversos segmentos de datos de paridad de modo que dos o más dispositivos de almacenamiento 150, 1602 en el conjunto de dispositivos de almacenamiento 1604 pueden fallar y la información de los segmentos de datos de paridad permite la recuperación de los segmentos de datos no disponibles o los segmentos de datos de paridad.

En otra realización, el módulo de generación de paridad del extremo frontal 2110 calcula el segmento de datos de paridad cuando opera dentro de un dispositivo de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 y/o un dispositivo de gestión de RAID de terceras partes. Por ejemplo, un servidor 112 separado del cliente 114 que transmite la petición de almacenamiento puede calcular el segmento de datos de paridad. En otra realización, el módulo de generación de paridad del extremo frontal 2110 opera dentro de un dispositivo de almacenamiento de paridad-espejo para calcular el segmento de datos de paridad. Por ejemplo, un controlador de almacenamiento 152 en el dispositivo de almacenamiento de paridad-espejo 1602 puede actuar como un controlador de almacenamiento maestro para el grupo de RAID formado por el conjunto de dispositivos de almacenamiento 1604.

En otra realización, el módulo de generación de paridad del extremo frontal 2110 calcula el segmento de datos de paridad y transmite a continuación el segmento de datos de paridad calculado a uno o más dispositivos de almacenamiento de paridad- espejo adicionales 1602 en un segundo conjunto de dispositivos de almacenamiento formando un espejo. Esta realización es ventajosa porque el código de control asociado con el cálculo del segmento de datos de paridad se realiza una vez, en lugar de hacerse para cada conjunto de dispositivos de almacenamiento 1604 con el beneficio adicional de reducir el tráfico de datos sobre la red 116.

El aparato 2100 puede incluir también, en una realización, un módulo de recuperación de segmentos de datos 2112 que recupera un segmento de datos almacenado sobre un dispositivo de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 si el dispositivo de almacenamiento 150 no está disponible y se recibe una petición para leer bien los segmentos de datos no disponibles o los datos que incluye el segmento de datos no disponible. El segmento de datos se recupera usando los segmentos de datos sobre dispositivos de almacenamiento disponibles 150 del conjunto de dispositivos de almacenamiento 1604, una combinación de los segmentos de datos de paridad y segmentos de datos sobre dispositivos de almacenamiento disponibles 150, 1602 del conjunto de dispositivos de almacenamiento 1604, o desde un dispositivo de almacenamiento de espejo que contiene una copia del segmento de datos. Típicamente, el dispositivo de almacenamiento de espejo es un dispositivo de almacenamiento 150 de un conjunto de dispositivos de almacenamiento de espejo que almacena una copia de los N

segmentos de datos. El módulo de recuperación de segmentos de datos 2112 puede operar y recuperar los segmentos de datos no disponibles desde el dispositivo de almacenamiento 150, un dispositivo de almacenamiento de paridad-espejo 1602, un dispositivo de gestión de RAID de terceras partes, un dispositivo de almacenamiento de espejo, etc.

En otra realización, el aparato 2100 incluye un módulo de reconstrucción de datos 2114 que almacena un segmento de datos recuperado sobre un dispositivo de almacenamiento de sustitución 150 en una operación de reconstrucción. Por ejemplo, si un dispositivo de almacenamiento 150 se convierte en no disponible debido a un fallo, pérdida de sincronización, etc. el módulo de reconstrucción de datos 2114 puede reconstruir un dispositivo de almacenamiento 150 para sustituir al dispositivo de almacenamiento no disponible 150. En una realización, el dispositivo de almacenamiento reconstruido 150 es el dispositivo de almacenamiento original 150 que se ha convertido en disponible.

El segmento de datos recuperado iguala a un segmento de datos no disponible almacenado sobre el dispositivo de almacenamiento no disponible 150 del conjunto de dispositivos de almacenamiento 1604. La operación de reconstrucción usualmente restaura uno o más segmentos de datos y segmentos de datos de paridad sobre el dispositivo de almacenamiento de sustitución 150 para igualar segmentos de datos y segmentos de datos de paridad almacenados anteriormente sobre el dispositivo de almacenamiento no disponible 150.

En una realización, el segmento de datos recuperado se recupera para la operación de reconstrucción usando los segmentos de datos disponibles sobre dispositivos de almacenamiento disponibles 150 del conjunto de dispositivos de almacenamiento 1604. En otra realización, el segmento de datos recuperado se recupera para la operación de reconstrucción usando una combinación de un segmento de datos de paridad desde uno o más dispositivos de almacenamiento de paridad-espejo 1602 y los segmentos de datos disponibles sobre dispositivos de almacenamiento disponibles 150 del conjunto de dispositivos de almacenamiento 1604. En otra realización, el segmento de datos recuperado se recupera para la operación de reconstrucción usando un segmento de datos de compaginación leídos desde un dispositivo de almacenamiento de paridad-espejo 1602. En otra realización más, el segmento de datos recuperado se recupera para la operación de reconstrucción usando un segmento de datos de compaginación desde un dispositivo de almacenamiento de espejo. El módulo de reconstrucción de datos 2114 podría operar y almacenar un segmento de datos recibido desde un cliente 114, un dispositivo de gestión de RAID de terceras partes, un dispositivo de almacenamiento 150, 1602, un dispositivo de almacenamiento de espejo, etc.

El aparato 2100 incluye, en otra realización, un módulo de reconstrucción de paridad 2116 que reconstruye el segmento de datos de paridad recuperado sobre un dispositivo de almacenamiento de sustitución 1602 en una operación de reconstrucción. Una operación de reconstrucción es sustancialmente similar a la operación de reconstrucción descrita en relación con el módulo de reconstrucción de datos 2114. El módulo de reconstrucción de paridad 2116 opera similar al módulo de reconstrucción de datos 2114 excepto que el módulo de reconstrucción de paridad 2116 reconstruye un segmento de datos de paridad recuperado. El segmento de datos de paridad recuperado compagina un segmento de datos de paridad no disponible almacenado sobre un dispositivo de almacenamiento de paridad-espejo no disponible 1602 asignado a la banda.

El segmento de datos de paridad se recupera, en diversas realizaciones, copiando el segmento de datos de paridad almacenado sobre un dispositivo de almacenamiento de paridad-espejo 1602 en un conjunto de dispositivos de almacenamiento de espejo, copiando el segmento de datos de paridad desde el dispositivo de almacenamiento de paridad-espejo 1602 en el conjunto de dispositivos de almacenamiento 1604 (si es idéntico al segmento de datos de paridad no disponible), generando el segmento de datos de paridad usando uno o más de los N segmentos de datos y los segmentos de datos de parida almacenados sobre los dispositivos de almacenamiento disponibles 150, 1602 del conjunto de dispositivos de almacenamiento 1604 y un dispositivo de almacenamiento de espejo que contiene una copia de un segmento de datos, etc. El módulo de reconstrucción de datos 2114 puede operar y almacenar un segmento de datos recuperado mientras que reside sobre un cliente 114, un dispositivo de gestión de RAID de terceras partes, un dispositivo de almacenamiento 150, un dispositivo de almacenamiento de espejo, etc.

Ventajosamente, el aparato 2100 no está limitado al almacenamiento de datos en los dispositivos de almacenamiento 150, 1602 para las particiones dedicadas a la operación de RAID distribuida del extremo frontal descrita en este documento. En cambio, un dispositivo de almacenamiento autónomo (por ejemplo 150a) puede recibir independientemente peticiones de almacenamiento desde un cliente 114 para almacenar datos con estructura de RAID o sin estructura de RAID en una o más regiones del dispositivo de almacenamiento 150a que también están disponibles para el almacenamiento de datos por el módulo de asociación de desmontaje 2104, el módulo de asociación de paridad-espejo 2106 y el módulo de generación de paridad del extremo frontal 2110.

En una realización, una o más peticiones de almacenamiento recibidas por el módulo receptor de peticiones de almacenamiento 2102 o transmitidas por el módulo transmisor de peticiones de almacenamiento 2108, identifican los dispositivos de almacenamiento 150 que comprende el conjunto de dispositivos de almacenamiento 1604 de la banda. Ventajosamente, la identificación del dispositivo de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 en las peticiones de almacenamiento facilita a un controlador de RAID de respaldo operar si un controlador maestro no está funcional. Por ejemplo, si los dispositivos de almacenamiento 150 del conjunto de

dispositivos de almacenamiento 1604 se identifican en peticiones de almacenamiento y el controlador maestro es un dispositivo de almacenamiento de paridad-espejo 1602 y no está disponible, otro dispositivo de almacenamiento de paridad-espejo 1602 u otro de los N dispositivos de almacenamiento 150a - n puede convertirse en el controlador maestro.

En una realización, el aparato 2100 incluye un módulo de alternancia de paridad 2118 que alterna, para cada banda, los dispositivos de almacenamiento 150 en el conjunto de dispositivos de almacenamiento 1604 que están designados como dispositivos de almacenamiento de paridad-espejo 1602 para la banda. Los beneficios del módulo de alternancia de paridad 2118 se han descrito anteriormente. En otra realización, los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 forman un grupo de pares y el aparato 2100 incluye un módulo de comunicación entre pares 2120 que transmite y recibe peticiones de almacenamiento dentro de los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604. El módulo de comunicaciones entre pares 2120 también puede transmitir y recibir peticiones de almacenamiento con dispositivos pares fuera del conjunto de dispositivos de almacenamiento 1604.

En una realización preferida, la petición de almacenamiento es una petición de objetos para almacenar un objeto desmontando los datos del objeto a través de los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604 usando los módulos 2102 - 2120 del aparato 2100. En otra realización, uno o más de los dispositivos de almacenamiento autónomos 150, 1602 del conjunto de dispositivos de almacenamiento 1604 están asignados dentro de un primer grupo de RAID para al menos una porción de un primer objeto o fichero y asignados dentro de un segundo grupo de RAID para el menos una porción de un segundo objeto o fichero. Por ejemplo, un dispositivo de almacenamiento 150a puede ser un controlador de RAID maestro para el conjunto de dispositivos de almacenamiento 1604 para una o más bandas y un segundo dispositivo de almacenamiento 150b puede ser un controlador de RAID maestro para un grupo de RAID que incluye algunos o todos los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604. Ventajosamente, el aparato 2100 permite flexibilidad en el agrupamiento de los dispositivos de almacenamiento 150, 1602 para formar grupos de RAID para diversos clientes 114.

La Figura 12 es un diagrama de flujo esquemático que ilustra una realización de un método 2200 para la RAID distribuida del extremo frontal de acuerdo con la presente invención. El método 2200 comienza 2202 y el módulo receptor de peticiones de almacenamiento 2102 recibe 2204 una petición de almacenamiento para almacenar datos en N dispositivos de almacenamiento 150a - n del conjunto de dispositivos de almacenamiento 1604. El módulo de asociación de desmontaje 2104 calcula 2206 un patrón de bandas para los datos y asocia 2208 los N segmentos de datos cada uno con uno de los N dispositivos de almacenamiento 150a - n.

El módulo de asociación de paridad-espejo 2106 asocia 2210 un conjunto de N segmentos de datos con uno o más dispositivos de almacenamiento de paridad-espejo 1602. El módulo transmisor de la petición de almacenamiento 2108 transmite 2212 una o más peticiones de almacenamiento a cada dispositivo de almacenamiento 150, 1602 en el conjunto de dispositivos de almacenamiento 1604. Cada petición de almacenamiento es suficiente para almacenar sobre el dispositivo de almacenamiento 150 el uno o más segmentos de datos asociados con el dispositivo de almacenamiento 150 que reciben la petición de almacenamiento. Los segmentos de los datos se transfieren a continuación a los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604 usando DMA, RDMA, difusión, transmisión multi-destino, etc. según se dirige por las peticiones de almacenamiento. Opcionalmente, el módulo de generación de paridad del extremo frontal 2110 calcula 2214 un segmento de datos de paridad para la banda y el método 2200 termina 2216.

Raid distribuida, del extremo frontal, compartida

La RAID tradicional usa una red de discos u otros dispositivos de almacenamiento con al menos una porción de cada dispositivo dedicada a la RAID y la formación del grupo de RAID. Un controlador de RAID gestiona las peticiones de almacenamiento para el grupo de RAID. Para los sistemas redundantes, el controlador de RAID tiene un controlador de RAID de respaldo listo para tomar el control si falla el controlador de RAID maestro o no está disponible. Las peticiones de almacenamiento desde múltiples clientes que buscan acceder a los mismos datos almacenados en la RAID usualmente se ejecutan de forma secuencial por orden de llegada.

Un sistema de RAID distribuido, del extremo frontal, como se ha descrito anteriormente en relación con el sistema 1600, el aparato 2100 y el método 2200 representado en las Figuras 10, 11 y 12 respectivamente, incluye dispositivos de almacenamiento autónomos 150 que pueden incluir cada uno un controlador de almacenamiento 152 que puede funcionar como un controlador de RAID distribuido y los dispositivos de almacenamiento 150 se pueden configurar cada uno en múltiples grupos de RAID solapantes que sirven a múltiples clientes 114. Ocasionalmente, dos clientes 114 pueden buscar acceder a los mismos datos. Si una petición de almacenamiento llega primero y se ejecuta entonces usualmente no hay ninguna inconsistencia de los datos. Si, por el contrario, dos o más peticiones de almacenamiento para los mismos datos llegan en el mismo momento o casi el mismo momento, los datos se pueden corromper.

Por ejemplo, si los datos se almacenan sobre cuatro dispositivos de almacenamiento 150 en un grupo de RAID, donde uno de los dispositivos de almacenamiento 150 se asigna como dispositivo de almacenamiento de paridad-espejo 1602, y un primer cliente 114 envía una petición de almacenamiento a un primer controlador de almacenamiento 152a que actúa como un controlador de RAID y un segundo cliente 114 envía una segunda petición de almacenamiento a un dispositivo de almacenamiento 150a que actúa como un segundo controlador de RAID y ambas peticiones de almacenamiento acceden a los mismos datos, el primer dispositivo de almacenamiento 150a puede comenzar la ejecución de la petición de almacenamiento sobre el primer dispositivo de almacenamiento 150a y a continuación los otros dispositivos de almacenamiento 150b - n del grupo de RAID. Al mismo tiempo, el segundo controlador de RAID sobre el segundo dispositivo de almacenamiento 150b puede comenzar la ejecución de la segunda petición de almacenamiento sobre otro dispositivo de almacenamiento (por ejemplo, 150b) y a continuación sobre los restantes dispositivos de almacenamiento 150a, c - n del grupo de RAID. Esta discordancia en la ejecución se puede causar por la distancia física entre los dispositivos de almacenamiento 150, discrepancias del tiempo de ejecución, etc. El resultado pueden ser datos corrompidos.

Lo que se necesita es un sistema, aparato, y método para una RAID distribuida, del extremo frontal compartida que maneje las peticiones de almacenamiento concurrentes que buscan acceso de los mismos datos. Ventajosamente, el sistema, aparato, y método controlarían el acceso a los datos de modo que una petición de almacenamiento se ejecuta antes de que se ejecute una segunda petición de almacenamiento.

La figura 10 es un diagrama de bloques esquemático que ilustra una realización que sirve como un sistema 1600 para una RAID distribuida del extremo frontal compartida de acuerdo con las presentes invenciones, además de la RAID progresiva y la RAID distribuida del extremo frontal. Las descripciones anteriores para los componentes representados en la Figura 10 relacionados con la RAID progresiva y la RAID distribuida del extremo frontal también son aplicables para una RAID distribuida del extremo frontal compartida. Como con la RAID distribuida del extremo frontal, el conjunto de dispositivos de almacenamiento 1604 forma un grupo de RAID e incluye dispositivos de almacenamiento 150 que son autónomos y capaces de recibir de forma independiente y servir peticiones de almacenamiento desde un cliente 114 sobre una red 116.

Con respecto a la RAID distribuida del extremo frontal compartida, el sistema 1600 incluye dos o más clientes 114 de modo que los dos o más clientes 114 envían cada uno una petición de almacenamiento con relación a los mismos datos. Las peticiones de almacenamiento son concurrentes ya que las peticiones de almacenamiento llegan de modo que una petición de almacenamiento no está completada antes de la llegada de la otra petición de almacenamiento. Entre los dispositivos de almacenamiento 150 dentro del conjunto de dispositivos de almacenamiento 1604, uno o más se designan como dispositivos de almacenamiento de paridad-espejo 1602 para una banda. Típicamente, el uno o más dispositivos de almacenamiento de paridad-espejo 1602 funcionan de forma sustancialmente similar a los otros dispositivos de almacenamiento 150.

En configuraciones típicas donde los dispositivos de almacenamiento de paridad-espejo designados 1602 alternan entre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604, los dispositivos de almacenamiento de paridad-espejo 1602 tienen esencialmente las mismas características que los otros dispositivos de almacenamiento 150 porque deben operar también como dispositivos de almacenamiento no de paridad-espejo. Las características similares son con respecto a la operación dentro de un grupo de RAID y de operación autónoma para la comunicación del cliente independiente 114 como se ha descrito anteriormente. En diversas realizaciones, los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 pueden diferir en otros aspectos no relacionados con el funcionamiento dentro del entorno de RAID descrito.

Los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 pueden ser independientes, agrupados dentro de uno o más servidores 112, puede residir cada uno en un servidor 112, se pueden acceder a través de uno o más servidores 112, etc. Uno o más clientes 114 pueden residir en servidores 112 que incluyen uno más dispositivos de almacenamiento 150, pueden residir en servidores separados 112, pueden residir en ordenadores, estaciones de trabajo, ordenadores portátiles, etc. que acceden a los dispositivos de almacenamiento 150 a través de la red de ordenadores 116 o similares.

En una realización, la red 116 incluye un bus del sistema y uno o más dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604 que comunican usando el bus del sistema. Por ejemplo, el bus del sistema puede ser un bus PCI-e, un bus de Conexión de Tecnología Avanzada Serie ("ATA serie"), un bus ATA en paralelo o similares. En otra realización, el bus del sistema es un bus externo tal como una interfaz de un pequeño sistema de ordenadores ("SCSI"), FireWire, Fiber Channel, USB, PCIe-AS, Infiniband o similares. Un experto en la materia apreciará otras configuraciones de sistema 1600 con dispositivos de almacenamiento 150 que son autónomos y son capaces de recibir de forma independiente y servir peticiones de almacenamiento desde un cliente 114 sobre una red 116.

La figura 13 es un diagrama de bloque esquemático que ilustra una realización de un aparato 2300 para una RAID distribuida del extremo frontal compartida de acuerdo con la presente invención. El aparato 2300 incluye, en diversas realizaciones, un módulo receptor de múltiples peticiones de almacenamiento 2302, un módulo de desmontaje 2304, un módulo de paridad-espejo 2306, un módulo secuenciador 2308, un módulo de validación de maestro 2310, un

módulo de determinación del maestro 2312, un módulo de error de maestro 2314, un módulo de generación de paridad 2316 y un módulo de alternancia de paridad 2318, que se describen a continuación.

El aparato 2300 incluye un módulo receptor de múltiples peticiones de almacenamiento 2302 que recibe al menos dos peticiones de almacenamiento desde al menos dos clientes 114 para almacenar datos en los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604. Los datos incluyen datos de un fichero o de un objeto. Las peticiones de almacenamiento concernientes al aparato 2300 tienen cada una al menos una porción de los datos en común y, además, son peticiones de almacenamiento concurrentes que llegan de modo que una petición de almacenamiento no se ha completado antes de la llegada de las peticiones de almacenamiento. Estas peticiones de almacenamiento concurrentes corren el riesgo de corromper datos comunes en el sistema de RAID distribuido del extremo frontal 1600. En una realización, las peticiones de almacenamiento concurrentes pueden proceder de un cliente 114. En otra realización las peticiones de almacenamiento concurrentes proceden de dos o más clientes 114.

Las múltiples peticiones de almacenamiento pueden actualizar uno o más segmentos de datos almacenados sobre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 donde los datos almacenados anteriormente se desmontan por el módulo de desmontaje 2304 en segmentos de datos almacenados sobre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604. En una realización, una petición de almacenamiento escribe los datos para la primera vez al grupo de RAID. En este caso, los datos usualmente habrían existido en otra parte y accedidos por uno o más clientes 114 y a continuación una petición de almacenamiento copia los datos al grupo de RAID mientras que otra petición de almacenamiento accede de forma concurrente a los datos.

Las múltiples peticiones de almacenamiento pueden comprender una petición de actualizar uno o más segmentos de datos almacenados sobre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 y uno o más peticiones de lectura con objetivo de al menos una porción de los datos en común. Si la petición de actualización no se completa, entonces las respuestas a las peticiones de lectura desde los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 pueden estar comprendidas por una combinación de datos preexistentes y actualizados que corrompen los datos.

El aparato 2300 incluye un módulo de desmontaje 2304 que calcula, para cada una de las peticiones de almacenamiento concurrentes, un patrón de bandas para los datos y escribe N segmentos de datos de una banda a N dispositivos de almacenamiento 150a - n dentro del conjunto de dispositivos de almacenamiento 1604. El patrón de bandas incluye una o más bandas y cada banda incluye un conjunto de N segmentos de datos. Cada uno de los N segmentos de datos se escribe a un dispositivo de almacenamiento separado 150 dentro del conjunto de dispositivos de almacenamiento 1604 y asignado a la banda. El aparato 2300 incluye un módulo de paridad-espejo 2306 que escribe, para cada una de las peticiones de almacenamiento concurrentes, un conjunto de N segmentos de datos de la banda a los dispositivos de almacenamiento 150 dentro del conjunto de dispositivos de almacenamiento 1604 designados como dispositivos de almacenamiento de paridad-espejo 1602. Los dispositivos de almacenamiento de paridad-espejo 1602 son en adición a los N dispositivos de almacenamiento 150a - n.

El módulo de desmontaje 2304 también se usa para calcular la identidad de uno o más dispositivos de almacenamiento 150a - n desde los que se leen uno o más segmentos de datos que son parte de un fichero u objeto.

El aparato 2300 incluye un módulo de secuenciador 2308 que asegura la terminación de una primera petición de almacenamiento desde un primer cliente 114 antes de ejecutar una segunda petición de almacenamiento desde un segundo cliente 114 donde las, al menos dos peticiones de almacenamiento concurrentes incluyen la primera y segunda peticiones de almacenamiento. En otras realizaciones, el módulo del secuenciador 2308 asegura la terminación de la primera petición de almacenamiento antes de ejecutar dos o más de otras peticiones de almacenamiento concurrentes. Ventajosamente, el módulo secuenciador 2308 facilita una ejecución ordenada de las peticiones de almacenamiento concurrentes para impedir la corrupción de datos. En una realización, el módulo secuenciador 2308 coordina la ejecución de las peticiones de almacenamiento concurrentes usando un controlador maestro al que deben acceder todas las peticiones de almacenamiento para los datos, usando un sistema de bloqueo, una confirmación de dos fases, u otros medios conocidos para los expertos en la materia. Algunos de los métodos usados por el módulo secuenciador 2308 se tratan a continuación.

En una realización, el módulo secuenciador 2308 asegura la terminación de la primera petición de almacenamiento antes de la ejecución de las peticiones de almacenamiento concurrentes recibiendo una confirmación desde cada uno de los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 que reciben una petición de almacenamiento en conjunción con la primera petición de almacenamiento antes de la ejecución de la segunda petición de almacenamiento. Usualmente, una confirmación confirma la terminación de una petición de almacenamiento. En una realización, cada uno de los dispositivos de almacenamiento 150 afectados por la petición de almacenamiento se escribe a cada uno de los dispositivos de almacenamiento 150 y se recibe una confirmación desde los mismos antes de que el módulo secuenciador 2308 comience la ejecución de una segunda petición de almacenamiento

La terminación de una petición de almacenamiento, en una realización puede incluir la terminación de una porción de una primera petición de almacenamiento dirigida a un dispositivo de almacenamiento único (por ejemplo, 150a) antes de la ejecución de una porción de una segunda petición de almacenamiento pendiente sobre el mismo dispositivo de almacenamiento 150a. El módulo secuenciador 2308 puede verificar de forma independiente la terminación de las porciones de la petición de almacenamiento sobre los dispositivos de almacenamiento 150. En esta realización, la escritura de segmentos de datos relacionados con una primera petición de almacenamiento no necesita mantenerse hasta que todos los segmentos de datos de la primera petición de almacenamiento se han completado. El módulo secuenciador 2308 puede coordinar las diversas ejecuciones que tienen lugar sobre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 para asegurar que los datos no se corrompen.

En una realización, la confirmación de la terminación de una petición de almacenamiento se recibe después de que el módulo de desmontaje 2304 y el módulo de paridad-espejo 2306 escriben cada uno los segmentos de datos relacionados con la petición de almacenamiento a los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604. En otra realización, la confirmación de la terminación de una petición de almacenamiento se recibe después de que el módulo de desmontaje 2304 y el módulo de paridad-espejo 2306 escriben cada uno los segmentos de datos relacionados con la petición de almacenamiento a los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 y cada uno de los dispositivos de almacenamiento 150, 1602 confirma que se han escrito los segmentos de datos.

En una realización, el módulo secuenciador 2308 selecciona una primera petición de almacenamiento para su ejecución seleccionando una petición de almacenamiento de entre las peticiones concurrentes que llegan primero. En otra realización, el módulo secuenciador 2308 selecciona una primera petición de almacenamiento para su ejecución seleccionando una petición de almacenamiento con el sello temporal más temprano. En otra realización, el módulo secuenciador 2308 selecciona una primera petición de almacenamiento para su ejecución seleccionando una petición de almacenamiento que usa algún criterio de selección. Por ejemplo, el módulo secuenciador 2308 puede seleccionar una petición de almacenamiento que se marca de alguna forma por un cliente solicitante 114 como de alta prioridad, se puede seleccionar una petición de almacenamiento de un cliente favorecido 114, etc. Un experto en la materia reconocerá otros modos en los que el módulo secuenciador 2308 puede seleccionar una primera petición de almacenamiento usando algunos criterios de selección.

En una realización, el módulo receptor de múltiples peticiones de almacenamiento 2302, el módulo de desmontaje 2304, el módulo de paridad-espejo 2306, y el módulo secuenciador 2308 son parte de un controlador maestro (no mostrado) que controla y sirve las peticiones de almacenamiento concurrentes. Todo o parte del controlador maestro puede residir y operar en un cliente 114, un dispositivo de gestión de RAID de terceras partes, un dispositivo de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604, o un controlador de almacenamiento 152 en un dispositivo de almacenamiento 150. Usando un controlador maestro para ejecutar peticiones de servicio para los datos, el módulo secuenciador 2308 puede ser conocedor de las peticiones de almacenamiento dirigidas a los datos y puede reconocer entonces las peticiones de almacenamiento concurrentes y puede secuenciar a continuación las peticiones de almacenamiento concurrentes de modo que los datos almacenados en los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 no se corrompen. Un experto en la materia reconocerá otras implementaciones de un controlador maestro que controla el servicio de una petición de almacenamiento dirigida a los datos.

En una realización, el controlador maestro es parte de un grupo de dos o más controladores maestros capaces de servir a las peticiones de almacenamiento concurrentes desde uno o más clientes 114 donde las peticiones de almacenamiento se dirigen en los datos almacenados sobre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604. Por ejemplo, un controlador maestro puede servir peticiones de almacenamiento para un primer cliente 114 y un segundo controlador maestro puede servir las peticiones de almacenamiento para un segundo cliente 114. El primer y el segundo clientes 114 pueden ambos tener acceso a los datos almacenados sobre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604, posibilitando de este modo peticiones de almacenamiento concurrentes. Un controlador maestro puede ser parte de un dispositivo de almacenamiento 150a mientras que el otro controlador maestro puede ser parte de un segundo dispositivo de almacenamiento 150b. En otra realización, el primer controlador maestro puede ser parte de primer conjunto de dispositivos de almacenamiento 1604a y el segundo controlador maestro puede ser parte de un conjunto de dispositivos de almacenamiento de espejo 1604b

Donde el controlador maestro es parte de un grupo de controladores maestros que acceden a los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604, el aparato 2300 puede incluir un módulo de validación maestro 2310 que confirma que un controlador maestro que sirve una petición de almacenamiento recibida está controlando la ejecución de la petición de almacenamiento por delante de la ejecución de una o más peticiones de almacenamiento concurrentes antes de la ejecución de la petición de almacenamiento recibida. En esta realización, las peticiones de almacenamiento concurrentes se reciben por otros controladores maestros y la petición de servicio tiene al menos una porción de los datos en común con las peticiones de almacenamiento concurrentes recibidas por los otros controladores maestros.

Por ejemplo, un controlador maestro puede recibir una petición de almacenamiento y el módulo de validación del maestro 2310 puede sondear a continuación a los otros controladores maestros antes de la ejecución de la petición de almacenamiento para verificar que el controlador maestro es aún el controlador maestro para los datos de la petición de almacenamiento. Parte de la validación puede incluir la verificación de que los controladores maestros pueden comunicar con cada uno de los otros de modo que el controlador maestro designado se verifica antes de la ejecución de la petición de almacenamiento. Esto puede ser útil en situaciones donde un controlador de RAID del extremo frontal se designa como maestro y otros de respaldo. En otro ejemplo, un controlador maestro puede recibir una petición de almacenamiento para leer un segmento de datos desde un fichero u objeto y el módulo de validación del maestro 2310 puede sondear a continuación a los otros controladores maestros para verificar que no hay ninguna actualización en marcha para el fichero u objeto. En otro ejemplo, un controlador maestro puede usar el módulo de validación del maestro 2310 para adquirir el control de los datos para la petición de almacenamiento.

Un modo de verificar que un controlador maestro es aún el maestro para la ejecución de una petición de almacenamiento es usar un esquema de sondeo de tres vías donde dos dispositivos / controladores deben estar disponibles para votar qué controlador es el maestro para que continúe una petición de almacenamiento. El esquema usa un dispositivo (no mostrado) que es una tercera parte para los controladores que compiten para ser maestros y mantiene un registro de qué controlador está asignado para ser maestro. Este dispositivo de verificación del maestro puede ser otro controlador, puede ser un cliente 114 sobre un servidor, etc. y es capaz de comunicar con los controladores en el grupo que pueden actuar como controlador maestro. Una porción del módulo de validación del maestro 2310 puede residir entonces sobre el dispositivo de verificación del maestro con una porción del módulo de verificación del maestro 2310 residiendo en cada controlador.

En un ejemplo, el sistema 1600 incluye un primer controlador de RAID distribuida del extremo frontal ("primer controlador"), un segundo controlador de la RAID distribuida del extremo frontal ("segundo controlador"), cada uno de los cuales puede ser el maestro, y un dispositivo de verificación del maestro separado. Los controladores primero y segundo y el dispositivo de verificación del maestro están todos en comunicación entre sí. El módulo de verificación del maestro 2310 puede designar al primer controlador como controlador maestro y el segundo controlador como de respaldo para los datos almacenados sobre los dispositivos de almacenamiento 150 del conjunto de dispositivos de almacenamiento 1604 y el módulo de verificación del maestro 2310 puede almacenar esta información del maestro sobre los controladores y el dispositivo de verificación del maestro. Siempre que se mantenga la comunicación entre el primer controlador, el segundo controlador y el dispositivo de verificación del maestro, el módulo de verificación del maestro 2310 puede confirmar que el primer controlador es el maestro.

Si el primer controlador maestro recibe una petición de almacenamiento y el segundo controlador de respaldo pasa a no disponible o se pierde la comunicación con el primer controlador y el dispositivo de verificación del maestro, el módulo de validación del maestro 2310 puede verificar mediante la comunicación entre el dispositivo de verificación del maestro y el primer controlador maestro que el primer controlador es aún maestro, y como tanto el primer controlador como el dispositivo de verificación del maestro confirman que el primer controlador es en efecto el controlador maestro, el módulo de validación del maestro 2310 puede permitir continuar la petición de almacenamiento. Las peticiones de almacenamiento recibidas por el segundo controlador de respaldo no continuarían porque a través del módulo de verificación de maestro 2310, reconoce que el segundo controlador no es el maestro.

Si por otra parte, el primer controlador maestro no está disponible o no puede comunicar con el segundo controlador de respaldo y el dispositivo de verificación del maestro y el segundo controlador de respaldo recibe una petición de almacenamiento, el módulo de validación del maestro 2310 puede reconocer que tanto el segundo controlador como el módulo de verificación del maestro no pueden comunicar con el primer controlador y el módulo de verificación del maestro 2310 puede designar al segundo controlador de respaldo para que sea el maestro y pueda continuar la petición de almacenamiento. El cambio de designación del maestro se puede grabar sobre el segundo controlador.

Si el primer controlador está operativo y tiene simplemente la comunicación perdida con el segundo controlador y el dispositivo de verificación de maestro, cualesquiera peticiones de almacenamiento para los datos recibidos por el primer controlador no se ejecutarán. Si se restaura la comunicación, el primer controlador no ejecutará aún las peticiones de almacenamiento porque tanto el segundo controlador como el dispositivo de verificación del maestro reconocen al segundo controlador como el maestro. Por supuesto esta designación de maestro se puede reiniciar. Un experto en la materia reconocerá una diversidad de medios estáticos y dinámicos para asignar y reasignar la designación del maestro a uno de los controladores maestros.

Si el dispositivo de verificación del maestro no está disponible y el primer controlador de almacenamiento recibe una petición de almacenamiento, la porción del módulo de verificación del maestro 2310 que opera sobre los controladores primero y segundo puede verificar que el primer controlador es el maestro y puede continuar la petición de almacenamiento. Si el segundo controlador recibe una petición de almacenamiento, la porción del módulo de verificación de maestro 2310 que opera sobre los controladores primero y segundo puede verificar que el primer controlador es el maestro y la petición de almacenamiento no continuará. En otras realizaciones, más de dos controladores son parte de un esquema de sondeo. Un experto en la materia reconocerá otros modos con los que el

módulo de verificación del maestro 2310 puede verificar que un controlador es el maestro antes de la ejecución de una petición de almacenamiento.

En otra realización, el aparato 2300 incluye un módulo de determinación del maestro 2312. Antes de enviar una petición de almacenamiento, el módulo de determinación del maestro 2312 envía una petición de determinación del maestro al grupo de controladores maestros. El grupo de controladores maestros identifica a continuación qué controlador está designado como maestro para una petición de almacenamiento y envía de vuelta una respuesta que identifica el controlador maestro para el módulo de determinación del maestro 2312. El módulo de determinación del maestro 2312 recibe la identificación del controlador maestro para la petición de almacenamiento y la dirige al dispositivo solicitante para que envíe la petición de almacenamiento al controlador maestro designado. En una realización, el módulo de determinación del maestro 2312 reside y opera en un cliente 114. En otra realización, el módulo de determinación del maestro 2312 reside y se ejecuta en un dispositivo de gestión de RAID de terceras partes. En otra realización, el módulo de determinación del maestro 2312 reside en un dispositivo de almacenamiento 150. En otra realización, el módulo de determinación del maestro 2312 está distribuido entre dos o más dispositivos de almacenamiento 150.

En una realización adicional, el aparato 2300 incluye un módulo de error de maestro 2314 que devuelve una indicación de error. El módulo de error de maestro 2314 devuelve una indicación de error, en una realización, si un módulo receptor de múltiples peticiones de almacenamiento 2302 que está controlado por un controlador maestro recibe una petición de almacenamiento no controlada por el controlador maestro.

En otra realización, el módulo de error de maestro 2314 devuelve una indicación de error si el módulo de determinación del maestro 2312 o el módulo de validación del maestro 2310 determina que el controlador maestro ya no es el maestro determinado en el momento de la terminación de la ejecución de la petición de almacenamiento. Esta realización usualmente ocurre cuando un controlador maestro comienza la ejecución de una petición de almacenamiento y pierde la comunicación con otros controladores maestros del grupo, o, en un esquema de sondeo, pierde la comunicación con los otros controladores maestros y el dispositivo de verificación del maestro. En otra realización, el módulo de error de maestro 2314 devuelve una indicación de error si un módulo receptor de múltiples peticiones de almacenamiento 2302 controlado por un controlador maestro recibe una petición de almacenamiento que no está controlada por el controlador maestro.

En otra realización, el controlador maestro controla las peticiones de almacenamiento a uno o más controladores maestros secundarios. Los controladores maestros secundarios controlan cada uno las peticiones de almacenamiento para los datos almacenados sobre los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604. En otra realización, el controlador maestro que controla el controlador maestro secundario es también un controlador maestro secundario para las peticiones de almacenamiento dirigidas para los datos almacenados sobre los dispositivos de almacenamiento 150, 1602 del conjunto de dispositivos de almacenamiento 1604.

En otra realización, el controlador maestro controla las peticiones de almacenamiento a uno o más controladores maestros secundarios y cada uno de los controladores maestros secundarios controla las peticiones de almacenamiento para los datos almacenados sobre los dispositivos de almacenamiento 150 de un conjunto de dispositivos de almacenamiento único para el controlador maestro secundario. El aparato 2300 es flexible de modo que cualquiera de los controladores maestros puede ser maestro para los otros controladores que actúan como controladores maestros secundarios. Algunos controladores maestros secundarios pueden compartir un conjunto de dispositivos de almacenamiento 1604 y otros pueden controlar un conjunto de dispositivos de almacenamiento diferente. En otras realizaciones, el controlador maestro puede ser un dispositivo de almacenamiento de paridad-espejo 1602 o uno de los N dispositivos de almacenamiento 150a - n.

En otra realización, un controlador maestro secundario se puede convertir a controlador maestro cuando el controlador maestro está fuera de línea o es incapaz de determinar que es el maestro designado. Un experto en la materia reconocerá una diversidad de medios estáticos y dinámicos para asignar y reasignar la designación de maestro de entre uno o más controladores maestros secundarios,

En una realización preferida, el aparato 2300 incluye un módulo de generación de paridad 2316 que calcula un segmento de datos de paridad para la banda y almacena el segmento de datos de paridad sobre un dispositivo de almacenamiento de paridad-espejo 1602. La banda de paridad se calcula a partir del conjunto de N segmentos de datos sobre el dispositivo de almacenamiento de paridad-espejo 1602. Esta realización es típica de RAID 5, RAID 6 o algunos otros niveles de RAID, pero usualmente no está incluida para RAID 0, RAID 1, RAID 10, etc.

En otra realización preferida, el aparato 2300 incluye un módulo de alternancia de paridad 2318 que alterna, para cada banda, cuál de los dispositivos de almacenamiento 150 dentro del conjunto de dispositivos de almacenamiento 1604 están asignados para ser uno o más dispositivos de almacenamiento de paridad- espejo 1602 para la banda. La rotación de los segmentos de datos de paridad por banda mejora el funcionamiento. El módulo de alternancia de paridad 2318 se puede usar en conjunción con el módulo de desmontaje 2304 para calcular la identidad de uno más

dispositivos de almacenamiento 150a - n desde los cuales se lee, escribe o actualiza uno o más segmentos de datos que son parte de un fichero u objeto.
Las funciones de los diversos módulos 2302- 2318 pueden residir juntas en un controlador maestro único o pueden estar distribuidas entre uno o más clientes 114, los dispositivos de gestión de RAID de terceras partes, y uno o más dispositivos de almacenamiento 150, 1602. Un experto en la técnica reconocerá diversas realizaciones donde las funciones descritas en este documento están distribuidas.

La Figura 14 es un diagrama de flujo esquemático que ilustra una realización de un método 2400 para una RAID distribuida del extremo frontal compartida de acuerdo con la presente invención. El método 2400 comienza 2402 y el módulo receptor de múltiples peticiones de almacenamiento 2302 recibe 2402 al menos dos peticiones de almacenamiento desde al menos dos clientes 114 para leer o almacenar datos en uno o más dispositivos de almacenamiento 150 de un conjunto de dispositivos de almacenamiento 1604. Los datos proceden de un fichero o de un objeto y las peticiones de almacenamiento tienen cada una al menos una porción de los datos en común y son concurrentes de modo que una petición de almacenamiento no está completa antes de la llegada de otra de las al menos dos peticiones de almacenamiento. El módulo de desmontaje 2304 calcula 2406 un patrón de banda para los datos, donde el patrón de bandas incluye una o más bandas y cada banda incluye un conjunto de N segmentos de datos. El módulo de desmontaje 2304 también lee o escribe 2408 los N segmentos de datos de una banda para los N dispositivos de almacenamiento 150a - n dentro del conjunto de dispositivos de almacenamiento 1604 cuando cada uno de los N segmentos de datos se escribe a un dispositivo de almacenamiento separado 150 o se lee desde el mismo.

Cuando la petición de almacenamiento es una operación de escritura, el módulo de paridad-espejo 2306 escribe 2410 un conjunto de N segmentos de datos de la banda a uno o más dispositivos de almacenamiento de paridad-espejo 1602 dentro del conjunto de dispositivos de almacenamiento 1604, donde los dispositivos de almacenamiento de paridad-espejo 1602están en adición a los N dispositivos de almacenamiento 150a - n. El módulo de paridad-espejo 2306 también puede leer 2410 los segmentos de datos almacenados en los dispositivos de paridad-espejo 1602 o un segmento de datos de paridad. El módulo secuenciador 2308 asegura 2412 la terminación de una primera petición de almacenamiento desde un primer cliente 114 antes de ejecutar una segunda petición de almacenamiento desde un segundo cliente 114 y el método 2400 termina 2416. Las peticiones de almacenamiento primera y segunda son peticiones de almacenamiento concurrentes.

La presente invención se puede realizar en otras formas específicas sin apartarse de su espíritu o características esenciales. Las realizaciones descritas se considerarán en todos los aspectos solo como ilustrativas y no como restrictivas. El ámbito de la invención se indica, por lo tanto por las reivindicaciones adjuntas más que por la descripción anterior.

ALMACENAMIENTO DE ESTADO SÓLIDO COMO MEMORIA CACHÉ PARA UN ALMACENAMIENTO NO VOLÁTIL DE ALTA CAPACIDAD

En general, la memoria caché es ventajosa porque los datos que se acceden a menudo o que se cargan como parte de una aplicación o sistema operativo se pueden almacenar en memoria caché con un acceso posterior mucho más rápido que cuando los datos se acceden a través de un dispositivo de almacenamiento no volátil de alta capacidad ("HCNV"), tal como una unidad de disco duro ("HDD"), una unidad óptica, un almacenamiento de cinta, etc. La memoria caché usualmente está incluida en un ordenador.

Algunos dispositivos de almacenamiento y sistemas incluyen memoria caché en los dispositivos de almacenamiento HCNV. Algunos dispositivos de almacenamiento HCNV contienen memoria caché de estado sólido no volátil; esto proporciona el beneficio de la reducción de tiempos de acceso pero solo puede proporcionar un funcionamiento consistente con la capacidad usualmente limitada de la interfaz del dispositivo de almacenamiento HCNV. Existen algunos dispositivos de almacenamiento de la memoria caché de estado sólido no volátil que están típicamente situados en la placa base; estos dispositivos no se pueden usar en entornos multi-cliente ya que no se proporciona la coherencia de la memoria caché. Algunos controladores de dispositivos de HCNV también incluyen memoria caché. Cuando se comparten los controladores de la memoria caché de HCNV redundantes entre múltiples clientes, se requieren algoritmos sofisticados de coherencia de la memoria caché para asegurar que los datos no se corrompen.

Usualmente, las memorias caché se implementan en DRAM, hacen de la capacidad de la memoria caché un objetivo primordial, y requieren potencia relativamente alta para el funcionamiento. Si se pierde la potencia que soporta la memoria caché volátil se pierden los datos almacenados en la memoria caché. Usualmente, se usa alguna batería de respaldo para evitar que los datos de pierdan en el caso de un fallo de potencia, con capacidad suficiente para descargar la memoria caché a una memoria no volátil antes de que falle la batería de respaldo. Además, los sistemas de batería de respaldo consumen potencia, requieren redundancia, impactan negativamente en la fiabilidad y consumen espacio. Las baterías también se tienen que mantener sobre una base regular y la batería de respaldo puede ser relativamente cara.

A partir de la discusión anterior, debería ser evidente que existe una necesidad de un aparato, sistema y método que gestionen los datos usando un almacenamiento de estado sólido como memoria caché. Ventajosamente, tal aparato, sistema, y método proporcionarían una memoria caché no volátil que consume poca potencia, proporciona una capacidad significativamente mayor y no requiere una batería de respaldo para mantener los datos almacenados en la memoria caché.

La Figura 15 es un diagrama de bloques esquemático que ilustra una realización de un sistema 3400 con almacenamiento de estado sólido 110 como memoria caché para un dispositivo de almacenamiento no volátil de alta capacidad 3404 de acuerdo con la presente invención. El sistema 3400 incluye un dispositivo de almacenamiento de estado sólido 102 con un controlador de almacenamiento 152 que incluye un controlador de almacenamiento de estado sólido 104 y un controlador de HCLV 3402, almacenamiento de estado sólido 110 y una interfaz de red 156. El sistema 3400 incluye un dispositivo solicitante 155 conectado al dispositivo de almacenamiento de estado sólido 102 a través de una red de ordenadores 116 y uno o más dispositivos de almacenamiento de HVNV 3404a - n. Un experto en la materia reconocerá que el sistema 3400 representado en la Figura 15 es meramente una realización y que pueden ser posibles muchas otras configuraciones que permiten al almacenamiento de estado sólido 110 ser una memoria caché para un dispositivo de almacenamiento.

El sistema 3400 incluye un dispositivo de almacenamiento de estado sólido 102 con una interfaz de red 156 y un controlador de almacenamiento 152. En otra realización, la interfaz de red 156 está fuera del dispositivo de almacenamiento de estado sólido 102. Por ejemplo, la interfaz de red 156 puede estar en un servidor 112 que puede ser que incluya o no el dispositivo de almacenamiento de estado sólido 102.

En la realización representada, el dispositivo de almacenamiento de estado sólido 102 incluye un controlador de almacenamiento 152 que incluye un controlador del almacenamiento de estado sólido 104 y un controlador del almacenamiento no volátil de alta capacidad ("HCNV") 3402. En otra realización, el dispositivo de almacenamiento de estado sólido 102 incluye un controlador del almacenamiento de estado sólido 104 y un controlador del almacenamiento HCNV 3402 que no están en el controlador de almacenamiento 152. En otras realizaciones, el dispositivo de almacenamiento de estado sólido 102 incluye un controlador del almacenamiento de estado sólido 104 que incluye un controlador de almacenamiento HCNV 3402 o viceversa.

En la realización representada, el sistema 3400 incluye un dispositivo de almacenamiento de estado sólido 102 con un almacenamiento de estado sólido integrado 110 y con dispositivos de almacenamiento HCNV externos 3404a - n. En otra realización, los controladores de almacenamiento 152, 104, 3402 pueden estar separados del almacenamiento de estado sólido 110. En otra realización, los controladores 152, 104, 3402 y el almacenamiento de estado sólido 110 están incluidos en un dispositivo de almacenamiento HCNV 3404. El dispositivo de almacenamiento HCNV 3404 también puede incluir una interfaz de red 156. Un experto en la materia reconocerá que son posibles otras muchas configuraciones. El dispositivo de almacenamiento de estado sólido 102, el controlador del almacenamiento de estado sólido 104, el almacenamiento de estado sólido 110, el bus I/O de almacenamiento 210, la interfaz de red 156, la red de ordenadores 116 y el dispositivo solicitante 155 son sustancialmente similares a otras realizaciones de los dispositivos y bus descritos anteriormente.

En una realización, el dispositivo solicitante 155 está conectado al dispositivo de almacenamiento de estado sólido 102, el controlador de almacenamiento 152, el controlador del almacenamiento de estado sólido 104, etc. a través de un bus del sistema. Las transferencias de datos entre el dispositivo solicitante 155 y el almacenamiento de estado sólido 110 pueden ocurrir sobre el bus del sistema.

Los dispositivos de almacenamiento HCNV 3404 son típicamente dispositivos de almacenamiento de alta capacidad que proporcionan almacenamiento no volátil y son usualmente más lentos para la escritura y la lectura de datos que el almacenamiento de estado sólido 110. Los dispositivos de almacenamiento HCNV 3404 también pueden ser más baratos por unidad de capacidad de almacenamiento que el almacenamiento de estado sólido 110. Los dispositivos de almacenamiento HCNV 3404 pueden ser una unidad de disco duro ("HDD"), una unidad óptica, un almacenamiento de cinta, y similares. La provisión de almacenamiento de estado sólido 110 como memoria caché para los dispositivos de los dispositivos de almacenamiento HCNV 3404 usualmente aumenta la velocidad del acceso y almacenamiento de datos. Un experto en la materia reconocerá otros beneficios del almacenamiento de estado sólido 110 como memoria caché para un dispositivo de almacenamiento HCNV 3404.

En una realización, los dispositivos de almacenamiento HCNV 3404 están conectados al controlador de almacenamiento 152 a través de una red de área de almacenamiento ("SAN"). En una realización, un controlador de SAN separado conecta los dispositivos de almacenamiento HCNV 3404 al controlador de almacenamiento 152. En otra realización el controlador de almacenamiento HCNV 3402 o el controlador de almacenamiento 152 actúan como un controlador de SAN. Un experto en la materia reconocerá otros modos en los que los dispositivos de almacenamiento HCNV 3404 se pueden conectar en una SAN.

La Figura 16 es un diagrama de bloques esquemático que ilustra una realización de un aparato 3500 con almacenamiento de estado sólido como memoria caché para un dispositivo de almacenamiento no volátil de alta capacidad de acuerdo con la presente invención. El aparato 3500 incluye un módulo del extremo frontal de memoria

caché 3502, un módulo del extremo posterior de memoria caché 3504, un controlador de almacenamiento de objetos 3506, un módulo de HCNV 3508, y un módulo de emulación del dispositivo estándar 3510, que se describen a continuación. Los módulos 3502 - 3510 del aparato 3500 se representan en un controlador de almacenamiento 152 con un controlador del almacenamiento de estado sólido 104 y un controlador del almacenamiento HCNV 3402, pero algunos o todos de cada módulo 3502 - 3510 pueden estar incluidos en el controlador de almacenamiento de estado sólido 104, el controlador de almacenamiento HCNV 3402, un servidor 112 un dispositivo de almacenamiento de HCNV 3404 u otra localización.

El aparato 3500 incluye un módulo del extremo frontal de la memoria caché 3502 que gestiona las transferencias de datos asociadas con una petición de almacenamiento, donde las transferencias de datos son entre el dispositivo solicitante 155 y el almacenamiento de estado sólido 110 que funciona como una memoria caché para uno o más dispositivos de almacenamiento HCNV 3404a - n. El aparato 3500 también incluye un módulo del extremo posterior de la memoria caché 3504 que gestiona las transferencias de datos entre el almacenamiento de estado sólido 110 y los dispositivos de almacenamiento HCNV 3404a - n. Las transferencias de datos pueden incluir datos, metadatos y/o índices de metadatos. El almacenamiento de estado sólido 110, como se ha descrito anteriormente es una red de elementos de almacenamiento de datos de estado sólido no volátil 216, 218, 220, que están usualmente dispuestos en bancos 214. En diversas realizaciones, el almacenamiento de estado sólido 110 puede ser memoria flash, nano memoria de acceso aleatorio ("nano RAM" o "NRAM"), RAM magneto-resistiva ("MRAM"), RAM dinámica ("DRAM"), RAM de cambio de fase ("PRAM") o similares.

Usualmente el módulo del extremo frontal de la memoria caché 3502, el módulo del extremo posterior de la memoria caché 3504 y el controlador del almacenamiento de estado sólido 104 operan de forma autónoma desde el dispositivo solicitante 155. Por ejemplo, el dispositivo solicitante 155 puede ver el controlador de almacenamiento 152 con el controlador de almacenamiento de estado sólido 104 y el controlador de almacenamiento HCNV 3402, junto con el almacenamiento asociado 110, 3404a - n, como un único dispositivo de almacenamiento. En otro ejemplo, el dispositivo solicitante 155 puede ver los dispositivos de almacenamiento HCNV 3404a - n y el almacenamiento de estado sólido 110 puede ser transparente.

En una realización, el controlador de almacenamiento de estado sólido 104 incluye un módulo controlador del almacenamiento de objetos 3506 que sirve las peticiones de objetos desde uno o más dispositivos solicitantes 155 y gestiona objetos de las peticiones de objetos dentro del almacenamiento de estado sólido 110. En la realización, el controlador de estado sólido 104, junto con el módulo controlador del almacenamiento de objetos 3506, gestionan las peticiones de objetos como se ha descrito anteriormente, y en particular como se describe en relación con el aparato 200 representado en la Figura 2A.

En otra realización, el aparato 3500 incluye un módulo RAID de HCNV 3508 que almacena datos en memoria caché en el almacenamiento de estado sólido 110 en dos o más dispositivos de almacenamiento HCNV 3404a - n en una red redundante de unidades independientes ("RAID") consistente con un nivel de RAID. En la realización, los datos aparecen para un dispositivo solicitante 155 como un todo de modo que la estructuración en RAID es oculta para el dispositivo solicitante 155. Por ejemplo, el módulo del extremo frontal de la memoria caché 3502 puede almacenar en memoria caché datos procedentes del dispositivo solicitante 155 en el almacenamiento de estado sólido 110 y el módulo del extremo posterior de memoria caché 3504 puede cooperar con el módulo de RAID de HCNV 3508 para desmontar los datos y almacenar los segmentos de datos y los segmentos de datos de paridad en los dispositivos de almacenamiento HCNV 3404a - n consistentes con un nivel de RAID. Un experto en la materia reconocerá otros modos en los que los datos procedentes de un dispositivo solicitante 155se puedan estructurar en RAID en los dispositivos de almacenamiento HCNV 3404a - n.

En otra realización, el almacenamiento de estado sólido 110 y los dispositivos de almacenamiento HCNV 3404a - n comprenden un dispositivo de almacenamiento híbrido dentro de un conjunto de dispositivos de almacenamiento híbridos que está configurado como un grupo de RAID. Por ejemplo, el conjunto de dispositivos de almacenamiento híbrido puede ser un conjunto de dispositivos de almacenamiento de RAID distribuido del extremo frontal 1604 y el dispositivo de almacenamiento híbrido puede ser un dispositivo de almacenamiento 150, 1602 en el conjunto de dispositivos de almacenamiento como se ha descrito anteriormente en relación con el sistema 1600, el aparato 2100, y el método 2200 representados en las Figuras 10, 11 y 12 respectivamente. En la realización, un segmento de datos en memoria caché en el almacenamiento de estado sólido 110 y almacenados más tarde sobre un dispositivo de HCNV 3404 es uno de los N segmentos de datos de una banda o un segmento de datos de paridad de una banda. Como en la RAID del extremo frontal, el dispositivo de almacenamiento híbrido recibe peticiones de almacenamiento desde uno o más clientes 114 independientes de los segmentos de datos de una banda de RAID.

En una realización adicional, el dispositivo de almacenamiento híbrido es un dispositivo de almacenamiento 150, 1602 de un grupo de RAID distribuido del extremo frontal 1604 que recibe dos o más peticiones de almacenamiento simultáneo desde dos o más clientes 114, como se ha descrito anteriormente en relación con la RAID del extremo frontal compartida como se ha descrito en relación con el sistema 1600, el aparato 2300 y el método 2400 representados en las Figuras 10, 13 y 14, respectivamente. Ventajosamente, esta realización asegura que la memoria caché redundante, compartida mantiene la coherencia sin complejos protocolos y algoritmos adicionales de coherencia.

En otra realización, el almacenamiento de estado sólido 110 y los dispositivos de almacenamiento HCNV 3204a - n comprenden un dispositivo de almacenamiento híbrido y el aparato 3500 incluye un módulo de emulación del dispositivo estándar 3510 que proporciona acceso al dispositivo de almacenamiento híbrido emulando un dispositivo estándar conectado al uno o más dispositivos solicitantes 155 antes de cargar los dispositivos solicitantes 155 con código específico para la operación del dispositivo de almacenamiento híbrido. En la realización, el dispositivo estándar se soporta por una normativa BIOS de la industria. Esta operación de arranque permite al dispositivo híbrido su reconocimiento y acceso por los dispositivos solicitantes 155 con funcionalidad limitada hasta las unidades específicas para el controlador de almacenamiento de estado sólido 104, el controlador de almacenamiento HCNV 3402, y posiblemente los otros módulos 3502 - 3510 del aparato 3500 se puede cargar sobre los dispositivos solicitantes 155.

En una realización, el dispositivo de almacenamiento de estado sólido 110 se puede dividir en dos o más regiones donde una o más de las particiones se usan como almacenamientos de estado sólido independientes del funcionamiento del almacenamiento de estado sólido como memoria caché para los dispositivos de almacenamiento HCNV 3404. Por ejemplo, algunas particiones del almacenamiento de estado sólido 110 se pueden acceder por los clientes 114 para almacenamiento de datos general mientras que una o más particiones se dedican como memoria caché par los dispositivos de almacenamiento HCNV 3404.

En una realización, más de un cliente 114 (o dispositivo solicitante 155) es capaz de enviar mensajes de control de la memoria caché al módulo del extremo frontal de memoria caché 3502 y el módulo del extremo posterior de la memoria caché 3504 para gestionar el estado de uno o más ficheros u objetos almacenados dentro del dispositivo de almacenamiento de estado sólido 110 y el uno o más dispositivos de almacenamiento HCNV 3404. Ventajosamente, la capacidad para los clientes 114 / dispositivos solicitantes 155 para gestionar la memoria caché en base a cada fichero, o por segmento de datos proporciona una gran acuerdo de flexibilidad para compartir un dispositivo de almacenamiento de estado sólido 110.

Numerosos mensajes de control de la memoria caché son permisibles y posibles. Por ejemplo, un mensaje de control de la memoria caché puede incluir un mensaje de control que cause que el módulo del extremo posterior de la memoria caché 3504 ancle una porción de un objeto o fichero en el almacenamiento de estado sólido 110. Otro mensaje de control de la memoria caché puede incluir un mensaje de control que cause que el módulo del extremo posterior de la memoria caché 3504 desancle una porción de un objeto o fichero en el almacenamiento de estado sólido 110. Otro mensaje de control de la memoria caché puede incluir un mensaje de control que cause que el módulo del extremo posterior de memoria caché 3504 transfiera una porción de un objeto o fichero desde el almacenamiento de estado sólido 110 al uno o más dispositivos de almacenamiento HCNV 3404. Otro mensaje de control de la memoria caché puede incluir un mensaje de control que cause que el módulo del extremo posterior de la memoria caché 3504 precargue una porción de un objeto o fichero al almacenamiento de estado sólido 110 desde el uno o más dispositivos de almacenamiento HCNV 3404. Otro mensaje de control de la memoria caché puede incluir un mensaje de control que cause que el módulo del extremo posterior de la memoria caché 3504 descargue una o más porciones del uno o más objetos o ficheros desde el almacenamiento de estado sólido a los dispositivos de almacenamiento HCNV 3404 para liberar una cantidad determinada de espacio de almacenamiento en el almacenamiento de estado sólido 110. Un experto en la materia reconocerá otros posibles mensajes de control de la memoria caché.

En una realización, los mensajes de control de la memoria caché se comunican mediante metadatos ("metadatos de control de la memoria caché") para el objeto o fichero. Los metadatos de control de la memoria caché, en una realización, son persistentes. En otra realización, los metadatos de control de la memoria caché se establecen mediante un conjunto de atributos en el momento de creación del fichero o el objeto. En la realización, los atributos se pueden heredar a través de la relación para una clase de objetos particular, características por defecto de los tipos de ficheros específicos, etc. En otra realización, los metadatos de control de la memoria caché se obtienen desde un sistema de gestión de ficheros u objetos. Un experto en la materia reconocerá otros modos en los que se pueden comunicar mensajes de control de la memoria caché mediante metadatos.

En una realización, el sistema 3400 incluye un elemento de almacenamiento de la memoria caché volátil. Por ejemplo, además del almacenamiento de estado sólido 110, el sistema 3400 también puede incluir una memoria de acceso aleatoria ("RAM") de algún tipo que es volátil. En esta realización, el módulo del extremo frontal de la memoria caché 3502 y el módulo del extremo posterior de la memoria caché 3504 almacenan algunos datos en el elemento de almacenamiento de la memoria caché volátil y gestionan los datos almacenados en el almacenamiento de estado sólido 110 y el elemento de almacenamiento de memoria caché volátil y el módulo de almacenamiento del extremo posterior 3504 también gestiona las transferencias de datos entre el elemento de almacenamiento de la memoria caché volátil, el almacenamiento de estado sólido y los dispositivos de almacenamiento HCNV. Por ejemplo, los datos que no son críticos o se pueden recuperar fácilmente desde otra fuente se pueden almacenar en memoria caché volátil mientras que otros datos se pueden almacenar en el almacenamiento de estado sólido 110 funcionando como memoria caché.

En una realización adicional, los metadatos y/o metadatos de índices para objetos y ficheros almacenados en los dispositivos de almacenamiento HCNV 3404 se mantienen dentro del dispositivo de almacenamiento de estado sólido 110 y en el elemento de almacenamiento de la memoria caché volátil. Como se ha descrito anteriormente en relación con el aparato 200 representado en la Figura 2A, ciertos metadatos se pueden almacenar en un elemento de almacenamiento de la memoria caché volátil y se pueden usar para reconstruir un índice si los datos en el elemento de almacenamiento caché volátil se pierden. En una realización, los metadatos y los metadatos de índices se almacenan en el almacenamiento de estado sólido 110 donde está incluido el elemento de almacenamiento de la memoria caché no volátil. Un experto en la materia reconocerá otros beneficios y modos de usar los elementos de almacenamiento de la memoria caché volátil junto con el almacenamiento de estado sólido 110 funcionando como memoria caché.

La Figura 17 es un diagrama de flujo esquemático que ilustra una realización de un método 3600 con almacenamiento de estado sólido como memoria caché para un dispositivo de almacenamiento no volátil de alta capacidad de acuerdo con la presente invención.
El método 3600 comienza 3602 y el módulo del extremo frontal de la memoria caché 3502 gestiona 3604 las transferencias de datos asociadas con una petición de almacenamiento, donde las transferencias de datos son entre un dispositivo solicitante 155 y el almacenamiento de estado sólido 110 funcionando como memoria caché para uno o más dispositivos de almacenamiento HCNV 3404a - n. El módulo del extremo posterior de la memoria caché 3504 gestiona 3606 las transferencias de datos entre el almacenamiento de estado sólido 110 y el uno o más dispositivos de almacenamiento HCNV 3404a-n y el método 3600 termina 3608. El método 3600 opera de forma sustancialmente similar a como se describió anteriormente con relación al aparato 3500 de la Figura 16.

La presente invención se puede realizar en otras formas específicas sin apartarse de sus características esenciales. Las realizaciones descritas se considerarán en todos los aspectos solo como ilustrativas y no restrictivas. Por lo tanto el ámbito de la invención se indica por las reivindicaciones adjuntas más que por la descripción anterior.

REIVINDICACIONES

1. Un aparato para gestionar el almacenamiento de datos sobre uno o más dispositivos de almacenamiento no volátiles de alta capacidad "HCNV" (150, 3404), comprendiendo el aparato:

un controlador de almacenamiento (152) que comprende un controlador de almacenamiento de estado sólido (104) y un controlador de almacenamiento HCNV (3402);
un módulo del extremo frontal de la memoria caché (3502) que gestiona las transferencias de datos asociadas con una petición de almacenamiento, funcionando las transferencias de datos entre un dispositivo solicitante y un almacenamiento de estado sólido como memoria caché para uno o más dispositivos de almacenamiento HCNV, comprendiendo las transferencias de datos uno o más de datos, metadatos e índices de metadatos, comprendiendo el almacenamiento de estado sólido una red de elementos de almacenamientos de datos de estado sólido no volátiles;
un módulo del extremo posterior de la memoria caché (3504) que gestiona las transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV; y
un módulo de almacenamiento secuencial (802) que almacena secuencialmente segmentos de datos de las transferencias de datos en divisiones de almacenamiento del almacenamiento de estado sólido por orden de procesamiento en el orden en el que los segmentos de datos llegan desde el dispositivo solicitante.

2. El aparato de la reivindicación 1, en el que el módulo del extremo frontal de la memoria caché, el módulo del extremo posterior de la memoria caché, y el controlador de almacenamiento de estado sólido operan de forma autónoma del dispositivo solicitante.

3. El aparato de la reivindicación 1, que comprende además un módulo de RAID de HCNV (3508) que almacena los datos almacenados en memoria caché en al almacenamiento de estado sólido en dos o más dispositivos de almacenamiento HCNV en una red redundante de unidades independientes, "RAID", consistentes con un nivel de RAID, en el que los datos aparecen al dispositivo solicitante como un todo.

4. El aparato de la reivindicación 1, en el que el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV comprende un dispositivo de almacenamiento híbrido dentro de un conjunto de dispositivos de almacenamiento híbrido que se configuran como un grupo de RAID, en el que un segmento de datos almacenado en memoria caché en el almacenamiento de estado sólido y almacenado más tarde sobre un dispositivo de HCNV comprende uno de N segmentos de datos de una banda o segmentos de datos de paridad de la banda, en el que el dispositivo de almacenamiento híbrido recibe peticiones de almacenamiento desde uno o más clientes independientes de los segmentos de datos de una banda de RAID.

5. El aparato de la reivindicación 4, en el que el dispositivo de almacenamiento híbrido es un dispositivo de almacenamiento de un grupo de RAID distribuido del extremo frontal compartido que recibe dos o más peticiones simultáneas de almacenamiento desde dos o más clientes.

6. El aparato de la reivindicación 1, en el que el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV comprenden un dispositivo de almacenamiento híbrido y comprenden además un módulo de emulación de un dispositivo normalizado (3510) que proporciona acceso al dispositivo de almacenamiento híbrido emulando un dispositivo normalizado conectado al uno o más dispositivos solicitantes antes de cargar el uno o más dispositivos solicitantes con código específico para la operación del dispositivo de almacenamiento híbrido, estando soportado el dispositivo normalizado por una normativa BIOS de la industria.

7. El aparato de la reivindicación 1, en el que el dispositivo de almacenamiento de estado sólido se puede dividir en dos o más regiones, en el que una o más particiones se pueden usar como almacenamiento de estado sólido independiente del almacenamiento de estado sólido que funciona como memoria caché para los dispositivos de almacenamiento HCNV.

8. El aparato de la reivindicación 1, en el que uno o más clientes (114) envían mensajes de control de la memoria caché al módulo del extremo del extremo frontal de la memoria caché y al módulo del extremo posterior de la memoria caché para gestionar un estado de uno o más ficheros u objetos almacenados dentro del dispositivo de almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV.

9. El aparato de la reivindicación 8, en el que los mensajes de control de la memoria caché se comunican mediante metadatos, "metadatos de control de la memoria caché", para el objeto o fichero.

10. El aparato de la reivindicación 9, en el que los metadatos de control de la memoria caché son uno de: persistentes, establecidos a través del conjunto de atributos en el momento de la creación del fichero u objeto; y obtenidos desde un sistema de gestión de ficheros u objetos.

11. El aparato de la reivindicación 1, en el que uno o más de los metadatos y metadatos de índices para los objetos y ficheros almacenados en los dispositivos de almacenamiento HCNM se mantienen dentro del dispositivo de almacenamiento de estado sólido.

12. El aparato de la reivindicación 1, en el que el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV comprenden un dispositivo de almacenamiento de modo que los dispositivos de almacenamiento HCNV están ocultos de la vista de un cliente conectado al dispositivo de almacenamiento y aparecen al dispositivo solicitante como un único dispositivo de almacenamiento.

13. Un sistema para la gestión del almacenamiento de datos sobre uno o más dispositivos de almacenamiento no volátiles de alta capacidad "HCNV", comprendiendo el sistema:

un almacenamiento de estado sólido (110) que comprende una red de elementos de almacenamiento de datos de estado sólido no volátiles;
uno o más dispositivos de almacenamiento HCNV (150, 3404); y
un controlador de almacenamiento (152) que comprende
un controlador del almacenamiento de estado sólido (104);
un controlador del dispositivo de almacenamiento HCNV (3402); y
el aparato de acuerdo con la reivindicación 1.

14. El sistema de la reivindicación 13 que comprende además una interfaz de red (156) conectada al controlador de almacenamiento, facilitando la interfaz de red las transferencias de datos entre el dispositivo solicitante y el controlador del almacenamiento de estado sólido a través de una red de ordenadores.

15. El sistema de la reivindicación 13, en el que el uno o más dispositivos de almacenamiento HCNV están conectados al controlador de almacenamiento a través de una red de área de almacenamiento ("SAN").

16. Un producto de programa de ordenador que comprende un medio legible por ordenador que tiene un código de programa utilizable por el ordenador que, cuando se ejecuta sobre un procesador, realiza operaciones para la gestión del almacenamiento de datos sobre uno o más dispositivos de almacenamiento no volátiles de alta capacidad "HCNV" (150, 3404), comprendiendo las operaciones del producto de programa de ordenador:

gestionar las transferencias de datos asociadas con una petición de almacenamiento usando un módulo del extremo frontal de la memoria caché (3502), las transferencias de datos entre un dispositivo solicitante y el almacenamiento de estado sólido que funciona como una memoria caché para uno o más dispositivos de almacenamiento HCNV, comprendiendo las transferencias de datos uno o más de datos, metadatos e índices de metadatos, comprendiendo el almacenamiento de estado sólido una red de elementos de almacenamiento de datos de estados sólido no volátil;
gestionar transferencias de datos entre el almacenamiento de estado sólido y el uno o más dispositivos de almacenamiento HCNV usando un módulo del extremo posterior de la memoria caché (3504); y
almacenar secuencialmente los segmentos de datos de las transferencias de datos en divisiones de almacenamiento del almacenamiento de estado sólido por orden de procesamiento en el orden que llegan los segmentos de datos desde el dispositivo solicitante.

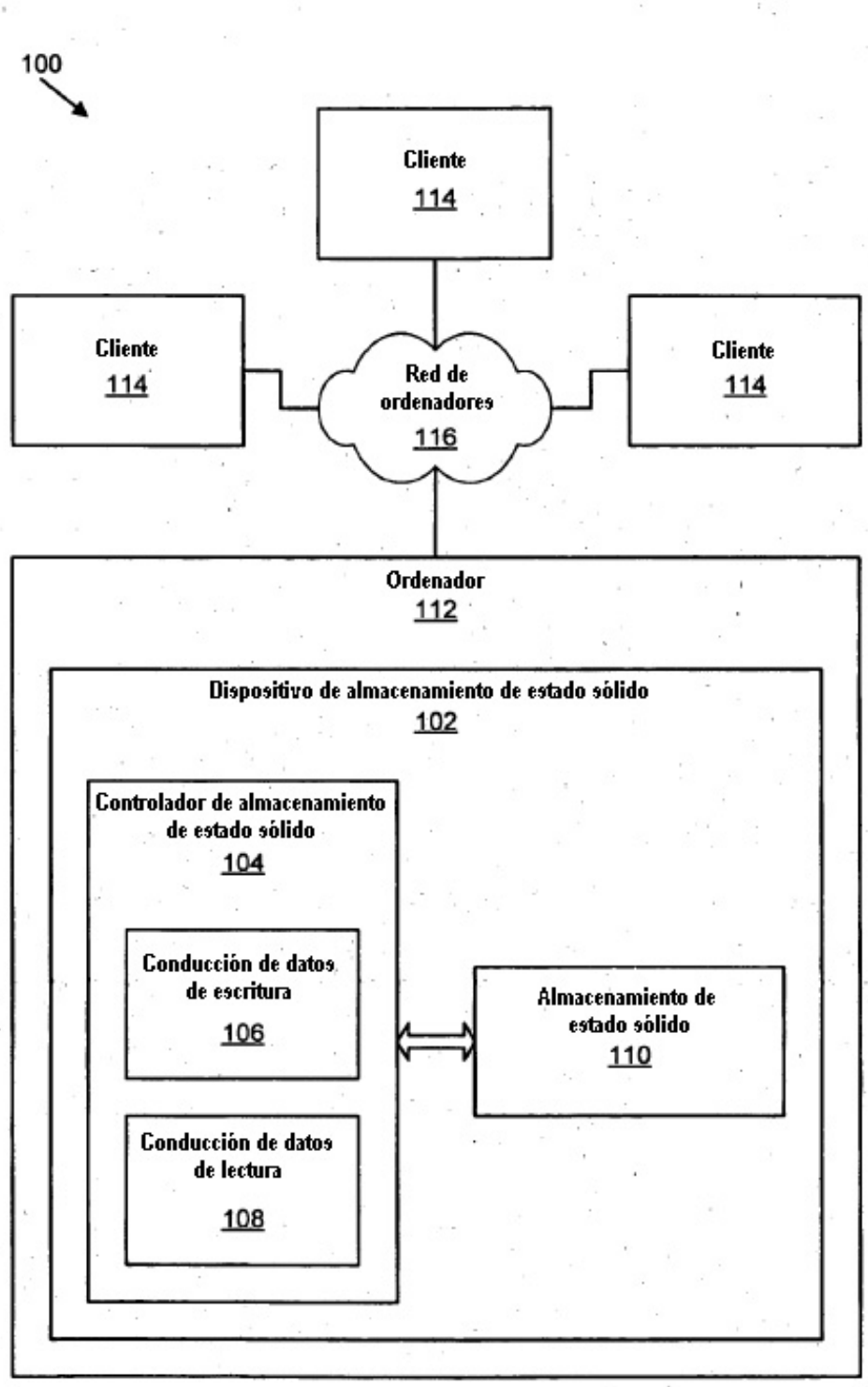
100
Cliente
114
Cliente
114
Red de ordenadores
116
Cliente
114
Ordenador
112
Dispositivo de almacenamiento de estado sólido
102
Controlador de almacenamiento de estado sólido
104
Conducción de datos de escritura
106
Conducción de datos de lectura
108
Almacenamiento de estado sólido
110

FIG. 1A

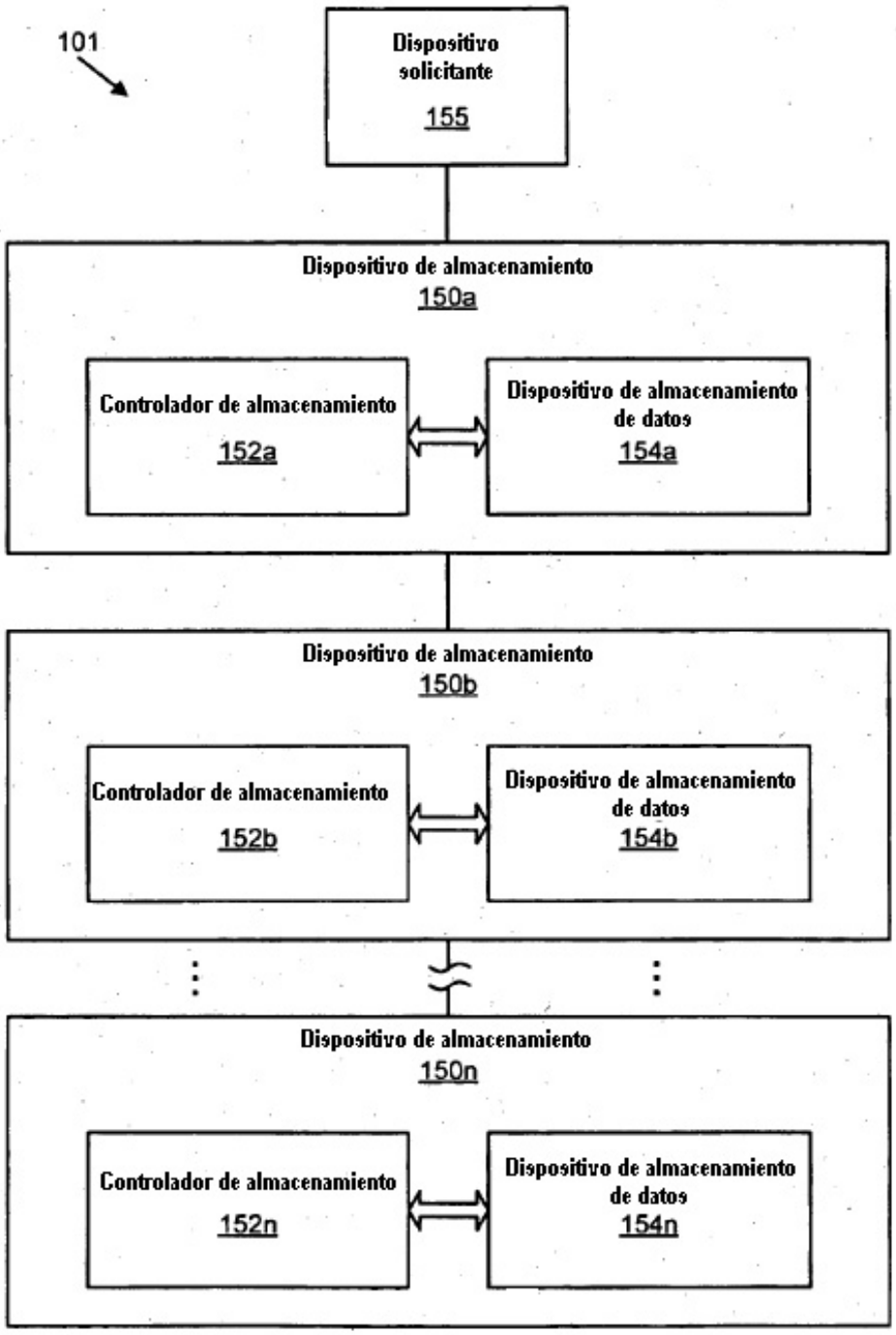
101
Dispositivo solicitante
155
Dispositivo de almacenamiento
150a
Controlador de almacenamiento
152a
Dispositivo de almacenamiento de datos
154a
Dispositivo de almacenamiento
150b
Controlador de almacenamiento
152b
Dispositivo de almacenamiento de datos
154b
Dispositivo de almacenamiento
150n
Controlador de almacenamiento
152n
Dispositivo de almacenamiento de datos
154n

FIG. 1B

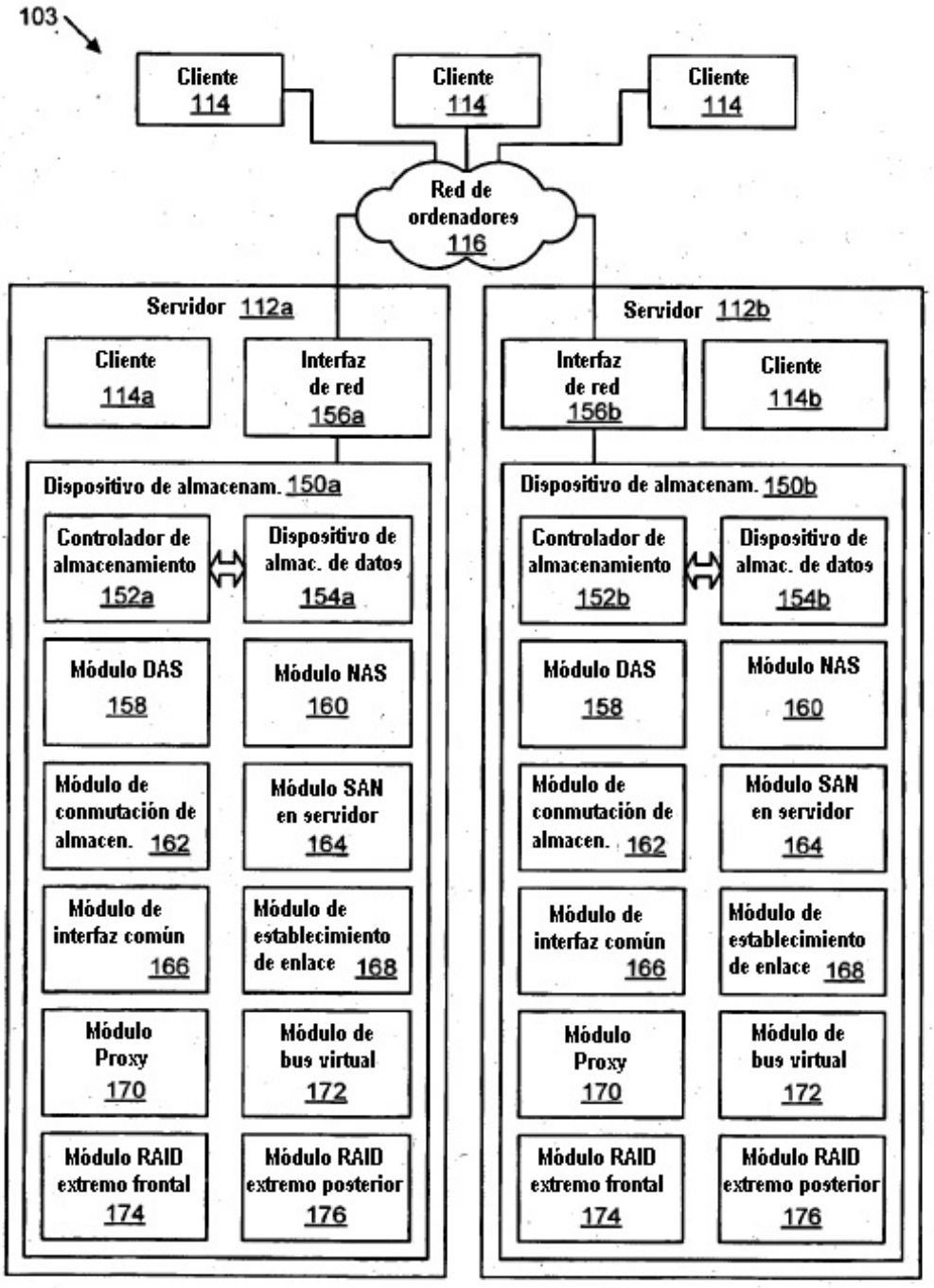
103
Cliente 114
Cliente 114
Cliente 114
Red de ordenadores 116
Servidor 112a
Cliente 114a
Interfaz de red 156a
Dispositivo de almacenam. 150a
Controlador de almacenamiento 152a
Dispositivo de almac. de datos 154a
Módulo DAS 158
Módulo NAS 160
Módulo de conmutación de almacen. 162
Módulo SAN en servidor 164
Módulo de interfaz común 166
Módulo de establecimiento de enlace 168
Módulo Proxy 170
Módulo de bus virtual 172
Módulo RAID extremo frontal 174
Módulo RAID extremo posterior 176
Servidor 112b
Interfaz de red 156b
Cliente 114b
Dispositivo de almacenam. 150b
Controlador de almacenamiento 152b
Dispositivo de almac. de datos 154b
Módulo DAS 158
Módulo NAS 160
Módulo de conmutación de almacen. 162
Módulo SAN en servidor 164
Módulo de interfaz común 166
Módulo de establecimiento de enlace 168
Módulo Proxy 170
Módulo de bus virtual 172
Módulo RAID extremo frontal 174
Módulo RAID extremo posterior 176

FIG. 1C

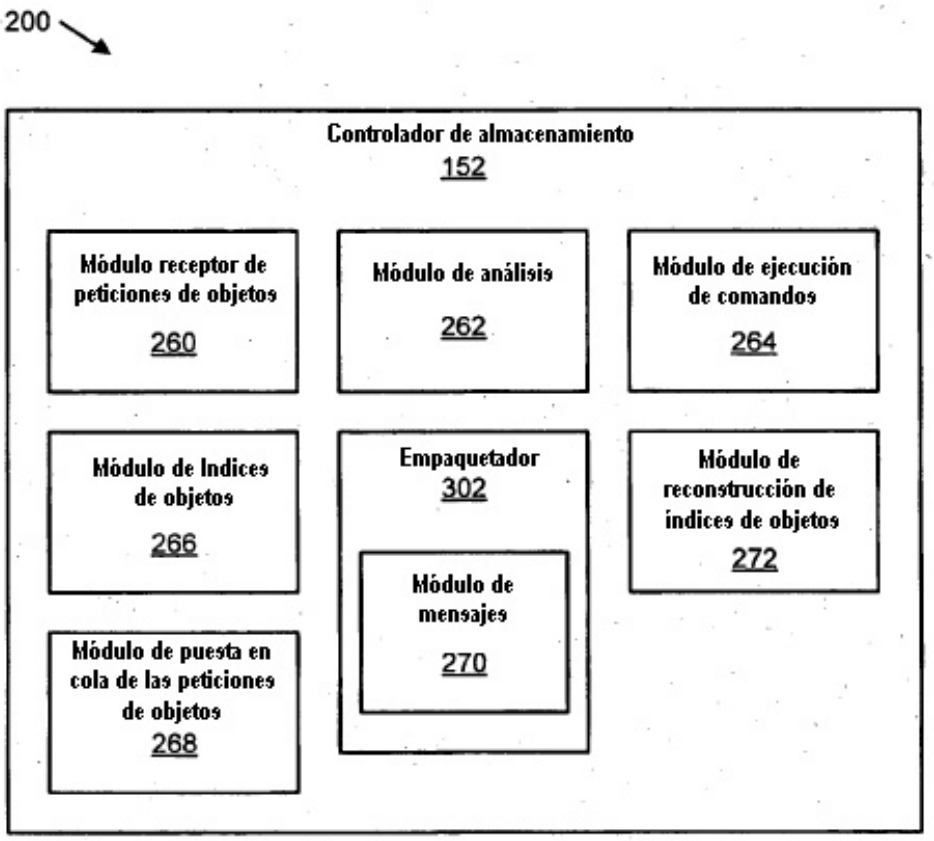
200
Controlador de almacenamiento
152
Módulo receptor de peticiones de objetos
260
Módulo de análisis
262
Módulo de ejecución de comandos
264
Módulo de Indices de objetos
266
Empaquetador
302
Módulo de mensajes
270
Módulo de reconstrucción de índices de objetos
272
Módulo de puesta en cola de las peticiones de objetos
268

FIG. 2A

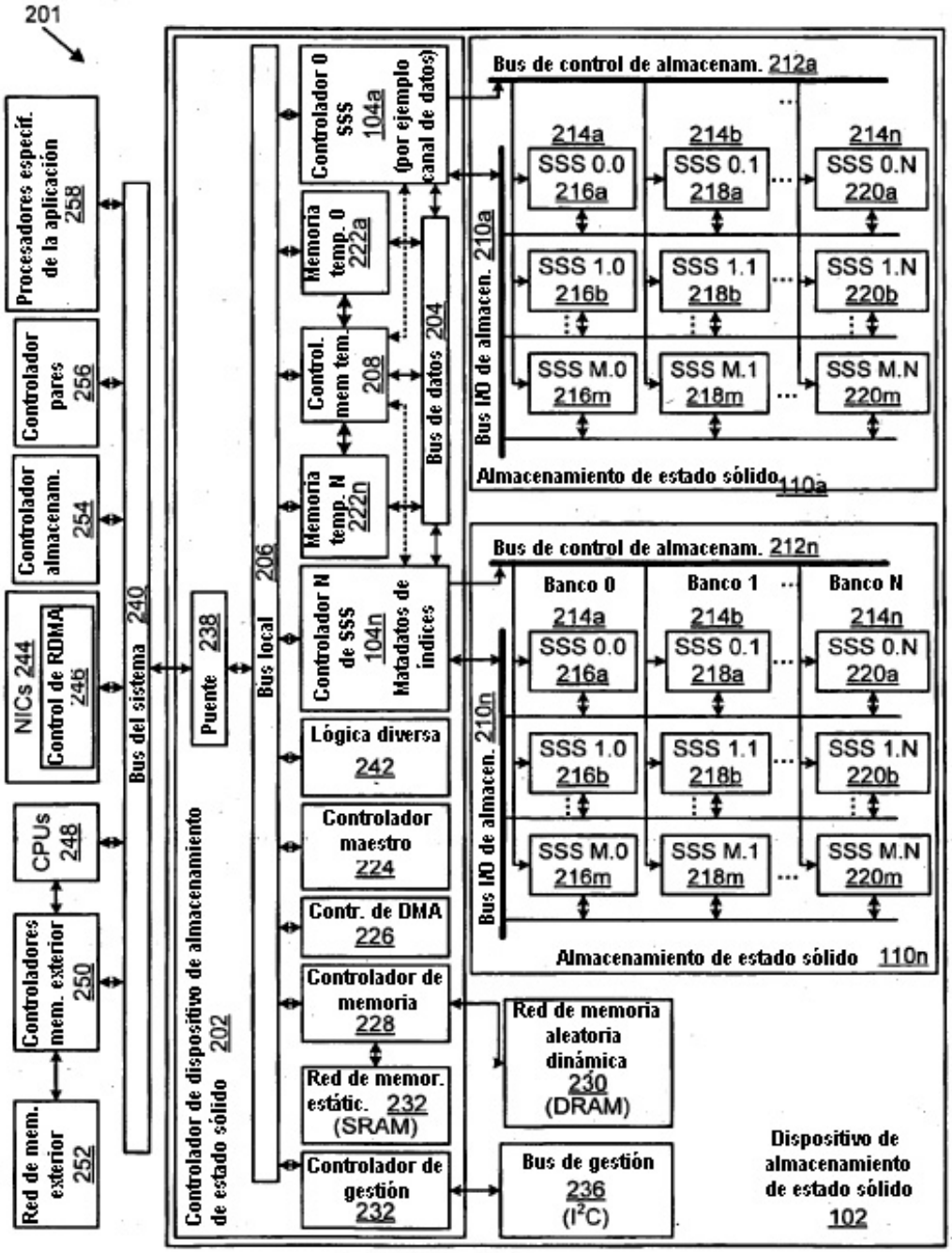
201
Procesadores específ. de la aplicación 258
Controlador pares 256
Controlador almacenam. 254
NICs 244
Control de RDMA 246
CPUs 248
Controladores mem. exterior 250
Red de mem. exterior 252
Bus del sistema 240
Puente 238
Bus local 206
Controlador de dispositivo de almacenamiento de estado sólido 202
Controlador 0 SSS 104a (por ejemplo canal de datos)
Memoria temp. 0 222a
Control. mem tem. 208
Memoria temp. N 222n
Bus de datos 204
Controlador N de SSS 104n Matadatos de índices
Lógica diversa 242
Controlador maestro 224
Contr. de DMA 226
Controlador de memoria 228
Red de memor. estátic. 232 (SRAM)
Controlador de gestión 232
Bus de control de almacenam. 212a
214a
214b
214n
SSS 0.0 216a
SSS 0.1 218a
SSS 0.N 220a
SSS 1.0 216b
SSS 1.1 218b
SSS 1.N 220b
SSS M.0 216m
SSS M.1 218m
SSS M.N 220m
Bus I/O de almacen. 210a
Almacenamiento de estado sólido 110a
Bus de control de almacenam. 212n
Banco 0
Banco 1
Banco N
Bus I/O de almacen. 210n
Almacenamiento de estado sólido 110n
Red de memoria aleatoria dinámica 230 (DRAM)
Bus de gestión 236 (I2C)
Dispositivo de almacenamiento de estado sólido 102

FIG. 2B

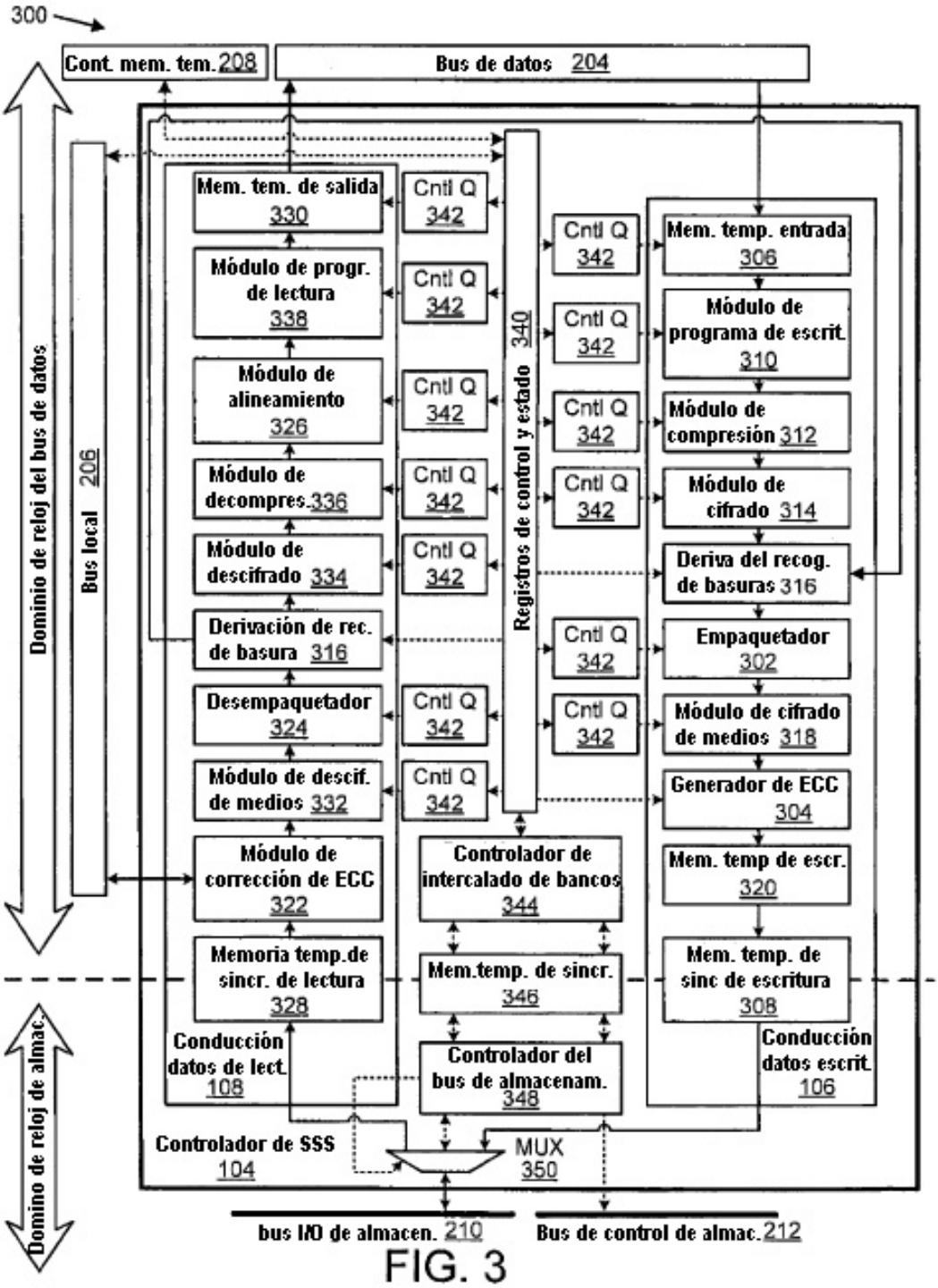
300
Cont. mem. tem. 208
Bus de datos 204
Dominio de reloj del bus de datos
Bus local 206
Mem. tem. de salida 330
Módulo de progr. de lectura 338
Módulo de alineamiento 326
Módulo de decompres. 336
Módulo de descifrado 334
Derivación de rec. de basura 316
Desempaquetador 324
Módulo de descif. de medios 332
Módulo de corrección de ECC 322
Memoria temp.de sincr. de lectura 328
Conducción datos de lect. 108
Cntl Q 342
Registros de control y estado 340
Controlador de intercalado de bancos 344
Mem.temp. de sincr. 346
Controlador del bus de almacenam. 348
Mem. temp. entrada 306
Módulo de programa de escrit. 310
Módulo de compresión 312
Módulo de cifrado 314
Deriva del recog. de basuras 316
Empaquetador 302
Módulo de cifrado de medios 318
Generador de ECC 304
Mem. temp de escr. 320
Mem. temp. de sinc de escritura 308
Conducción datos escrit. 106
Controlador de SSS 104
MUX 350
Domino de reloj de almac.
bus I/O de almacen. 210
Bus de control de almac. 212

FIG. 3

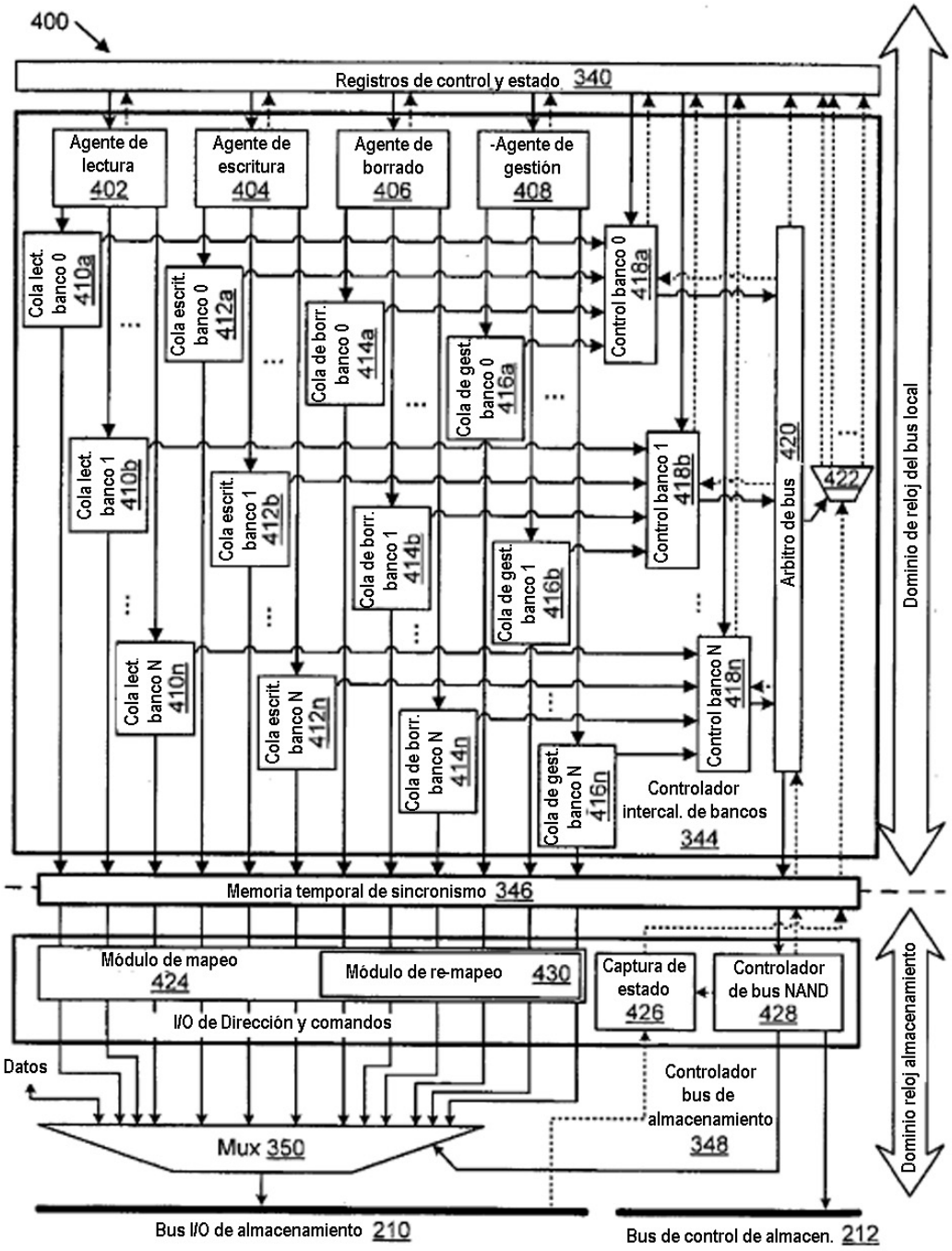
400
Registros de control y estado 340
Agente de lectura 402
Agente de escritura 404
Agente de borrado 406
-Agente de gestión 408
Cola lect. banco 0 410a
Cola escrit. banco 0 412a
Cola de borr. banco 0 414a
Cola de gest. banco 0 416a
Control banco 0 418a
Cola lect. banco 1 410b
Cola escrit. banco 1 412b
Cola de borr. banco 1 414b
Cola de gest. banco 1 416b
Control banco 1 418b
Cola lect. banco N 410n
Cola escrit. banco N 412n
Cola de borr. banco N 414n
Cola de gest. banco N 416n
Control banco N 418n
Arbitro de bus 420
422
Dominio de reloj del bus local
Controlador intercal. de bancos 344
Memoria temporal de sincronismo 346
Módulo de mapeo 424
Módulo de re-mapeo 430
Captura de estado 426
Controlador de bus NAND 428
I/O de Dirección y comandos
Datos
Controlador bus de almacenamiento 348
Mux 350
Dominio reloj almacenamiento
Bus I/O de almacenamiento 210
Bus de control de almacen. 212

FIG. 4A

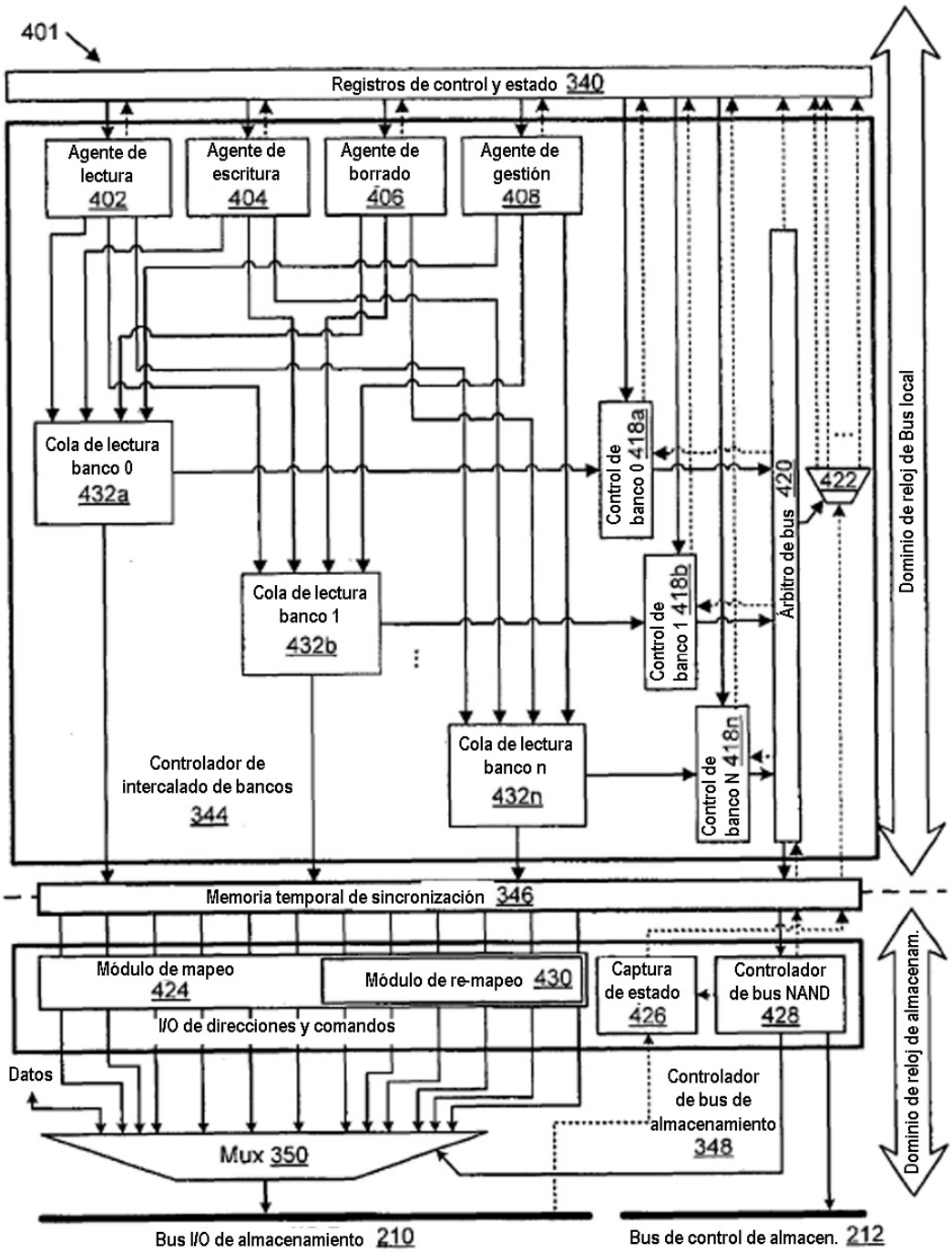
401
Registros de control y estado 340
Agente de lectura 402
Agente de escritura 404
Agente de borrado 406
Agente de gestión 408
Cola de lectura banco 0 432a
Cola de lectura banco 1 432b
Cola de lectura banco n 432n
Controlador de intercalado de bancos 344
Control de banco 0 418a
Control de banco 1 418b
Control de banco N 418n
Arbitro de bus 420
422
Dominio de reloj de Bus local
Memoria temporal de sincronización 346
Módulo de mapeo 424
Módulo de re-mapeo 430
I/O de direcciones y comandos
Captura de estado 426
Controlador de bus NAND 428
Datos
Controlador de bus de almacenamiento 348
Mux 350
Dominio de reloj de almacenam.
Bus I/O de almacenamiento 210
Bus de control de almacen. 212

FIG. 4B

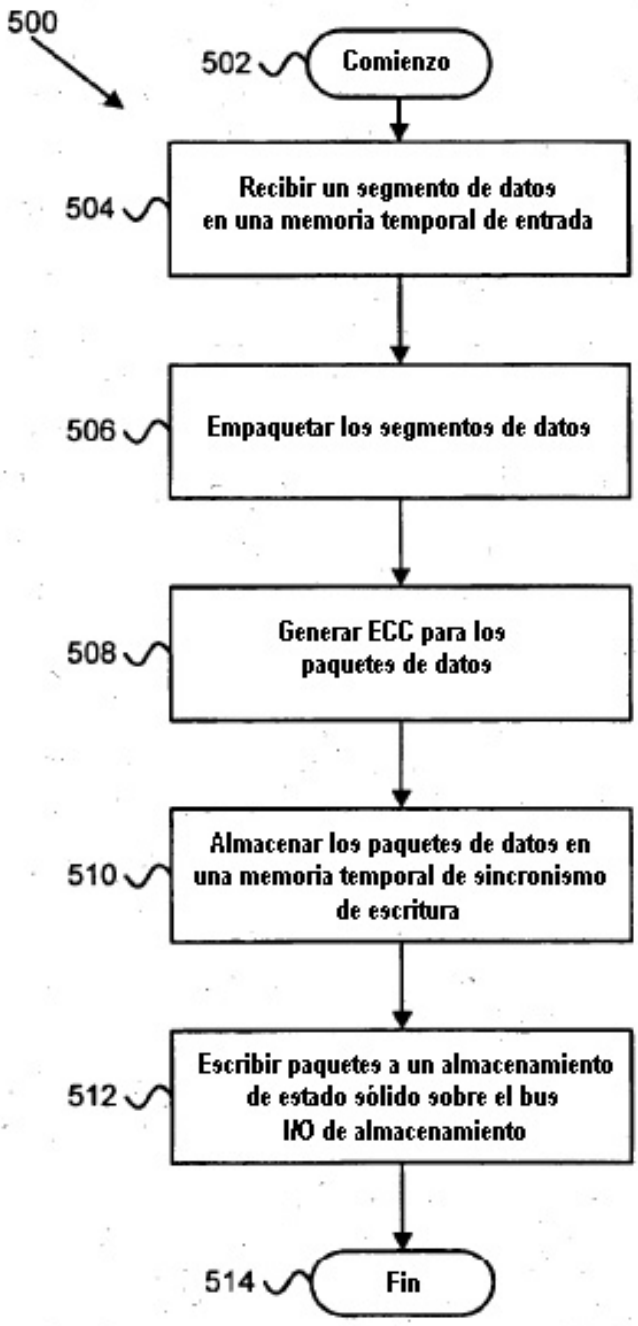
500
502
Comienzo
504
Recibir un segmento de datos en una memoria temporal de entrada
506
Empaquetar los segmentos de datos
508
Generar ECC para los paquetes de datos
510
Almacenar los paquetes de datos en una memoria temporal de sincronismo de escritura
512
Escribir paquetes a un almacenamiento de estado sólido sobre el bus I/O de almacenamiento
514
Fin

FIG. 5A

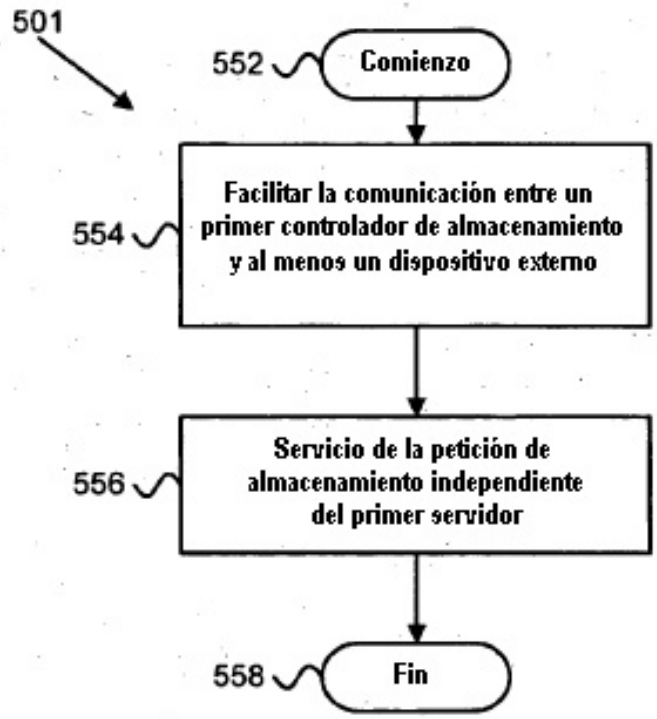
501
552
Comienzo
554
Facilitar la comunicación entre un primer controlador de almacenamiento y al menos un dispositivo externo
556
Servicio de la petición de almacenamiento independiente del primer servidor
558
Fin

FIG. 5B

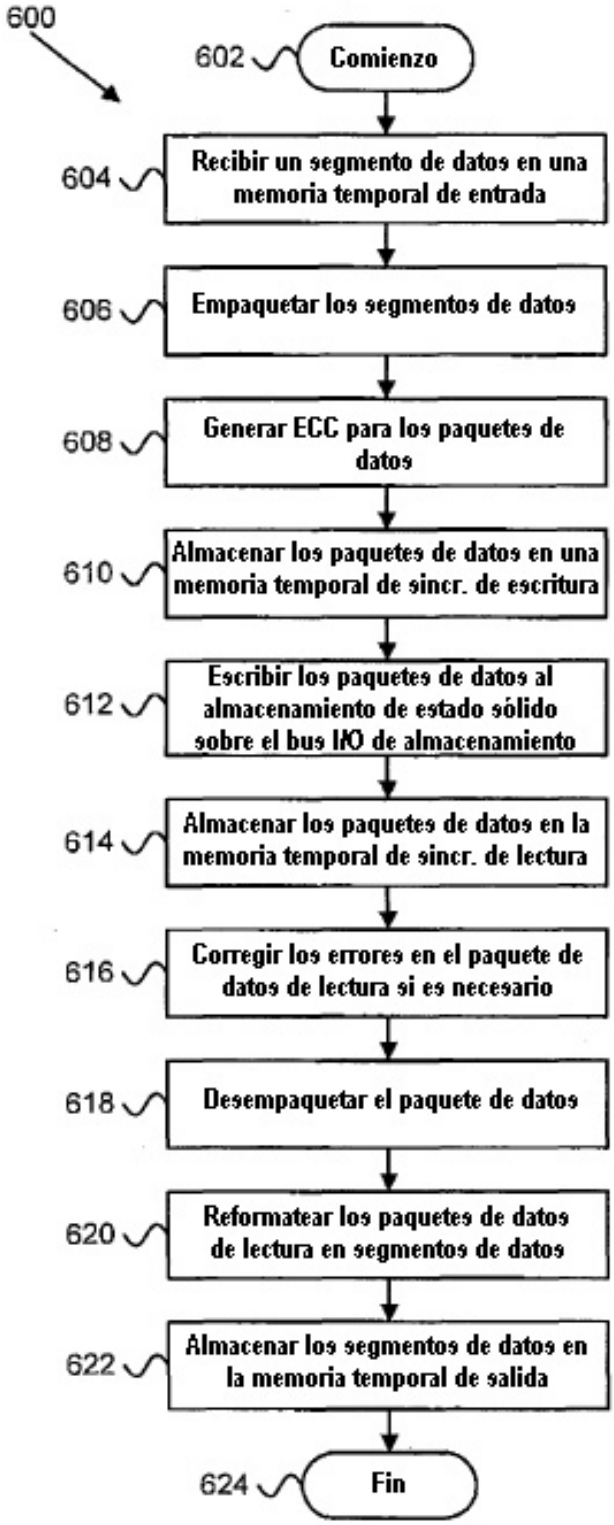
600
602
Comienzo
604
Recibir un segmento de datos en una memoria temporal de entrada
606
Empaquetar los segmentos de datos
608
Generar ECC para los paquetes de datos
610
Almacenar los paquetes de datos en una memoria temporal de sincr. de escritura
612
Escribir los paquetes de datos al almacenamiento de estado sólido sobre el bus I/O de almacenamiento
614
Almacenar los paquetes de datos en la memoria temporal de sincr. de lectura
616
Corregir los errores en el paquete de datos de lectura si es necesario
618
Desempaquetar el paquete de datos
620
Reformatear los paquetes de datos de lectura en segmentos de datos
622
Almacenar los segmentos de datos en la memoria temporal de salida
624
Fin

FIG. 6

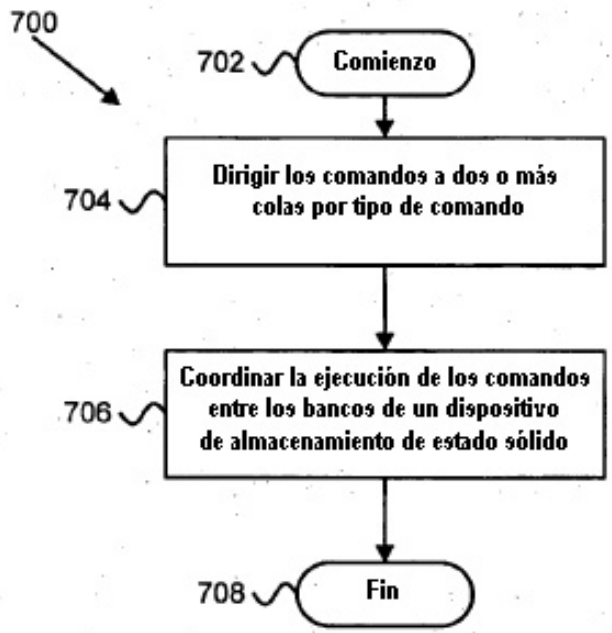
700
702
Comienzo
704
Dirigir los comandos a dos o más colas por tipo de comando
706
Coordinar la ejecución de los comandos entre los bancos de un dispositivo de almacenamiento de estado sólido
708
Fin

FIG. 7

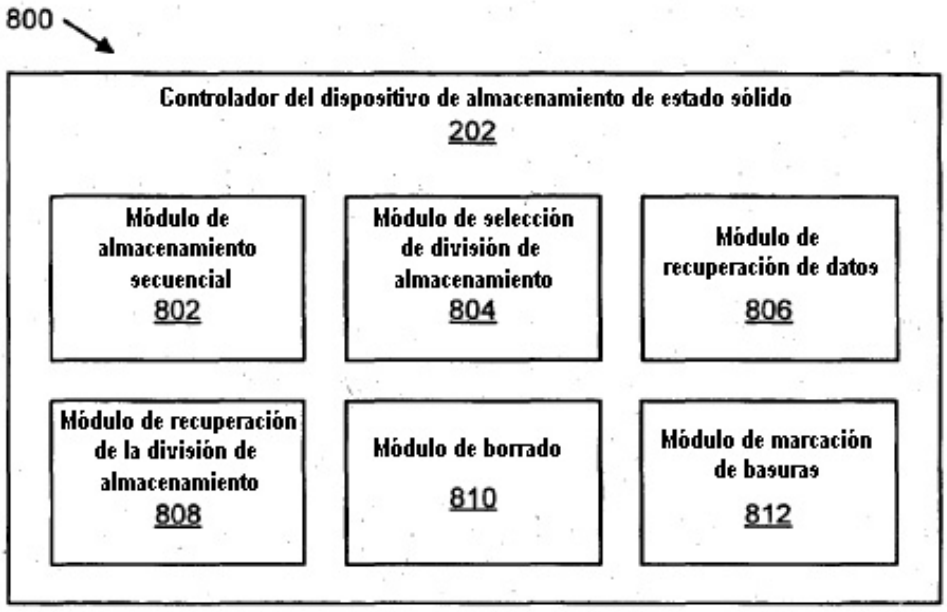
800
Controlador del dispositivo de almacenamiento de estado sólido
202
Módulo de almacenamiento secuencial
802
Módulo de selección de división de almacenamiento
804
Módulo de recuperación de datos
806
Módulo de recuperación de la división de almacenamiento
808
Módulo de borrado
810
Módulo de marcación de basuras
812

FIG. 8

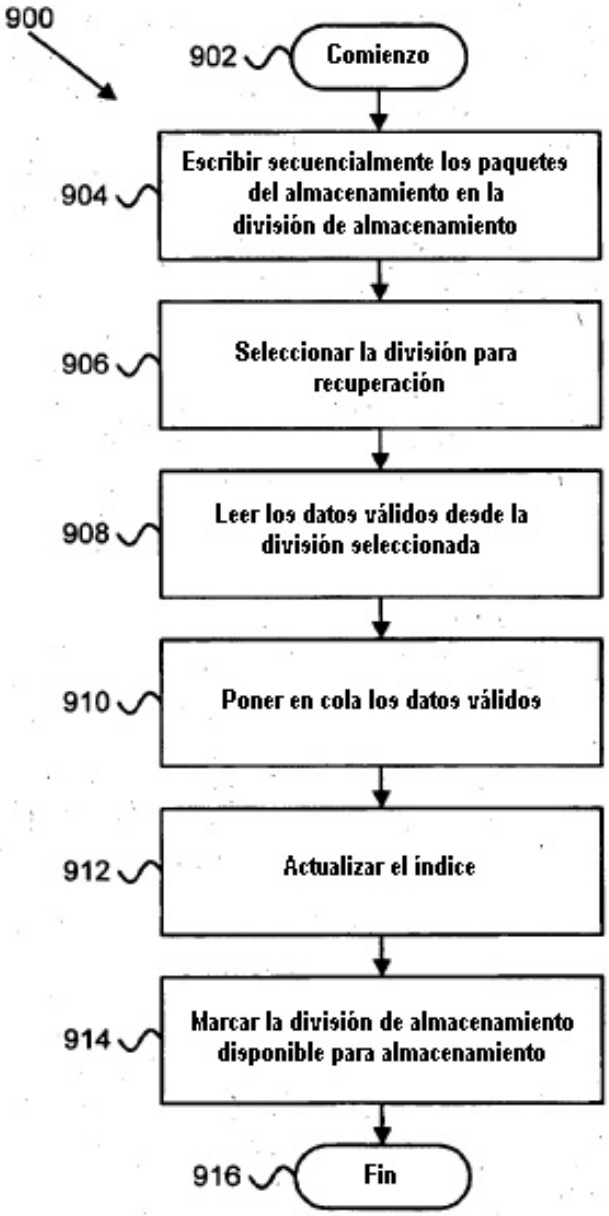
900
902
Comienzo
904
Escribir secuencialmente los paquetes del almacenamiento en la división de almacenamiento
906
Seleccionar la división para recuperación
908
Leer los datos válidos desde la división seleccionada
910
Poner en cola los datos válidos
912
Actualizar el índice
914
Marcar la división de almacenamiento disponible para almacenamiento
916
Fin

FIG. 9

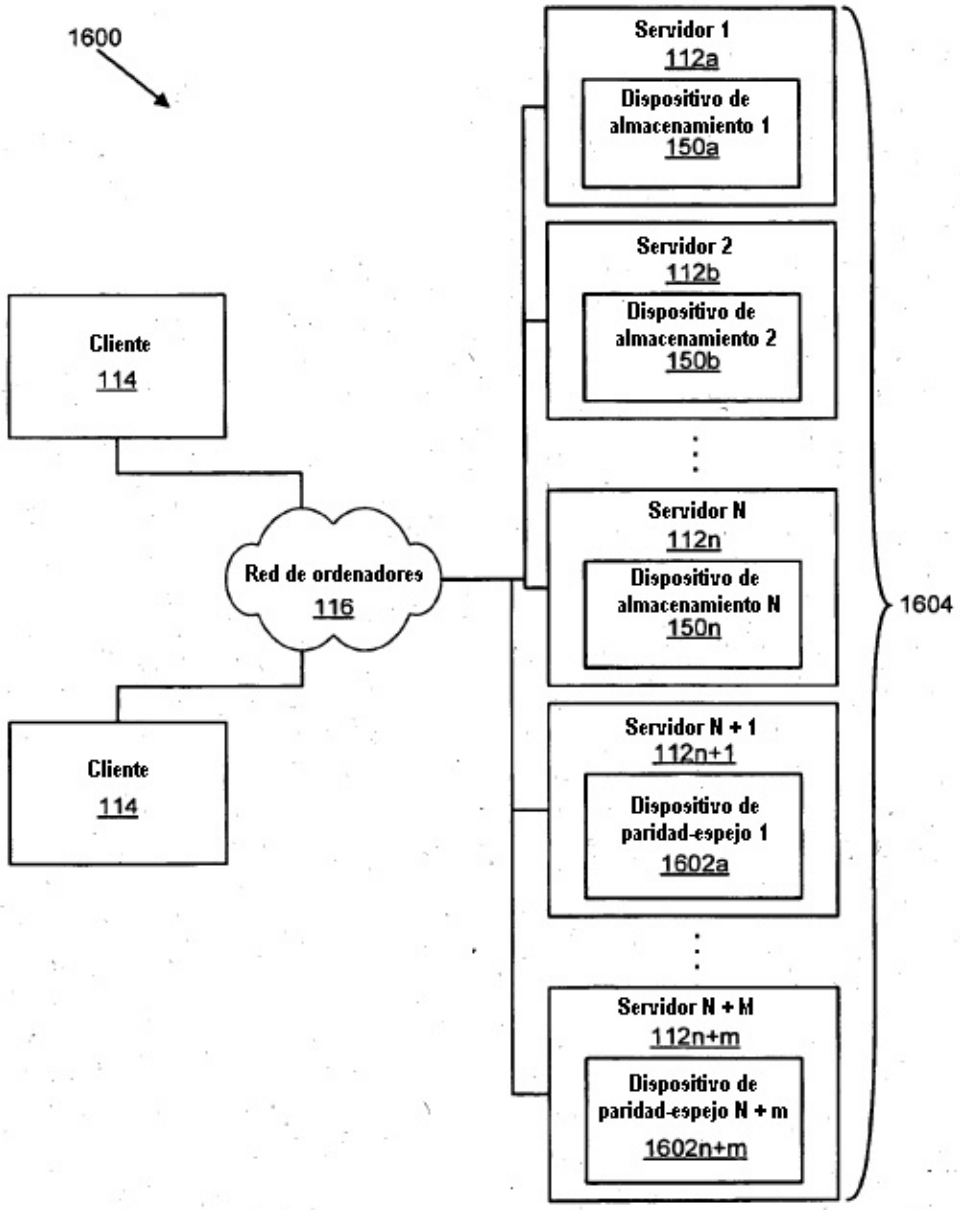
1600
Servidor 1
112a
Dispositivo de almacenamiento 1
150a
Servidor 2
112b
Dispositivo de almacenamiento 2
150b
Cliente
114
Servidor N
112n
Dispositivo de almacenamiento N
150n
Red de ordenadores
116
1604
Servidor N + 1
112n+1
Dispositivo de paridad-espejo 1
1602a
Cliente
114
Servidor N + M
112n+m
Dispositivo de paridad-espejo N + m
1602n+m

FIG. 10


Aparato de RAID distribuida del extremo frontal
2100
Módulo receptor de peticiones de almacenamiento
2102
Módulo de asociación de desmontaje
2104
Módulo de asociación de paridad-espejo
2106
Módulo transmisor de peticiones de almacenamiento
2108
Módulo de generación de paridad del extremo frontal
2110
Módulo de recuperación de segmentos de datos
2112
Módulo de reconstrucción de datos
2114
Módulo de reconstrucción de paridad
2116
Módulo de alternancia de paridad
2118
Módulo de comunicaciones entre pares
2120

FIG. 11

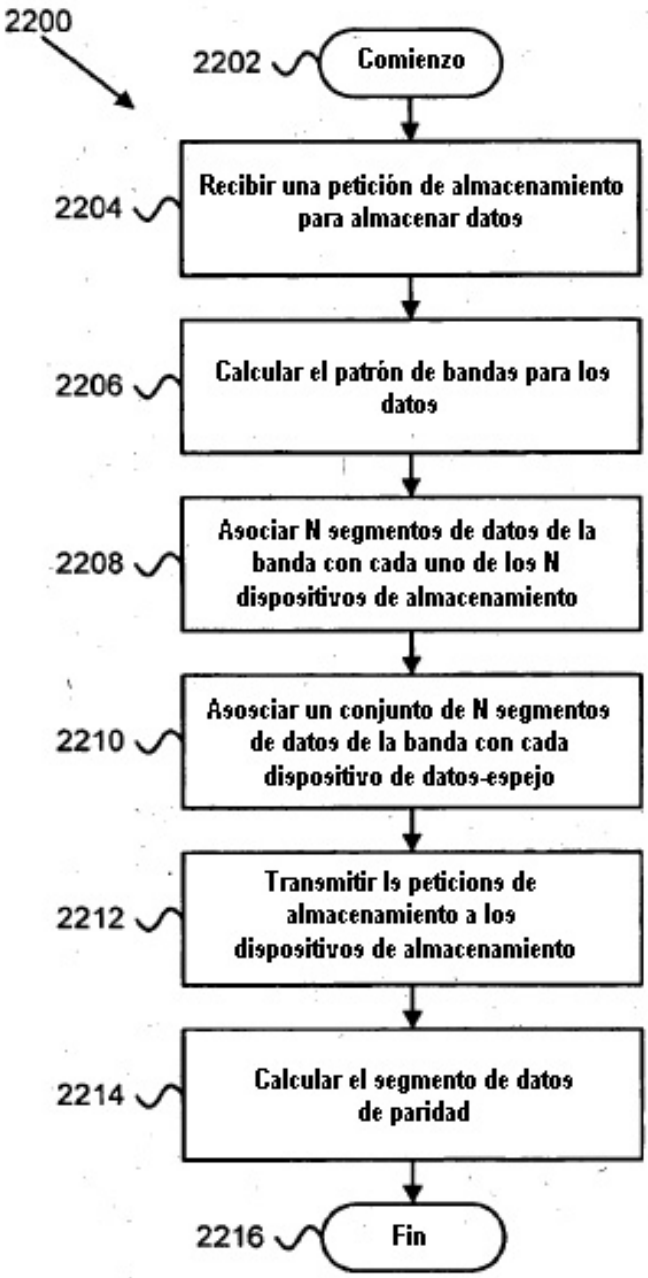
2200
2202
Comienzo
2204
Recibir una petición de almacenamiento para almacenar datos
2206
Calcular el patrón de bandas para los datos
2208
Asociar N segmentos de datos de la banda con cada uno de los N dispositivos de almacenamiento
2210
Asociar un conjunto de N segmentos de datos de la banda con cada dispositivo de datos-espejo
2212
Transmitir la peticions de almacenamiento a los dispositivos de almacenamiento
2214
Calcular el segmento de datos de paridad
2216
Fin

FIG. 12

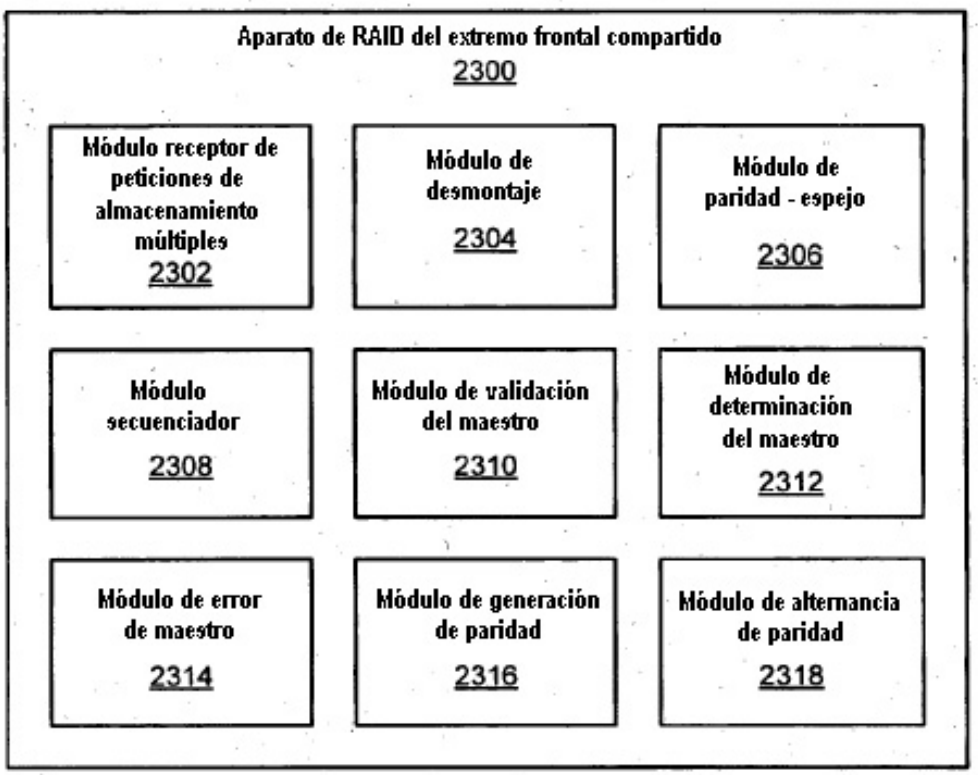
Aparato de RAID del extremo frontal compartido
2300
Módulo receptor de peticiones de almacenamiento múltiples
2302
Módulo de desmontaje
2304
Módulo de paridad - espejo
2306
Módulo secuenciador
2308
Módulo de validación del maestro
2310
Módulo de determinación del maestro
2312
Módulo de error de maestro
2314
Módulo de generación de paridad
2316
Módulo de alternancia de paridad
2318

FIG. 13

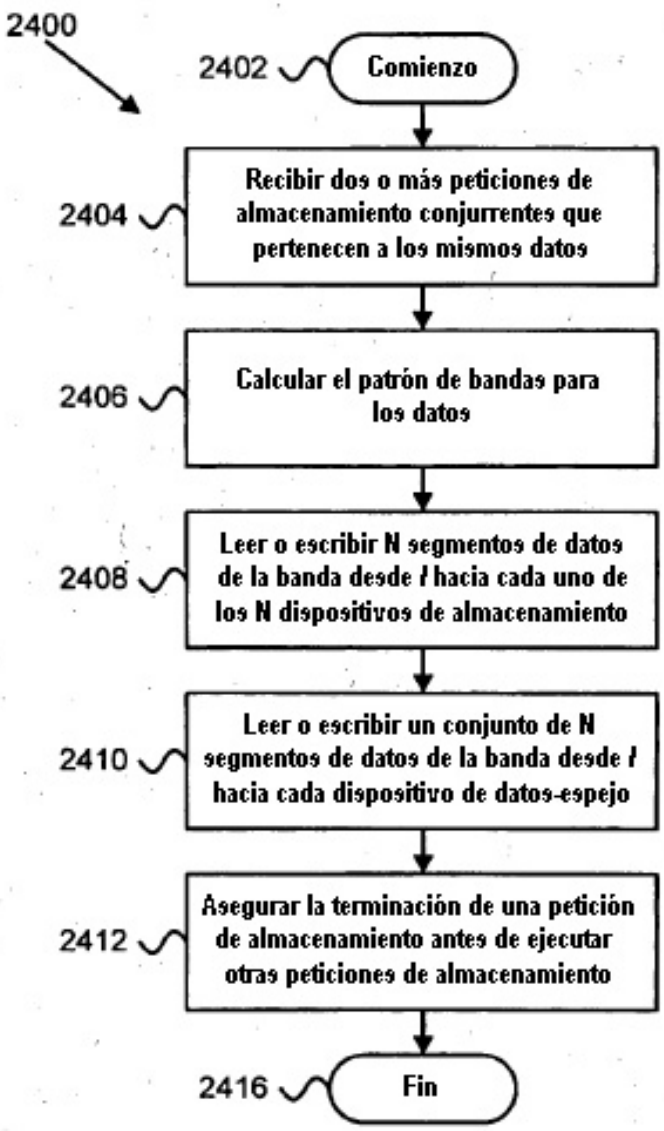
2400
2402
Comienzo
2404
Recibir dos o más peticiones de almacenamiento conjurrentes que pertenecen a los mismos datos
2406
Calcular el patrón de bandas para los datos
2408
Leer o escribir N segmentos de datos de la banda desde *I* hacia cada uno de los N dispositivos de almacenamiento
2410
Leer o escribir un conjunto de N segmentos de datos de la banda desde *I* hacia cada dispositivo de datos-espejo
2412
Asegurar la terminación de una petición de almacenamiento antes de ejecutar otras peticiones de almacenamiento
2416
Fin

FIG. 14

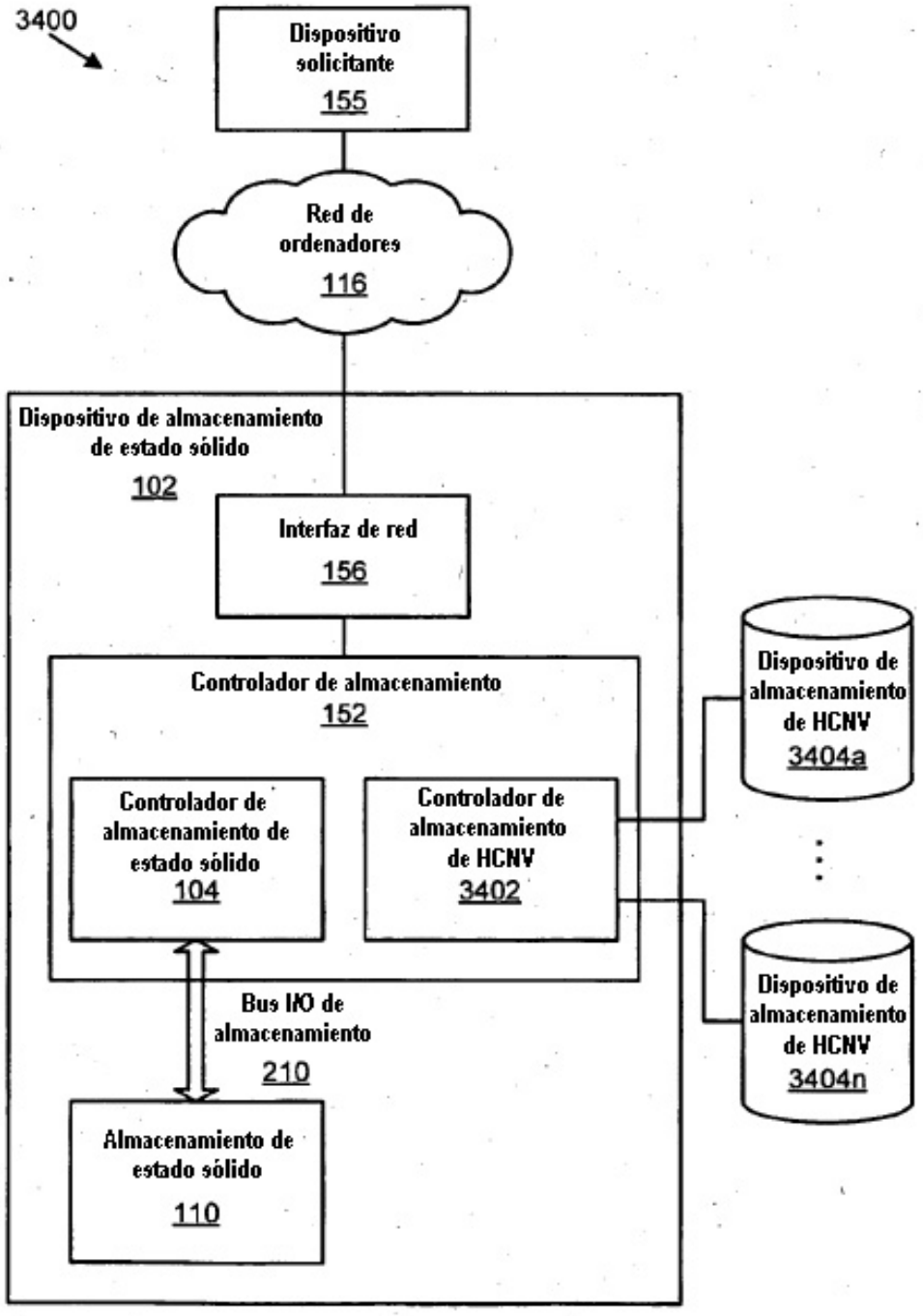
3400
Dispositivo solicitante
155
Red de ordenadores
116
Dispositivo de almacenamiento de estado sólido
102
Interfaz de red
156
Controlador de almacenamiento
152
Controlador de almacenamiento de estado sólido
104
Controlador de almacenamiento de HCNV
3402
Bus I/O de almacenamiento
210
Almacenamiento de estado sólido
110
Dispositivo de almacenamiento de HCNV
3404a
Dispositivo de almacenamiento de HCNV
3404n

FIG. 15

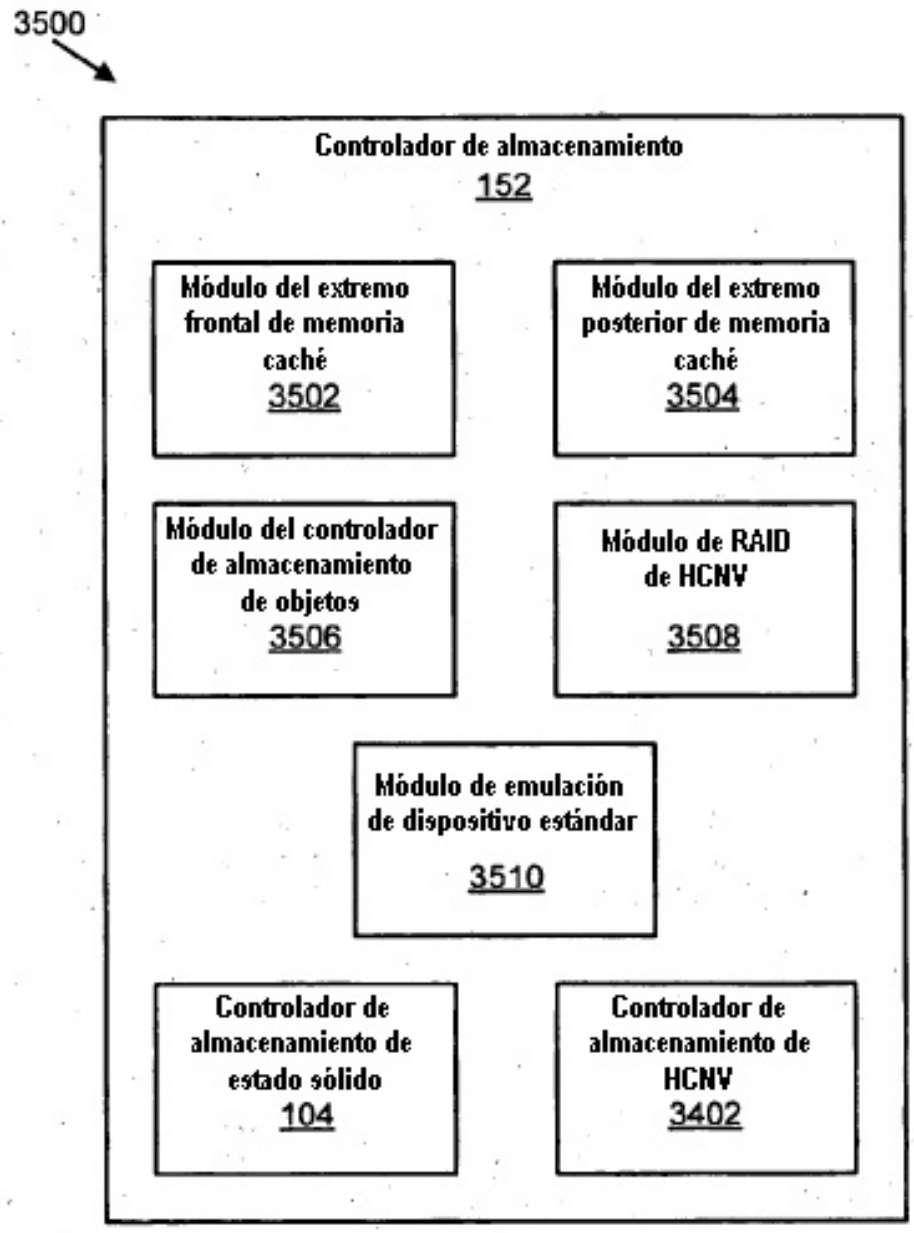
3500
Controlador de almacenamiento
152
Módulo del extremo frontal de memoria caché
3502
Módulo del extremo posterior de memoria caché
3504
Módulo del controlador de almacenamiento de objetos
3506
Módulo de RAID de HCNV
3508
Módulo de emulación de dispositivo estándar
3510
Controlador de almacenamiento de estado sólido
104
Controlador de almacenamiento de HCNV
3402

FIG. 16

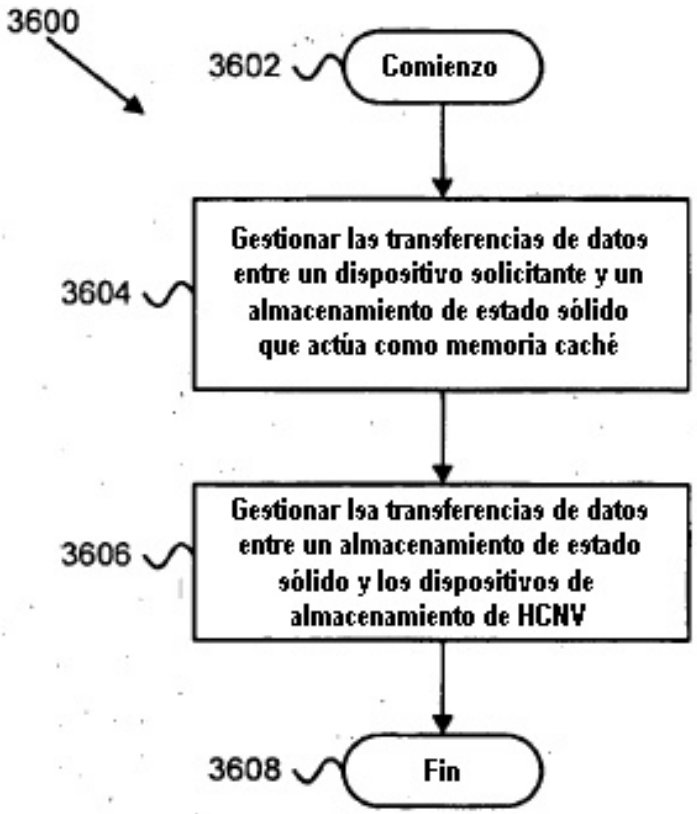
3600
3602
Comienzo
Gestionar las transferencias de datos entre un dispositivo solicitante y un almacenamiento de estado sólido que actúa como memoria caché
3604
Gestionar lsa transferencias de datos entre un almacenamiento de estado sólido y los dispositivos de almacenamiento de HCNV
3606
3608
Fin

FIG. 17