(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 632 558**

(51) Int. CI.:

G06F 12/00 (2006.01)
G06F 17/40 (2006.01)
G06F 9/50 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **13.02.2013 PCT/US2013/025798**

(87) Fecha y número de publicación internacional: **06.09.2013 WO13130262**

(96) Fecha de presentación y número de la solicitud europea: **13.02.2013 E 13755733 (6)**

(97) Fecha y número de publicación de la concesión europea: **19.04.2017 EP 2820549**

(54) Título: **Arbitración de propiedad de disco en un grupo de almacenamiento**

(30) Prioridad:

28.02.2012 US 201213407428

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
14.09.2017

(73) Titular/es:

MICROSOFT TECHNOLOGY LICENSING, LLC (100.0%)
One Microsoft Way
Redmond, Washington 98052-6399, US

(72) Inventor/es:

PADMANABAN, SAI SUDHIR ANANTHA;
KUZNETSOV, VYACHESLAV;
WARWICK, ALAN y
D'AMATO, ANDREA

(74) Agente/Representante:

CARPINTERO LÓPEZ, Mario

DESCRIPCIÓN

Arbitración de propiedad de disco en un grupo de almacenamiento

Antecedentes

1. Antecedentes y técnica relevante

Los sistemas informáticos y la tecnología relacionada afectan a muchos aspectos de la sociedad. De hecho, la capacidad de los sistemas informáticos para procesar información ha transformado la manera en la que vivimos y trabajamos. Los sistemas informáticos ahora realizan comúnmente una multitud de tareas (por ejemplo, procesamiento de textos, planificación, contabilidad, etc.) que antes de la llegada de los sistemas informáticos se realizaban manualmente. Más recientemente, los sistemas informáticos se han acoplado entre sí y a otros dispositivos electrónicos para formar tanto redes informáticas alámbricas como inalámbricas a través de las cuales los sistemas informáticos y otros dispositivos electrónicos pueden transferir datos electrónicos. Por consiguiente, el rendimiento de muchas tareas informáticas se distribuye a través de un número de diferentes sistemas informáticos y/o un número de diferentes entorno informáticos.

La agrupación se refiere al agrupamiento de múltiples sistemas informáticos, denominados en el presente documento como nodos. En ocasiones, un clúster emplea almacenamiento compartido para posibilitar que las aplicaciones que se ejecutan en cualquiera de los nodos accedan a los mismos datos. El almacenamiento compartido posibilita la migración tras error de aplicaciones de nodo a nodo en el clúster. Por ejemplo, si un nodo falla, las aplicaciones que se ejecutan en el nodo fallido pueden conmutarse a través de otro nodo donde continúan ejecutándose. Puesto que los datos usados por las aplicaciones migradas tras error se almacenan en un almacenamiento compartido accesible desde cualquier nodo, las aplicaciones pueden continuar ejecutándose (por ejemplo acceder a los mismos datos) después de que se conmutan a otro nodo. De esta manera, la migración tras error es esencialmente transparente desde una perspectiva del usuario.

Para implementar un clúster, es necesario mantener consistencia a través de los nodos. Por ejemplo, los datos de configuración del clúster deberían permanecer consistentes a través de los nodos incluso aunque cada nodo tenga acceso a los datos de configuración. Siempre que cada nodo pueda comunicar con los otros nodos en el clúster, puede mantenerse la consistencia. En algunos clústeres, cada nodo almacena una copia de los datos de configuración del clúster, y un servicio de clúster sincroniza los datos a través de los nodos.

Surge un problema cuando tiene lugar una partición de red que evita por lo tanto que los nodos comuniquen. La Figura 1 (Técnica anterior) ilustra un ejemplo donde ha tenido lugar una partición 107 de red en un clúster 100 que evita que una primera partición de nodos (nodos 101-102) comuniquen con una segunda partición de nodos (nodos 103-104). El almacenamiento 106 compartido permanece accesible a cada nodo del clúster; sin embargo, los nodos en la primera partición no pueden comunicar con los nodos en la segunda partición. Cuando tiene lugar una división de este tipo en el clúster, es necesario que únicamente una partición continúe funcionado como un clúster para asegurar que se mantiene la consistencia.

Para asegurar que únicamente una partición en un clúster se continúe ejecutando (es decir, mantenga el acceso al almacenamiento 106 compartido), el servicio de clúster requiere que una partición tenga quórum. En general, tener quórum significa que la partición comprende una mayoría de los elementos en el clúster. Dependiendo de la configuración del clúster, los elementos pueden incluir los nodos del clúster y posiblemente un disco (en ocasiones denominado como un testigo de disco) o una compartición de fichero. Los testigos de disco se usan cuando existe un número par de nodos para evitar la situación donde pudiera ocurrir un empate (por ejemplo si únicamente se consideraran los nodos cuando se determina la mayoría).

En la Figura 1, tanto la primera como la segunda partición tienen el mismo número de nodos. Por consiguiente, cualesquiera partición que tenga la propiedad de un almacenamiento 106 compartido tendrá quórum (es decir 2 nodos + 1 disco). Cuando un nodo tiene propiedad de un disco, tiene acceso de escritura al disco. Otros nodos pueden escribir también en el disco, pero el propietario puede controlar qué nodos pueden tener acceso. De esta manera, se le proporciona control a un único nodo sobre quién puede acceder al disco. En la Figura 1, antes de la aparición de la partición 107 de red, el nodo 101 tiene propiedad del almacenamiento 106 compartido.

En general, un nodo propietario permite el acceso a nodos que tiene conocimiento que son miembros del clúster y están dentro de la partición del nodo propietario (por ejemplo cuando tiene lugar una partición). Por lo tanto, cuando tiene lugar una partición, el nodo que tiene propiedad del disco es responsable de evitar que nodos fuera de la mayoría accedan al almacenamiento compartido. Por ejemplo, en la Figura 1, el nodo 101, como propietario, evitaría que los nodos 103-104 accedan al almacenamiento 106 compartido después de que tiene lugar la partición 107 de red.

La propiedad de disco se determina comúnmente y se gestiona usando un protocolo SCSI conocido como Reserva Persistente (PR). PR es un mecanismo de defensa/desafío. En PR, cada nodo tiene una clave única conocida como una clave de registro. Para obtener la propiedad de un disco, un nodo debe tener su clave registrada, y a continuación debe **obtener** una reserva con la clave.

La Figura 2A (Técnica anterior) ilustra dos tablas que se usan en PR. La Figura 2A representa el estado de las dos tablas cuando el nodo 101 es el propietario actual del disco (almacenamiento 106 compartido). Aunque se muestran dos tablas en este ejemplo, PR puede implementarse usando una única tabla u otra estructura de datos para almacenar información similar.

Para tener acceso de escritura en un disco, un nodo debe registrar su clave de registro. Un nodo registra su clave añadiéndola a la tabla 201 de registro. Por ejemplo, la tabla 201 de registro muestra que los nodos 101-104 tienen cada uno su clave registrada. Por consiguiente, los nodos 101-104 tienen acceso de escritura al almacenamiento 106 compartido.

Una vez registrado, un nodo puede a continuación intentar reservar su clave añadiéndola a la tabla 202 de reserva. Por ejemplo, la tabla 202 de reserva muestra que el nodo 101 ha reservado satisfactoriamente su clave haciendo de esta manera al nodo 101 el propietario del disco. Una vez que un nodo reserva su clave, puede usar su clave para controlar el disco.

Sin embargo, si ya existe una reserva cuando un nodo (denominado como un desafiador) intenta reservar su clave, la reserva fallará. Por ejemplo, debido a que la clave del nodo 101 ya está reservada, cualquier otro intento del nodo para reservar su clave fallará. Como parte de esta reserva fallida, un desafiador recibe la clave de reserva actual (por ejemplo la clave del nodo 101).

Cuando ya existe una reserva, un nodo debe adelantarse en primer lugar a la reserva antes de que pueda reservar su propia clave. Para posibilitar que un nodo propietario (denominado como el defensor) defienda su propiedad de un disco, las reglas PR requieren que un desafiador espere un periodo de tiempo especificado (generalmente 6 segundos) después de que una reserva falle antes de emitir un comando previo. Después de este periodo de tiempo, el desafiador se reactiva y emite un comando previo para intentar eliminar la clave del defensor de la tabla de reserva. El comando previo especifica la clave del desafiador así como la clave de reserva actual (la clave del defensor reservada en la tabla de reserva que el desafiador recibió cuando falló la reserva). Por ejemplo, el comando previo especifica la clave para eliminar de la tabla 202 de reserva (la clave del defensor), así como la clave para reservar en la tabla 202 de reserva (la clave del desafiador).

Para que tenga éxito el comando previo, la clave del desafiador debe aún estar en la tabla 201 de registro. Para defender su propiedad (es decir para provocar que el comando previo del desafiador falle), el defensor debe eliminar la clave de registro del desafiador de la tabla 201 de registro antes de que el desafiador emita un comando previo. En general, cada 3 segundos, el defensor se reactivará y comprobará la tabla 202 de reserva para determinar si se ha realizado algún desafío. Puesto que el defensor tiene propiedad del disco, puede eliminar cualquier otra clave de registro del nodo de la tabla 201 de registro.

Cuando el nodo 101 ve la clave del nodo 104 en la tabla 201 de registro, el nodo 101 eliminará la clave de registro del nodo 104 de la tabla 201 de registro. La Figura 2B (Técnica anterior) representa el estado de las dos tablas después de que el nodo 101 se ha defendido contra el desafío del nodo 104. Como se muestra, la clave del nodo 104 ya no se enumera en la tabla 201 de registro.

Por consiguiente, cuando el nodo 104 se reactiva y emite un comando previo, el comando previo fallará puesto que el defensor ya ha eliminado la clave de registro del nodo 104. Puesto que la clave de registro del nodo 104 se ha eliminado de la tabla 201 de registro, el nodo 104 ya no tiene acceso al almacenamiento 106 compartido.

Puesto que el defensor entra en inactividad durante una duración más corta que el desafiador, siempre que el defensor permanezca funcional y no tenga razón para defender su propiedad, mantendrá la propiedad eliminado periódicamente las claves de registro de cualquier nodo que desafíe propiedad. Sin embargo, si el defensor falla o pierde acceso de otra manera al disco, fallará al eliminar la clave de registro de un desafiador permitiendo de esta manera que el desafiador se adelante satisfactoriamente a la propiedad (es decir eliminar las claves del defensor de la tabla de registro y de reserva).

Además de eliminar una clave de registro del desafiador en una defensa, un defensor eliminará también periódicamente las claves de registro de cualquier nodo que no reconozca como que es parte del clúster activo. En otras palabras, un nodo en un clúster se actualiza periódicamente con respecto a los nodos con los que puede comunicar. Si un nodo propietario recibe una notificación de que no puede comunicar con otro nodo en el clúster (por ejemplo debido a una partición de red), el nodo propietario eliminará la clave de registro del otro nodo para evitar que el otro nodo acceda al clúster de almacenamiento.

Por consiguiente, en PR convencional, hay dos maneras generales en las que se eliminará una clave de registro de un nodo: (1) cuando el nodo propietario recibe notificación de que no puede comunicar con el nodo; y (2) cuando el nodo desafía insatisfactoriamente la propiedad del almacenamiento compartido.

Cuando tiene lugar una partición de red (es decir cuando un nodo tiene conocimiento de que no puede comunicar con cada nodo en el clúster), un nodo en una partición que no tiene la propiedad del disco en general empezará a desafiar para la propiedad del disco. Por ejemplo, en la Figura 1, después de que tiene lugar la partición 107 de red, uno o más de los nodos 103-104 pueden comenzar un desafío para adelantarse a la propiedad del nodo 101 del

almacenamiento 106 compartido (puesto que los nodos 103 y 104 necesitan la propiedad del almacenamiento 106 compartido para obtener quórum (para de esta manera poder continuar funcionando dentro del clúster).

En la Figura 1, la partición 107 de red no evita que ninguno de los nodos acceda al almacenamiento 106 compartido (es decir únicamente evita que los nodos 101-102 comuniquen con los nodos 103-104). Por consiguiente, el nodo 101 continuará defendiendo su propiedad del almacenamiento 106 compartido de los desafíos del nodo 103 o del nodo 104.

Sin embargo, si el nodo 101 fallara o perdiera de otra manera la conexión al almacenamiento 106 compartido (o determinara que no debería defender puesto que no está en una partición que pudiera tener quórum), los desafíos del nodo 103 o 104 tendrían éxito (puesto que el nodo 101 no podría eliminar las claves de registro del nodo 103 o del nodo 104). Como resultado, cuando el desafiador (el nodo 103 o el nodo 104) se reactiva, su clave de registro se enumerará aún en la tabla 201 de registro permitiéndole de esta manera adelantarse a la propiedad del nodo 101. El desafiador tomará a continuación la propiedad del almacenamiento 106 compartido. Como propietario, el nodo comenzará defendiendo su propiedad como se ha descrito anteriormente (por ejemplo si el nodo 101 o 102 comenzó desafiando la propiedad).

El ejemplo anterior describe un clúster que proporciona a los nodos y a al almacenamiento con votos para determinar la propiedad. Pueden existir también otros esquemas de votación que usan las técnicas PR anteriormente descritas. Estos esquemas incluyen que únicamente vote el nodo (donde únicamente los nodos tienen un voto), y voto de nodo + compartición de fichero (donde los nodos y una compartición de fichero votan). Votar únicamente el nodo se usa comúnmente cuando el clúster incluye un número impar de nodos. Votar el nodo + compartición de fichero es similar al voto de nodo + almacenamiento anteriormente descrito, pero se usa cuando se usa una compartición de fichero para almacenamiento compartido.

PR como se ha descrito anteriormente funciona correctamente dentro de muchas configuraciones de almacenamiento típicas. Sin embargo, PR, como se ha descrito anteriormente, no es siempre satisfactorio cuando se usa en otros tipos de configuraciones de almacenamiento (por ejemplo cuando un clúster emplea discos virtuales como almacenamiento compartido). En el sistema operativo Windows, los discos virtuales se denominan como “Espacios de almacenamiento”. En un espacio de almacenamiento, se agregan múltiples discos físicos en un grupo de almacenamiento. El grupo de almacenamiento puede a continuación dividirse en uno o más “espacios” lógicos (o discos virtuales). Cada espacio aparece en las aplicaciones como un dispositivo de almacenamiento físico incluso aunque el espacio esté virtualizado y pueda realmente abarcar muchos dispositivos de almacenamiento físico diferentes.

Por ejemplo, la Figura 3 (Técnica anterior) ilustra un clúster 300 que es similar al clúster 100 de la Figura 1 excepto que el almacenamiento 106 compartido se ha sustituido por el grupo 306 de almacenamiento. El grupo 306 de almacenamiento comprende tres dispositivos 310-312 de almacenamiento físicos. A partir de este grupo, el usuario puede crear uno o más espacios. La Figura 3 muestra que el usuario ha creado un único espacio 307. El espacio 307 puede tratarse, desde la perspectiva de las aplicaciones en cada nodo en el clúster, como un disco físico normal.

Cuando se usan espacios, los nodos del clúster necesitan tener acceso a cada disco físico en el grupo subyacente puesto que los espacios pueden extenderse entre los discos físicos. Por ejemplo, en el clúster 300, los datos escritos en el espacio 307 podrían almacenarse físicamente en cualquiera de tres dispositivos de almacenamiento físico en el grupo 306 de almacenamiento. Cuando se usan espacios, un único nodo tiene propiedad del grupo de almacenamiento (lo que significa que el nodo tiene propiedad de cada disco físico en el grupo).

Un problema particular provocado aplicando técnicas PR convencionales con espacios es que en cualquier momento que un desafiador intenta adelantarse a la propiedad del grupo a partir del cual se crean los espacios, el propietario del grupo eliminará la clave del desafiador provocando de esta manera que cualquier E/S a cualquier espacio desde las aplicaciones en el desafiador falle (puesto que la clave del desafiador debe registrarse con un disco físico para posibilitar que las aplicaciones en el desafiador escriban en el disco físico).

Provocar que la E/S de un desafiador falle puede ser un resultado incorrecto de un desafío. Por ejemplo, puede notificarse a un desafiador de una partición de red antes de al defensor. En respuesta a la partición de red, el desafiador comienza un desafío para el grupo. Si el desafiador está en una partición que tiene quórum, el resultado correcto sería que el desafiador ganara el desafío para tener la propiedad del grupo (para permitir de esta manera que la partición del desafiador continúe ejecutándose).

Sin embargo, si el defensor no ha sido notificado de la partición de red (y por consiguiente, no se ha notificado de que está en una partición que no tiene quórum), el defensor defenderá satisfactoriamente su propiedad del grupo. Usando técnicas PR convencionales como se ha descrito anteriormente, esta defensa incluye eliminar la clave de registro del desafiador de modo que cualquier escritura del desafiador fallará. El defensor continuará defendiendo satisfactoriamente su propiedad hasta que reciba la notificación de la partición de red. Por consiguiente, hasta que el defensor reciba la notificación de la partición de red, el clúster no comenzará operando apropiadamente en la partición con quórum (es decir la partición del desafiador). En otros escenarios, la aplicación de las técnicas PR

convencionales también conduce a resultados indeseables.

El documento US-2003/065782 describe un mecanismo de reserva persistente para recursos en una red de área de almacenamiento.

Breve resumen

La presente invención se extiende a procedimientos, sistemas y productos de programa informático para implementar técnicas de reserva persistente para establecer la propiedad de uno o más discos físicos. Estas técnicas de reserva persistente pueden emplearse para determinar la propiedad de discos físicos en un grupo de almacenamiento así como en cualquier otra configuración de almacenamiento. Usando las técnicas de reserva persistente de la presente invención, cuando tiene lugar una partición de red, un defensor de un disco físico no elimina una clave de registro del desafiador hasta que el defensor recibe notificación de que el desafiador ya no está en la partición del defensor. De esta manera, la E/S pendiente de las aplicaciones que se ejecutan en el desafiador no fallará debido a que se elimine la clave del desafiador hasta que pueda resolverse la propiedad correcta del disco físico.

En una realización, un primer nodo se defiende contra el intento del otro nodo de adelantarse a la reserva persistente del primer nodo en un dispositivo de almacenamiento. Después de que una partición de red que evita que el primer nodo comunique con otro nodo en el clúster, y antes de que se notifique al primer nodo de la partición de red, el primer nodo detecta que otro nodo en el clúster ha intentado reservar el dispositivo de almacenamiento compartido por los nodos del clúster. La detección comprende identificar que el otro nodo ha cambiado la clave de registro del otro nodo en una estructura de datos de registro.

El primer nodo cambia la clave de registro del primer nodo, registra la clave de registro cambiada en la estructura de datos de registro, y reserva la clave de registro cambiada en una estructura de datos de reserva.

En otra realización, un segundo nodo intenta eliminar una reserva persistente del primer nodo en un dispositivo de almacenamiento para obtener una reserva persistente para el segundo nodo. El segundo nodo recibe una notificación de que ha tenido lugar una partición de red que evita que el segundo nodo comunique con el primer nodo.

El segundo nodo intenta reservar la clave de registro del segundo nodo para obtener una reserva persistente en el dispositivo de almacenamiento. El intento para reservar incluye que el segundo nodo lea la clave de registro del primer nodo que se almacena en una estructura de datos de reserva y almacenar la clave del primer nodo. El intento para reservar también incluye cambiar el segundo nodo la clave de registro del segundo nodo y registrar la clave de registro cambiada. El intento para reservar también incluye entrar en inactividad el segundo nodo durante una duración de tiempo especificada antes de emitir un comando previo para eliminar la reserva persistente del primer nodo.

Este resumen se proporciona para introducir una selección de conceptos en una forma simplificada que se describe a continuación en la descripción detallada. Este resumen no se pretende para identificar características clave o características esenciales de la materia objeto reivindicada, ni se pretende para usarse como una ayuda al determinar el alcance de la materia objeto reivindicada.

Se expondrán características y ventajas adicionales de la invención en la descripción que sigue, y en parte serán evidentes a partir de la descripción, o pueden aprenderse mediante la puesta en práctica de la invención. Las características y ventajas de la invención pueden realizarse y obtenerse por medio de los instrumentos y combinaciones particularmente señaladas en las reivindicaciones adjuntas. Estas y otras características de la presente invención se harán más completamente evidentes a partir de la siguiente descripción y reivindicaciones adjuntas, o pueden aprenderse mediante la puesta en práctica de la invención como se expone en lo sucesivo.

Breve descripción de los dibujos

Para describir la manera en la que las ventajas y características anteriormente indicadas y otras de la invención pueden obtenerse, se describirá una descripción más particular de la invención brevemente por referencia a realizaciones específicas de la misma que se ilustran en los dibujos adjuntos. Entendiendo que estos dibujos representan únicamente realizaciones de la invención y no se han de considerar por lo tanto limitantes de su alcance, la invención se describirá y explicará con especificidad adicional y detalle a través del uso de los dibujos adjuntos en los que:

la Figura 1 ilustra un clúster de nodos típica en la que se usan técnicas de reserva persistente convencionales.
Las Figuras 2A-2B ilustran tablas ejemplares usadas al implementar reserva persistente convencional.
La Figura 3 ilustra un clúster de nodos que emplea un disco virtual para almacenamiento compartido.
La Figura 4 ilustra otro clúster de nodos que emplea una pluralidad de discos virtuales para almacenamiento compartido.
Las Figuras 5A-5F ilustran tablas ejemplares usadas al implementar las técnicas de reserva persistente de la presente invención;

la Figura 6 ilustra un formato ejemplar de una clave de registro que puede usarse para reserva persistente de acuerdo con la presente invención;
la Figura 7 ilustra un diagrama de flujo de un procedimiento ejemplar para que un primer nodo se defienda contra el intento de otro nodo de adelantarse a la reserva persistente del primer nodo en un dispositivo de almacenamiento; y
la Figura 8 ilustra un diagrama de flujo de un procedimiento ejemplar para que un segundo nodo intente eliminar una reserva persistente del primer nodo en un dispositivo de almacenamiento para obtener una reserva persistente para el segundo nodo.

Descripción detallada

La presente invención se extiende a procedimientos, sistemas y productos de programa informático para implementar técnicas de reserva persistente para establecer la propiedad de uno o más discos físicos. Estas técnicas de reserva persistente pueden emplearse para determinar la propiedad de discos físicos en un grupo de almacenamiento así como en cualquier otra configuración de almacenamiento. Usando las técnicas de reserva persistente de la presente invención, cuando tiene lugar una partición de red, un defensor de un disco físico no elimina una clave de registro del desafiador hasta que el defensor recibe notificación de que el desafiador ya no está en la partición del defensor. De esta manera, la E/S pendiente de las aplicaciones que se ejecutan en el desafiador no fallará debido a que se elimine la clave del desafiador hasta que pueda resolverse la propiedad correcta del disco físico.

En una realización, un primer nodo se defiende contra el intento del otro nodo de adelantarse a la reserva persistente del primer nodo en un dispositivo de almacenamiento. Después de que una partición de red que evita que el primer nodo comunique con otro nodo en el clúster, y antes de que se notifique al primer nodo de la partición de red, el primer nodo detecta que otro nodo en el clúster ha intentado reservar el dispositivo de almacenamiento compartido por los nodos del clúster. La detección comprende identificar que el otro nodo ha cambiado la clave de registro del otro nodo en una estructura de datos de registro.

El primer nodo cambia la clave de registro del primer nodo, registra la clave de registro cambiada en la estructura de datos de registro, y reserva la clave de registro cambiada en una estructura de datos de reserva.

En otra realización, un segundo nodo intenta eliminar una reserva persistente del primer nodo en un dispositivo de almacenamiento para obtener una reserva persistente para el segundo nodo. El segundo nodo recibe una notificación de que ha tenido lugar una partición de red que evita que el segundo nodo comunique con el primer nodo.

El segundo nodo intenta reservar la clave de registro del segundo nodo para obtener una reserva persistente en el dispositivo de almacenamiento. El intento para reservar incluye que el segundo nodo lea la clave de registro del primer nodo que se almacena en una estructura de datos de reserva y almacenar la clave del primer nodo. El intento para reservar también incluye cambiar el segundo nodo la clave de registro del segundo nodo y registrar la clave de registro cambiada. El intento para reservar también incluye entrar en inactividad el segundo nodo durante una duración de tiempo especificada antes de emitir un comando previo para eliminar la reserva persistente del primer nodo.

Las realizaciones de la presente invención pueden comprender o utilizar un ordenador de fin especial o de fin general que incluye hardware informático, tal como, por ejemplo, uno o más procesadores y memoria de sistema, como se analiza en mayor detalle a continuación. Las realizaciones dentro del alcance de la presente invención incluyen también medio legible por ordenador físico y otros para llevar o almacenar instrucciones y/o estructuras de datos ejecutables por ordenador. Tal medio legible por ordenador puede ser cualquier medio disponible que pueda accederse mediante un sistema informático de fin general o de fin especial. Medio legible por ordenador que almacena instrucciones ejecutables por ordenador son medios de almacenamiento informático (dispositivos). Medio legible por ordenador que lleva instrucciones ejecutables por ordenador son medios de transmisión. Por lo tanto, a modo de ejemplo, y no como limitación, las realizaciones de la invención pueden comprender al menos dos diferentes tipos de manera distinta de medio legible por ordenador: medio de almacenamiento informático (dispositivos) y medio de transmisión.

Medio de almacenamiento informático (dispositivos) incluyen RAM, ROM, EEPROM, CD-ROM, unidades de estado sólido (“SSD”) (por ejemplo, basadas en RAM), memoria Flash, memoria de cambio de fase (“PCM”), otros tipos de memoria, otro almacenamiento de disco óptico, almacenamiento de disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio que pueda usarse para almacenar medios de código de programa deseado en forma de instrucciones o estructuras de datos ejecutables por ordenador y que puede accederse por un ordenador de fin general o de fin especial.

Una “red” se define como uno o más enlaces de datos que posibilitan el transporte de datos electrónicos entre sistemas informáticos y/o módulos y/u otros dispositivos electrónicos. Cuando la información se transfiere o se proporciona a través de una red u otra conexión de comunicaciones (ya sea alámbrica, inalámbrica o una combinación de alámbrica o inalámbrica) a un ordenador, el ordenador observa de manera apropiada la conexión

como un medio de transmisión. Medio de transmisión puede incluir una red y/o enlaces de datos que pueden usarse para llevar o medios de código de programa deseados en forma de instrucciones o estructuras de datos ejecutables por ordenador y que pueden accederse por un ordenador de fin general o de fin especial. Deberían incluirse también combinaciones de lo anterior dentro del alcance de medio legible por ordenador.

Además, tras alcanzar diversos componentes de sistema informático, los medios de código de programa en forma de instrucciones o estructuras de datos ejecutables por ordenador pueden transferirse automáticamente de medio de transmisión a medio de almacenamiento informático (dispositivos) (o viceversa). Por ejemplo, instrucciones o estructuras de datos ejecutables por ordenador recibidas a través de una red o enlace de datos pueden almacenarse en memoria intermedia en RAM dentro de un módulo de interfaz de red (por ejemplo, un “NIC”), y a continuación transferirse eventualmente a RAM de sistema informático y/o a medio de almacenamiento informático menos volátil (dispositivos) en un sistema informático. Por lo tanto, debería entenderse que medio de almacenamiento informático (dispositivos) pueden incluirse en componentes de sistema informático que también (o incluso principalmente) utilizan medios de transmisión.

Instrucciones ejecutables por ordenador comprenden, por ejemplo, instrucciones y datos que, cuando se ejecutan en un procesador, provocan que un ordenador de fin general, ordenador de fin especial, o dispositivo de procesamiento de fin especial realice una cierta función o grupo de funciones. Las instrucciones ejecutables por ordenador pueden ser, por ejemplo, binarios, instrucciones de formato intermedio tal como lenguaje ensamblador o incluso código fuente. Aunque se ha descrito la materia objeto en lenguaje específico a características estructurales y/o actos metodológicos, se ha de entender que la materia objeto definida en las reivindicaciones adjuntas no está necesariamente limitada a las características descritas o actos descritos anteriormente. En su lugar, las características y actos descritos se desvelan como formas ejemplares para implementar las reivindicaciones.

Los expertos en la materia apreciarán que la invención puede ponerse en práctica en entornos informáticos de red con muchos tipos de configuraciones de sistemas informáticos, incluyendo, ordenadores personales, ordenadores de sobremesa, ordenadores portátiles, procesadores de mensajes, dispositivos portátiles, sistemas multiprocesador, electrónica de consumo basada en microprocesador o programable, PC de red, miniordenadores, ordenadores centrales, teléfonos móviles, PDA, tabletas, buscapersonas, encaminadores, conmutadores y similares. La invención puede ponerse también en práctica en entornos de sistemas distribuidos donde sistemas informáticos locales y remotos, que están enlazados (ya sean enlaces de datos alámbricos, enlaces de datos inalámbricos o mediante una combinación de enlaces de datos alámbricos e inalámbricos) a través de una red, ambos realizan tareas. En un entorno de sistema distribuido, los módulos de programa pueden localizarse tanto en dispositivos de almacenamiento de memoria locales como remotos.

Aunque se describe la presente invención como se implementa en el clúster 400 mostrado en la Figura 4, se ha de entender que las técnicas PR pueden implementarse también en otras arquitecturas informáticas que incluyen múltiples nodos. El clúster 400 incluye cuatro nodos interconectados (nodos 401-404). Cada nodo está conectado a cada uno de los dispositivos 410-412 de almacenamiento. Los dispositivos 410-412 de almacenamiento comprenden el grupo 420 de almacenamiento a partir del cual se han credo múltiples discos virtuales (o espacios) 430a-430n. Por consiguiente, las aplicaciones en cada nodo pueden acceder a cada uno de los discos 430a-430n virtuales. Se supone que se determina quórum en el clúster 400 usando un esquema de voto de nodos + almacenamiento. Por consiguiente, cada uno de los nodos 401-404 y el grupo 420 tiene un voto. Un único nodo tiene propiedad de cada dispositivo de almacenamiento en el grupo 420 en un tiempo dado.

Adicionalmente, las Figuras 5A-5F ilustran diversos estados de una tabla 501 de registro y una tabla 502 de reserva que pueden mantenerse para implementar las técnicas PR de la presente invención. Un conjunto separado de tabla de registro y de reserva se mantiene para cada uno de los dispositivos 410-412 de almacenamiento. En la siguiente descripción, se supondrá que la tabla 501 de registro y la tabla 502 de reserva pertenecen al dispositivo 410 de almacenamiento, pero debería entenderse que podría seguirse un procedimiento similar para determinar la propiedad de cada dispositivo de almacenamiento. También, incluso aunque se muestran dos tablas, la presente invención puede implementarse usando una única tabla o cualquier número de otras estructuras de datos. Por consiguiente, la estructura de datos usada para almacenar las claves no es esencial para la invención.

Las Figuras 5A-5F también incluyen los nodos 401 y 404 para ilustrar cómo los nodos almacenan claves actualmente reservadas durante el procedimiento PR de la presente invención. Por supuesto, aunque no se muestra, cada nodo también almacena su clave de registro.

La Figura 5A representa el estado de las dos tablas antes de la aparición de la partición 405 de red. En la Figura 5A, la tabla 501 de registro incluye cuatro claves de registro, una para cada uno de los nodos 401-404 que indican que cada nodo en el clúster puede escribir en el dispositivo 410 de almacenamiento. La tabla 502 de reserva también incluye la clave de registro del nodo 401 que indica que el nodo 401 tiene una reserva en el dispositivo 410 de almacenamiento (es decir el nodo 401 es el propietario actual del disco físico).

Cuando tiene lugar la partición 405 de red, cada nodo en el clúster se notificará de la partición. Sin embargo, no se notificará necesariamente a cada nodo al mismo tiempo. En este ejemplo, el nodo 404 se ha notificado de la partición 405 de red y por consiguiente ha comenzado un desafío para la propiedad de cada dispositivo de

almacenamiento en el grupo 420 (puesto que la partición del nodo 404 necesita la propiedad del grupo 420 para tener quórum).

El nodo 401, sin embargo, en el momento del desafío del nodo 404 no ha sido notificado aún de la partición 405 de red. Como tal, el nodo 401 pensará que está aún en la misma partición que el nodo 404. En PR convencional, cuando el nodo 404 desafía la propiedad del dispositivo 410 de almacenamiento, el hecho de que el nodo 401 aún piense que el nodo 404 está en la misma partición no tendría importancia, y el nodo 401 eliminaría la clave de registro del nodo 404 de la tabla 501 de registro provocando por lo tanto que la E/S del nodo 404 fallara.

En contraste, en la presente invención, se aplican diferentes técnicas PR para asegurar que el nodo 401 no elimina la clave de registro del nodo 404 hasta que el nodo 401 tiene conocimiento de que el nodo 404 ya no está en la misma partición. Las Figuras 5B-5E ilustran cómo se implementan las técnicas PR de la presente invención.

La Figura 5B representa el estado de las dos tablas después de que el nodo 404 ha comenzado un desafío para la propiedad del dispositivo 410 de almacenamiento. Por consiguiente, la Figura 5B puede representar el periodo de tiempo mientras el nodo 404 está inactivo y el nodo 401 no se ha reactivado aún para defender su propiedad.

El nodo 404 intenta reservar realizando las siguientes tareas. El nodo 404 incrementa (o cambia de otra manera) su propia clave de registro y la registra en la tabla 501 de registro (en lugar de su clave de registro antigua). Sin embargo, puesto que el nodo 401 es el propietario actual, el intento del nodo 404 para reservar inicialmente fallará requiriendo por lo tanto que el nodo 404 entre en inactividad antes de intentar adelantarse a la reserva del nodo 401. Puesto que la reserva falló, el nodo 404 también lee la clave de reserva actual (clave del nodo 401 en la tabla 502 de reserva) y la almacena.

Por consiguiente, la Figura 5B, muestra que tabla 501 de registro se ha cambiado sustituyendo la clave de registro antigua del nodo 404 por la clave de registro incrementada del nodo 404.

En la Figura 5B, se muestra el nodo 404 como almacena la clave de reserva actual, la clave de registro del nodo 401.

La Figura 5C representa el estado de las dos tablas después de que el nodo 401 se ha reactivado para defender su propiedad, pero antes de que se haya notificado al nodo 401 de la partición 405 de red. En una defensa, un nodo elimina las claves de registro de cualquiera de los nodos que no se reconocen como que son parte de la partición del nodo propietario. Por consiguiente, cuando el nodo 401 observa la clave de registro incrementada del nodo 404 en la tabla 501 de registro, el nodo 401 no eliminará la clave de registro del nodo 404 puesto que el nodo 401 aún cree que el nodo 404 está en la misma partición.

En lugar de eliminar la clave de registro del nodo 404 de la tabla 501 de registro (evitando por lo tanto que el nodo 404 escriba en el dispositivo 410 de almacenamiento), en su lugar el nodo 401 incrementa su propia clave de registro, registra la clave incrementada, y reserva la clave incrementada. Esta es la forma de defensa del nodo propietario cuando un nodo que no tiene conocimiento que está en una partición diferente desafía la propiedad.

Por consiguiente, la Figura 5C muestra que la tabla 501 de registro y la tabla 502 de reserva ahora cada una incluyen la clave de registro incrementada del nodo 401 en lugar de la clave anterior del nodo 401. En resumen, puesto que el nodo 401 no tiene conocimiento aún de que el nodo 401 y el nodo 404 ya no están en la misma partición, la respuesta del nodo 401 al desafío del nodo 404 implica incrementar y reservar su clave de registro en lugar de eliminar la clave de registro del nodo 404 de la tabla 501 de registro.

La Figura 5D representa el estado de las dos tablas después de que el nodo 404 se ha reactivado y emitido el comando 530 previo. El comando 530 previo incluye la clave de registro actual del nodo 404 y la clave de reserva que el nodo 404 leyó antes de entrar en inactividad. Puesto que el nodo 401 ha incrementado su clave de registro y reservado la clave incrementada, la clave de reserva mantenida por el nodo 404 no coincidirá con la clave de reserva actual almacenada en la tabla 502 de reserva (Nodo_401_clave_0001 != Nodo_401_clave_0002). Por lo tanto, el intento de adelanto del nodo 404 fallará, y el nodo 401 seguirá siendo el propietario del dispositivo 410 de almacenamiento. La Figura 5D muestra en consecuencia que ambas tablas siguen siendo iguales como en la Figura 5C puesto que el nodo 401 sigue siendo el propietario pero no ha eliminado el registro del nodo 404.

En algún tiempo más tarde, el nodo 404 (o de manera similar, el nodo 403) puede comenzar otro desafío de la manera descrita con respecto a las Figuras 5A-5B. La Figura 5E representa el estado de las dos tablas después de que el nodo 404 ha comenzado otro desafío. Como se muestra, la tabla 501 de registro ahora incluye la clave incrementada del nodo 404 (Nodo_404_clave_0003). El nodo 404 tiene almacenada también la clave de reserva actual (Nodo_401_clave_0002).

Antes de que el nodo 401 se reactive para defender su propiedad, se ha notificado al nodo 401 de la partición 405 de red. Por consiguiente, el nodo 401 responde de manera diferente al desafío del nodo 404. En particular, el nodo 401 ahora tiene conocimiento de que el nodo 404 no está en la misma partición, y como tal, el nodo 401 elimina la clave de registro del nodo 404 de la tabla 501 de registro lo que provoca que el desafío del nodo 404 falle así como evita que el nodo 404 escriba en el dispositivo 410 de almacenamiento.

La Figura 5F, por lo tanto, representa el estado de las dos tablas después de que se ha notificado al nodo 401 de la partición 405 de red y después de que el nodo 401 se ha defendido contra el desafío del nodo 404. Como se muestra, la tabla 501 de registro no incluye la clave de registro del nodo 404. Aunque se muestra el nodo 404 almacenando aún la clave de reserva actual, un intento de adelanto por el nodo 404 fallaría puesto que la clave de registro del nodo 404 no está en la tabla 501 de registro.

Aunque la descripción anterior usa el ejemplo de incrementar una clave, la presente invención puede implementarse también cambiando una clave de cualquier otra manera para notificar a otro nodo de un desafío. Por ejemplo, en lugar de incrementar su clave, el nodo 404 podría cambiar un bit. En esencia, el cambio de la clave de registro actúa como una manera para que los nodos continúen comunicando incluso aunque la partición de red haya evitado que los nodos comuniquen directamente. Por consiguiente, cualquier medio de modificación de la clave de registro para comunicar un desafío de un nodo puede usarse en la presente invención.

Se observa que la clave de registro del nodo 404 puede eliminarse también incluso sin que el nodo 404 comience un desafío. En cualquier momento que un nodo 401 se reactiva para defender la propiedad, comprueba en primer lugar cualesquiera claves de registro de nodos que no reconoce como que son parte de la misma partición. Por ejemplo, si el nodo 401 se reactiva después de la partición 405 de red y antes de que el nodo 404 comience un desafío, el nodo 401 eliminará la clave del nodo 404 (así como la clave del nodo 403) de la tabla 501 de registro puesto que los nodos 404 y 403 ya no están en la misma partición.

Por consiguiente, un nodo propietario no elimina la clave de registro del otro nodo hasta que el nodo propietario tiene conocimiento de que el otro nodo no está en la misma partición que el nodo propietario. De esta manera, un nodo propietario no eliminará la clave del otro nodo cuando tiene lugar una partición de red hasta que el nodo propietario tiene conocimiento acerca de la partición de red y puede responder en consecuencia (por ejemplo defendiendo o no defendiendo su propiedad).

Por supuesto, durante cualquier defensa dada, el nodo propietario puede eliminar ambas claves de registro de cualquier nodo que no reconoce como que es parte de la misma partición, así como incrementar, registrar y reservar su clave de registro en respuesta a un desafío desde un nodo que no tiene conocimiento aún que está en otra partición. Usando el mismo ejemplo de las Figuras 5A-5F, si los nodos 403 y 404 se desafiaron al mismo tiempo, pero el nodo 401 había únicamente notificado de que el nodo 404 ya no estaba más en la misma partición, el nodo 401 eliminaría la clave del nodo 404, mientras deja la clave del nodo 403 y aumenta su propia clave.

De manera similar, si en cualquier momento, el nodo 401 tiene conocimiento de que estaba en una partición que no tenía quórum (por ejemplo si la partición 405 de red separó el nodo 401 de los nodos 402-404), el nodo 401 no defendería su propiedad permitiendo de esta manera que uno de los nodos en la otra partición se adelantara satisfactoriamente a la propiedad del nodo 401.

Como se ha mencionado anteriormente, el procedimiento descrito con respecto a las Figuras 5A-5F se realiza para cada dispositivo de almacenamiento en un grupo de almacenamiento. Por ejemplo, el nodo 404 desafiaría la propiedad de los dispositivos 411 y 412 de almacenamiento de la misma manera. Un único nodo, sin embargo, debería tener en general la propiedad de cada dispositivo de almacenamiento en el grupo. Para asegurar que un único nodo obtiene y mantiene la propiedad de cada dispositivo de almacenamiento, el procedimiento descrito anteriormente se lleva a cabo en cada dispositivo de almacenamiento es un orden predefinido.

En otras palabras, cada nodo tiene conocimiento de un orden en el cual debería desafiarse cada dispositivo de almacenamiento. Este orden puede determinarse, por ejemplo, basándose en un identificador asociado con el dispositivo de almacenamiento (por ejemplo identificador único global de grupo). Por ejemplo, cuando el nodo 404 se reactiva para emitir los comandos previos, puede emitir los comandos previos en un orden especificado (tal como adelantándose en el dispositivo 410 de almacenamiento, a continuación dispositivo 411 de almacenamiento, a continuación el dispositivo 412 de almacenamiento).

Si cualquier comando previo falla, el nodo desafiante dejará de desafiar la propiedad. Por ejemplo, si el comando previo del nodo 404 falló en el dispositivo 410 de almacenamiento, el nodo 404 no intentaría adelantarse a la propiedad de los dispositivos 411 y 412 de almacenamiento. Emitiendo los comandos previos en un orden especificado, la situación puede evitarse donde un nodo gana la propiedad de algunos dispositivos de almacenamiento en el grupo, mientras uno más otros nodos ganan la propiedad de otros dispositivos de almacenamiento en el grupo.

Además de asegurar que un nodo propietario o nodo desafiante es parte de una partición que tiene quórum (o podría tener quórum obteniendo la propiedad del grupo), un nodo puede verificar también que un quórum de discos en el grupo son accesibles antes de comenzar una defensa de o desafío para los dispositivos de almacenamiento del grupo. Por ejemplo, cuando el nodo 401 se reactiva para defender su propiedad, puede enumerar en primer lugar todos los dispositivos de almacenamiento en el grupo 420. Si el número de dispositivos de almacenamiento enumerado es menor que una mayoría de los dispositivos de almacenamiento en el grupo (por ejemplo menor que 2 de los dispositivos 410-412 de almacenamiento), el nodo 401 puede dejar su defensa. De manera similar, cuando el nodo 404 intenta un desafío, puede enumerar también los dispositivos de almacenamiento y dejar el desafío si un

quórum de los dispositivos de almacenamiento no es accesible. Un dispositivo de almacenamiento puede ser inaccesible si el dispositivo de almacenamiento falla o deja de otra manera de operar correctamente.

La Figura 6 ilustra un formato ejemplar para una clave 600 de registro. La clave 600 de registro incluye cuatro secciones: una sección 601 de identificador, una sección 602 de revisión, una sección 603 de identificador de nodo y una sección 604 de firma. La sección 601 de identificador incluye un identificador del dispositivo de almacenamiento para el que se usa la clave de registro. La sección 602 de revisión es la porción de la clave que se incrementa como se ha descrito anteriormente. La sección 603 de identificador de nodo incluye un identificador del nodo al cual pertenece la clave. La porción 604 de firma incluye una firma única generada mediante el correspondiente nodo.

La Figura 7 ilustra un diagrama de flujo de un procedimiento 700 de ejemplo para que un primer nodo se defienda contra el intento de otro nodo de adelantarse a la reserva persistente del primer nodo en un dispositivo de almacenamiento. El procedimiento 700 se describirá con respecto a las Figuras 4 y 5A-5F.

El procedimiento 700 incluye un acto 701 de, después de que una partición de red que evita que el primer nodo comunique con otro nodo en el clúster, y antes de que se notifique al primer nodo de la partición de red, el primer nodo detectar que otro nodo en el clúster ha intentado reservar el dispositivo de almacenamiento compartido por los nodos del clúster. La detección comprende identificar que el otro nodo ha cambiado la clave de registro del otro nodo en una estructura de datos de registro. Por ejemplo, el nodo 401 puede detectar que el nodo 404 ha cambiado su clave de registro en la tabla 501 de registro que pertenece al dispositivo 410 de almacenamiento mientras que el nodo 401 tiene una reserva persistente en dispositivo 410 de almacenamiento.

El procedimiento 700 incluye un acto 702 de, el primer nodo cambiar la clave de registro del primer nodo, registrar la clave de registro cambiada en la estructura de datos de registro y reservar la clave de registro cambiada en la estructura de datos de reserva. Por ejemplo, el nodo 401 puede cambiar (por ejemplo incrementar) su clave de registro, registrar la clave de registro cambiada en la tabla 501 de registro, y reservar la clave de registro cambiada en la tabla 502 de reserva.

La Figura 8 ilustra un diagrama de flujo de un procedimiento 800 de ejemplo para que un segundo nodo intente eliminar una reserva persistente del primer nodo en un dispositivo de almacenamiento para obtener una reserva persistente para el segundo nodo. El procedimiento 800 se describirá con respecto a las Figuras 4 y 5A-5F.

El procedimiento 800 incluye un acto 801 de recibir el segundo nodo una notificación de que ha tenido lugar una partición de red que evita que el segundo nodo comunique con el primer nodo. Por ejemplo, el nodo 404 puede recibir la notificación de que ha tenido lugar la partición 405 de red.

El procedimiento 800 incluye un acto 802 del segundo nodo intentar reservar la clave de registro del segundo nodo para obtener una reserva persistente en el dispositivo de almacenamiento. El acto 802 incluye los subactos 802a-802c.

El subacto 802a incluye que el segundo nodo lea la clave de registro del primer nodo que se almacena en una estructura de datos de reserva y almacenar la clave del primer nodo. Por ejemplo, el nodo 404 puede leer la clave de registro del nodo 401 en la tabla 502 de reserva.

El subacto 802b incluye cambiar el segundo nodo la clave de registro del segundo nodo y registrar la clave de registro cambiada. Por ejemplo, el nodo 404 puede cambiar su clave de registro y registrar la clave de registro cambiada en la tabla 501 de registro.

El subacto 802c incluye entrar en inactividad el segundo nodo durante una duración de tiempo especificada antes de emitir un comando previo para eliminar la reserva persistente del primer nodo. Por ejemplo, el nodo 404 puede entrar en inactividad durante al menos dos veces la duración que el nodo defensor entra en inactividad (por ejemplo seis segundos si el nodo 401 entra en inactividad durante tres segundos) antes de reactivar y emitir un comando previo para eliminar la reserva del nodo 401 en el dispositivo 410 de almacenamiento.

La presente invención puede realizarse en otras formas específicas sin alejarse de sus características esenciales. Las realizaciones descritas se han de considerar en todos los aspectos únicamente como ilustrativas y no restrictivas. El alcance de la invención se indica, por lo tanto, mediante las reivindicaciones adjuntas en lugar de mediante la descripción anterior. Todos los cambios que entran dentro del significado y rango de equivalencia de las reivindicaciones se han de abarcar dentro de su alcance.

REIVINDICACIONES

1. Un primer nodo de un clúster, teniendo propiedad el primer nodo de una reserva persistente en un dispositivo de almacenamiento compartido por nodos del clúster, un procedimiento para que el primer nodo se defienda contra el intento de otro nodo para adelantarse a la reserva persistente del primer nodo, comprendiendo el procedimiento:

después de que una partición de red que evita que el primer nodo comunique con otro nodo en el clúster, y antes de que se notifique al primer nodo de la partición de red, detectar el primer nodo que otro nodo en el clúster ha intentado reservar el dispositivo de almacenamiento compartido por los nodos del clúster, comprendiendo la detección identificar que el otro nodo ha cambiado la clave de registro del otro nodo en una estructura de datos de registro; y
cambiar el primer nodo la clave de registro del primer nodo, registrar la clave de registro cambiada en la estructura de datos de registro, y reservar la clave de registro cambiada en una estructura de datos de reserva.

2. El procedimiento de la reivindicación 1, que comprende adicionalmente:

recibir el primer nodo la notificación de la partición de red; y
eliminar el primer nodo la clave de registro del otro nodo de la estructura de datos de registro para evitar que el otro nodo acceda al dispositivo de almacenamiento.

3. El procedimiento de la reivindicación 2, que comprende adicionalmente:

antes de eliminar la clave de registro del otro nodo, determinar el primer nodo que el primer nodo es parte de una partición que tiene quórum.

4. El procedimiento de la reivindicación 1, que comprende adicionalmente:

recibir el primer nodo la notificación de la partición de red;
determinar el primer nodo que el primer nodo es parte de una partición que no tiene quórum; y
fallando al defender el primer nodo contra el intento de otro nodo de adelantarse a la reserva persistente del primer nodo.

5. El procedimiento de la reivindicación 1, en el que la clave de registro cambiada del primer y el otro nodo comprende una versión incrementada de la clave de registro del primer y el otro nodo respectivamente.

6. El procedimiento de la reivindicación 1, en el que la estructura de datos de registro y la estructura de datos de reserva son cada una la misma estructura de datos o estructuras de datos separadas.

7. En un segundo nodo de un clúster, compartiendo el clúster un dispositivo de almacenamiento para el que un primer nodo en el clúster tiene una reserva persistente, un procedimiento para que el segundo nodo intente eliminar la reserva persistente del primer nodo para obtener una reserva persistente para el segundo nodo, comprendiendo el procedimiento:

recibir el segundo nodo una notificación de que ha tenido lugar una partición de red que evita que el segundo nodo comunique con el primer nodo; e
intentar el segundo nodo reservar la clave de registro del segundo nodo para obtener una reserva persistente en el dispositivo de almacenamiento, comprendiendo el intento para reservar:

leer el segundo nodo la clave de registro del primer nodo que se almacena en una estructura de datos de reserva y almacenar la clave del primer nodo;
cambiar el segundo nodo la clave de registro del segundo nodo y registrar la clave de registro cambiada; y
entrar en inactividad el segundo nodo durante una duración de tiempo especificada antes de emitir un comando previo para eliminar la reserva persistente del primer nodo.

8. El procedimiento de la reivindicación 7, que comprende adicionalmente:

reactivar el segundo nodo después de la duración de tiempo especificada;
emitir el segundo nodo un comando previo para intentar eliminar la reserva persistente del primer nodo, incluyendo el comando previo la clave de registro cambiada del segundo nodo, y la clave de registro del primer nodo que se leyó desde la estructura de datos de reserva; y
recibir el segundo nodo la notificación de que el comando previo ha fallado debido a que el primer nodo ha cambiado la clave de registro del primer nodo de manera que la clave de registro del primer nodo en el comando previo no coincide con la versión cambiada actual de la clave de registro del primer nodo que se reserva en la estructura de datos de reserva.

9. El procedimiento de la reivindicación 7, que comprende adicionalmente:

reactivar el segundo nodo después de la duración de tiempo especificada;
emitir el segundo nodo un comando previo para intentar eliminar la reserva persistente del primer nodo,

incluyendo el comando previo la clave de registro cambiada del segundo nodo, y la clave de registro del primer nodo que se leyó desde la estructura de datos de reserva; y
recibir el segundo nodo la notificación de que el comando previo ha tenido éxito debido a que la clave de registro cambiada del segundo nodo se incluye en la estructura de datos de registro y la clave de registro del primer nodo en el comando previo coincide con la clave de registro del primer nodo en la estructura de datos de reserva.

10. El procedimiento de la reivindicación 7, en el que el dispositivo de almacenamiento es parte de un grupo de almacenamiento que incluye una pluralidad de dispositivos de almacenamiento.

11. El procedimiento de la reivindicación 10, que comprende adicionalmente:

reactivar el segundo nodo después de la duración de tiempo especificada;
enumerar el segundo nodo todos los dispositivos de almacenamiento en el grupo que son accesibles; y
fallar el segundo nodo al emitir un comando previo si el número de dispositivos de almacenamiento accesible es menor que una mayoría de los dispositivos de almacenamiento en el grupo.

12. El procedimiento de la reivindicación 10, que comprende adicionalmente:

reactivar el segundo nodo después de la duración de tiempo especificada;
emitir el segundo nodo un comando previo en cada dispositivo de almacenamiento en el grupo en un orden especificado.

13. El procedimiento de la reivindicación 12, que comprende adicionalmente:

si cualquier comando previo falla, dejar el segundo nodo de emitir comandos previos en cualquier dispositivo de almacenamiento adicional en el grupo.

14. El procedimiento de la reivindicación 7, que comprende adicionalmente:

antes de que el segundo nodo intente reservar la clave de registro del segundo nodo, determinar el segundo nodo que el segundo nodo está en una partición que tiene quórum, o que tendría quórum si el segundo nodo obtuvo una reserva persistente en el dispositivo de almacenamiento.

15. Un clúster de nodos que comprende:

un primer nodo que tiene propiedad de una reserva persistente en cada uno de una pluralidad de dispositivos de almacenamiento en un grupo de dispositivos de almacenamiento compartido por los nodos del clúster, realizando el primer nodo lo siguiente para defender su reserva persistente en cada dispositivo de almacenamiento en el grupo:

después de que una partición de red evita que el primer nodo comunique con un segundo nodo en el clúster, y antes de que se notifique al primer nodo de la partición de red, detectar que el segundo nodo en el clúster pretende tener propiedad de cada dispositivo de almacenamiento en el grupo, comprendiendo la detección identificar que el segundo nodo ha añadido una versión incrementada de la clave de registro del segundo nodo a una estructura de datos de registro para cada dispositivo de almacenamiento; y
para cada dispositivo de almacenamiento, incrementar la clave de registro del primer nodo, registrar la clave de registro incrementada en la estructura de datos de registro y reservar la clave de registro incrementada en una estructura de datos de reserva;

el segundo nodo que realiza lo siguiente para intentar obtener una reserva persistente en cada dispositivo de almacenamiento en el grupo tras ser notificado de la partición de red:

para cada dispositivo de almacenamiento, leer la clave de registro del primer nodo que se almacena en la estructura de datos de reserva y almacenar la clave del primer nodo;
para cada dispositivo de almacenamiento, incrementar la clave de registro del segundo nodo,

y registrar la clave de registro incrementada; y
entrar en inactividad durante una duración de tiempo especificada antes de emitir un comando previo en cada uno de los dispositivos de almacenamiento para intentar adelantarse a la reserva persistente del primer nodo en cada dispositivo de almacenamiento.

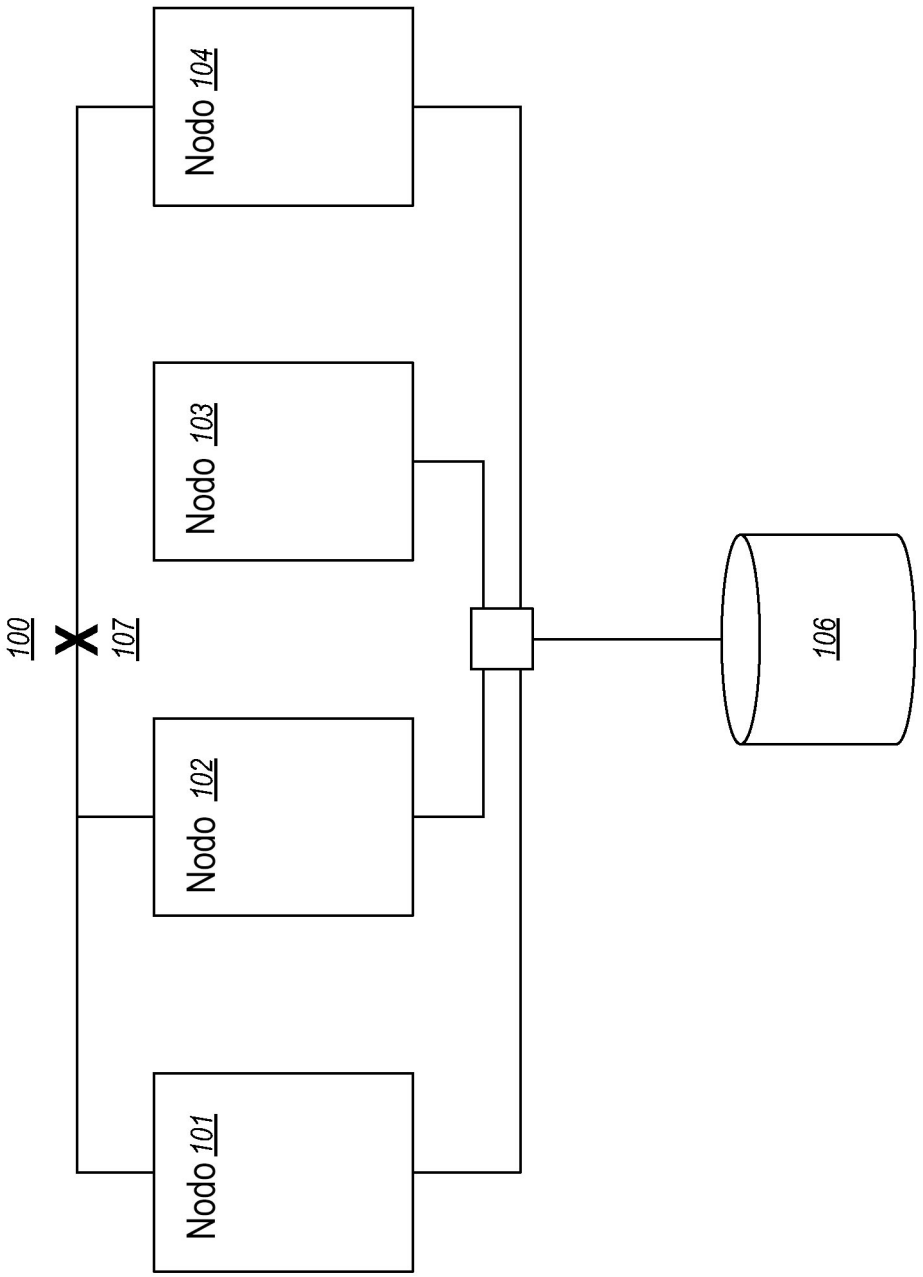

Figura 1
(Técnica Anterior)

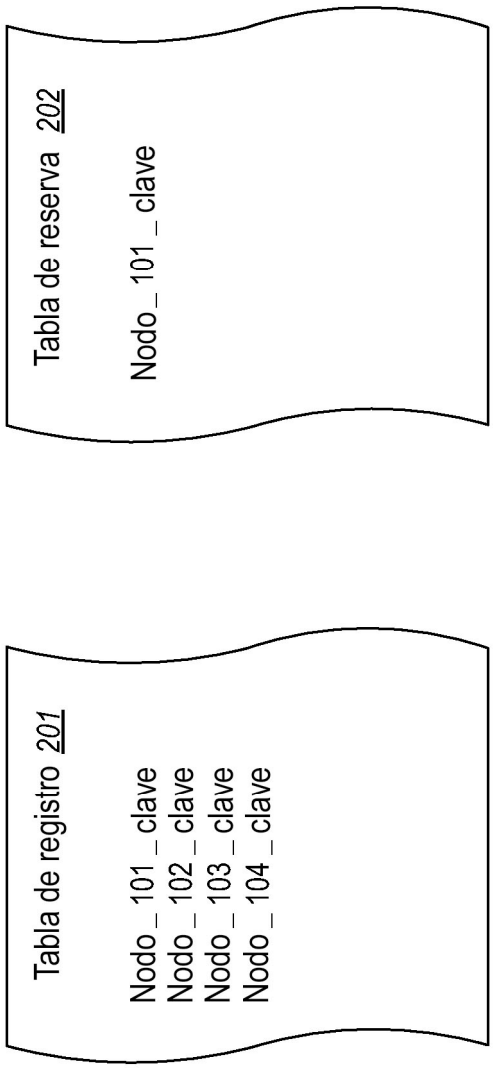

Figura 2A
(Técnica Anterior)

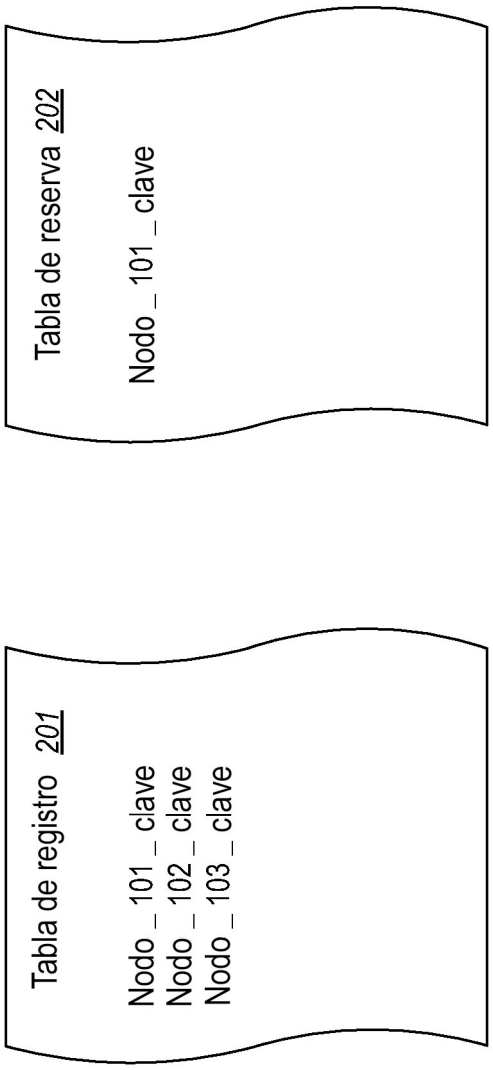

Figura 2B
(Técnica Anterior)

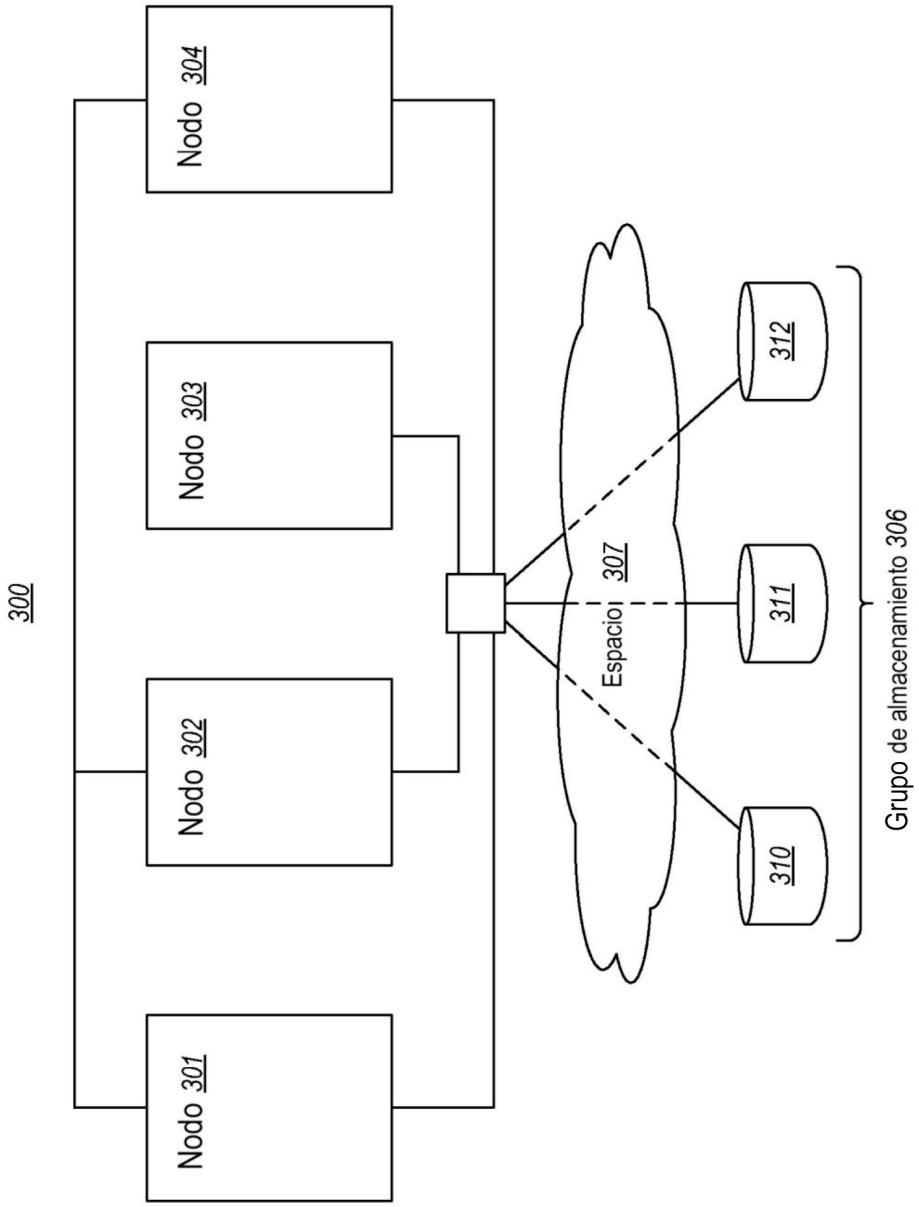

Figura 3 (Técnica Anterior)

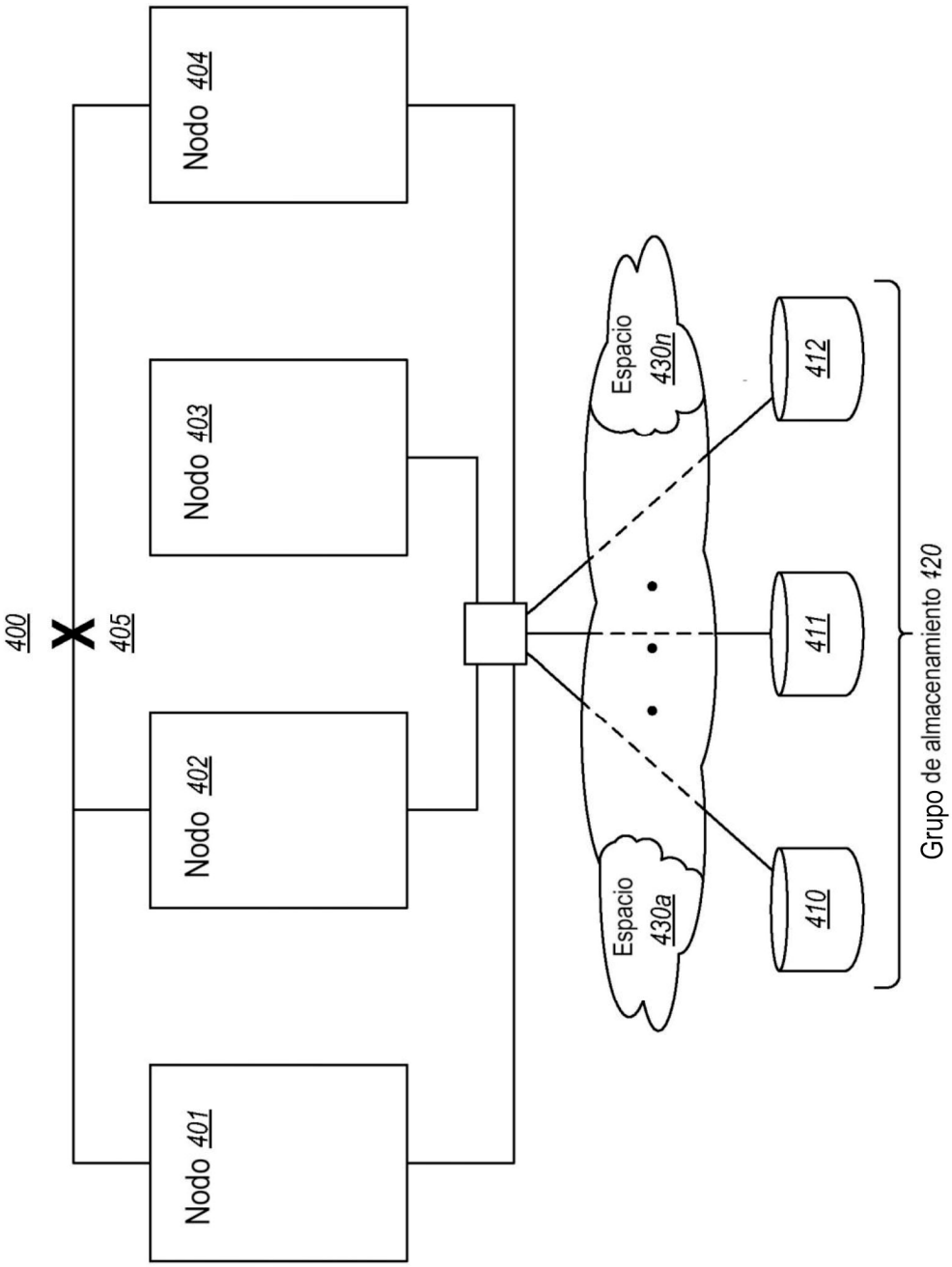
Nodo 404
Nodo 403
Nodo 402
Nodo 401
400
405
Espacio 430n
Espacio 430a
412
411
410
Grupo de almacenamiento 420

Figura 4

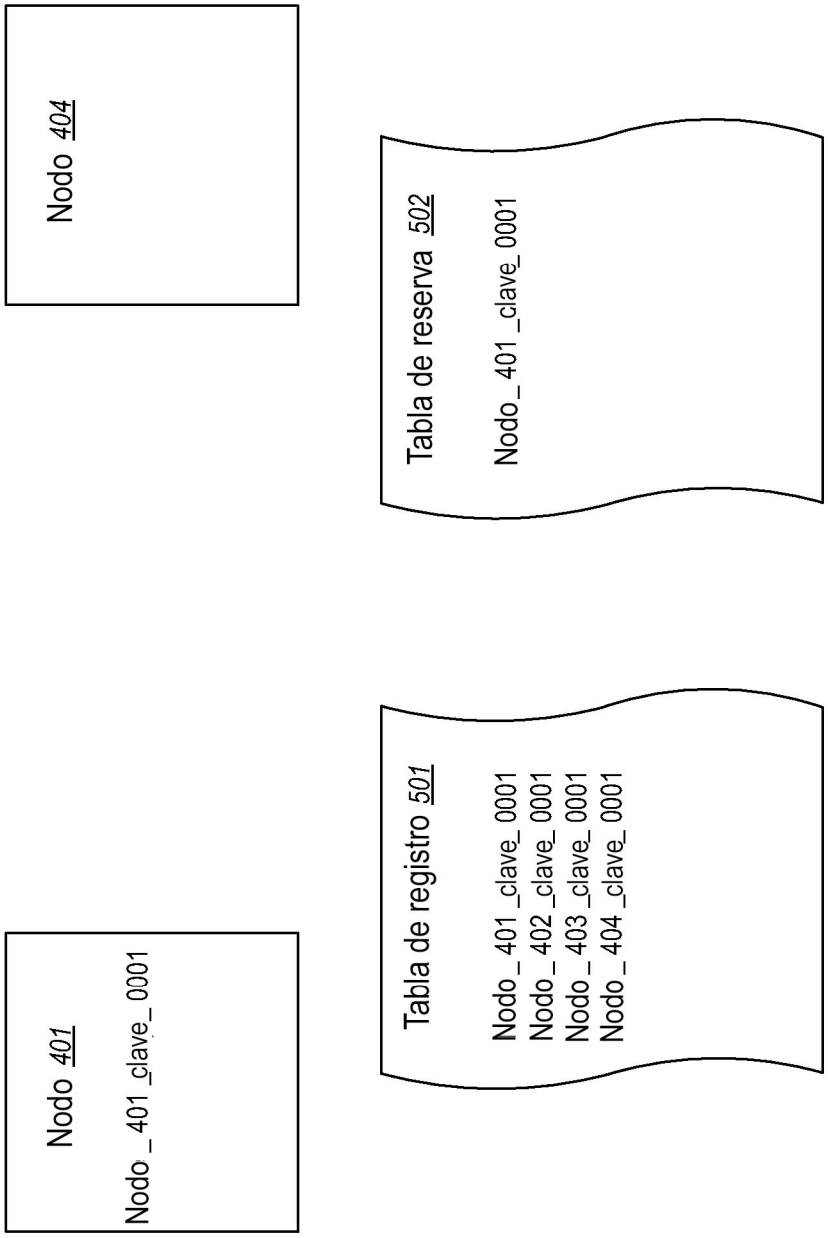
Nodo 401
Nodo _ 401 _clave_ 0001
Nodo 404
Tabla de registro 501
Nodo _ 401 _clave_ 0001
Nodo _ 402 _clave_ 0001
Nodo _ 403 _clave_ 0001
Nodo _ 404 _clave_ 0001
Tabla de reserva 502
Nodo_ 401 _clave_ 0001

Figura 5A

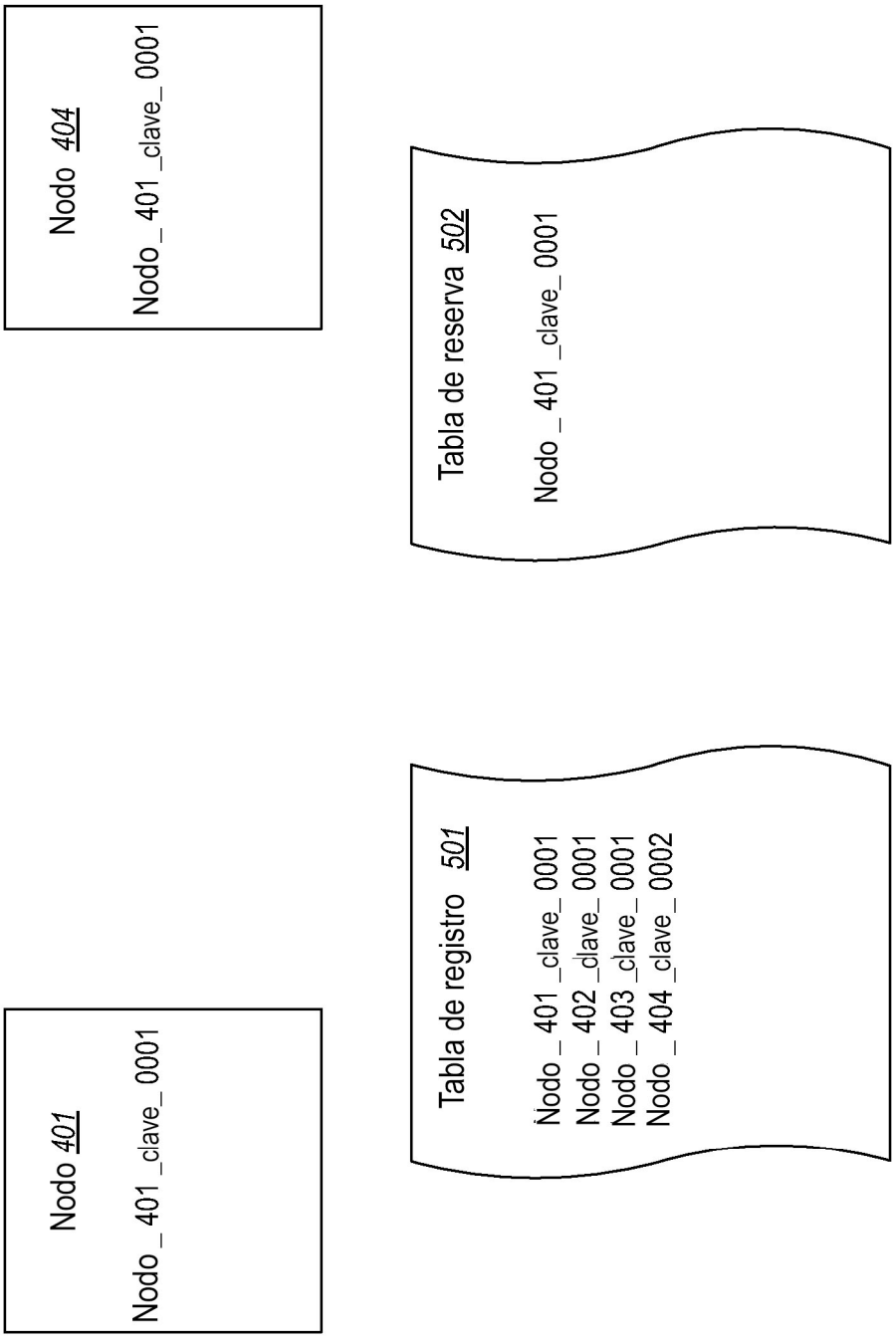
Nodo 401
Nodo _ 401 _clave_ 0001
Nodo 404
Nodo _ 401 _clave_ 0001
Tabla de registro 501
Nodo _ 401 _clave_ 0001
Nodo _ 402 _clave_ 0001
Nodo _ 403 _clave_ 0001
Nodo _ 404 _clave_ 0002
Tabla de reserva 502
Nodo _ 401 _clave_ 0001

Figura 5B

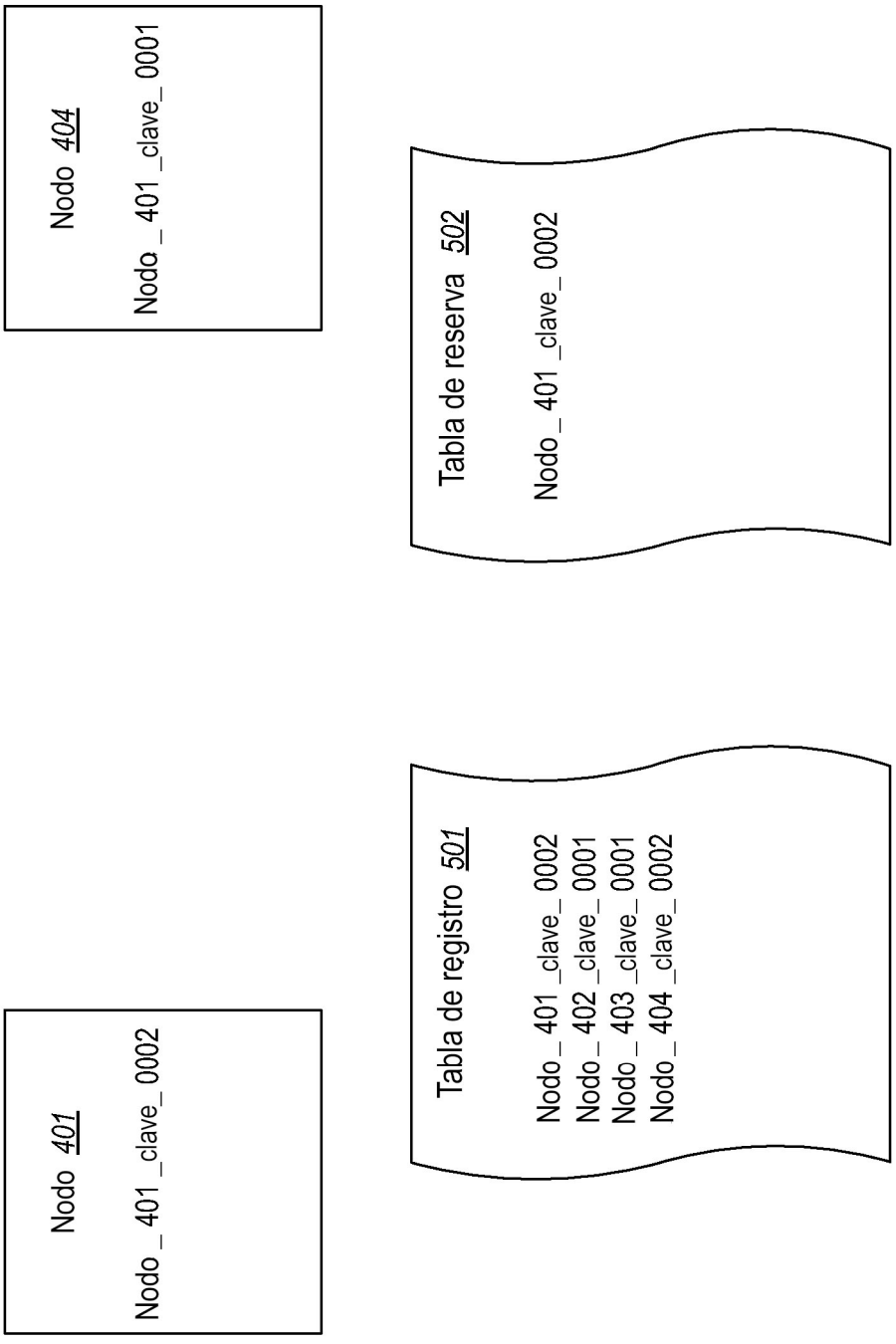
Nodo 404
Nodo _ 401 _clave_ 0001
Tabla de reserva 502
Nodo _ 401 _clave_ 0002
Nodo 401
Nodo _ 401 _clave_ 0002
Tabla de registro 501
Nodo _ 401 _clave_ 0002
Nodo _ 402 _clave_ 0001
Nodo _ 403 _clave_ 0001
Nodo _ 404 _clave_ 0002

Figura 5C

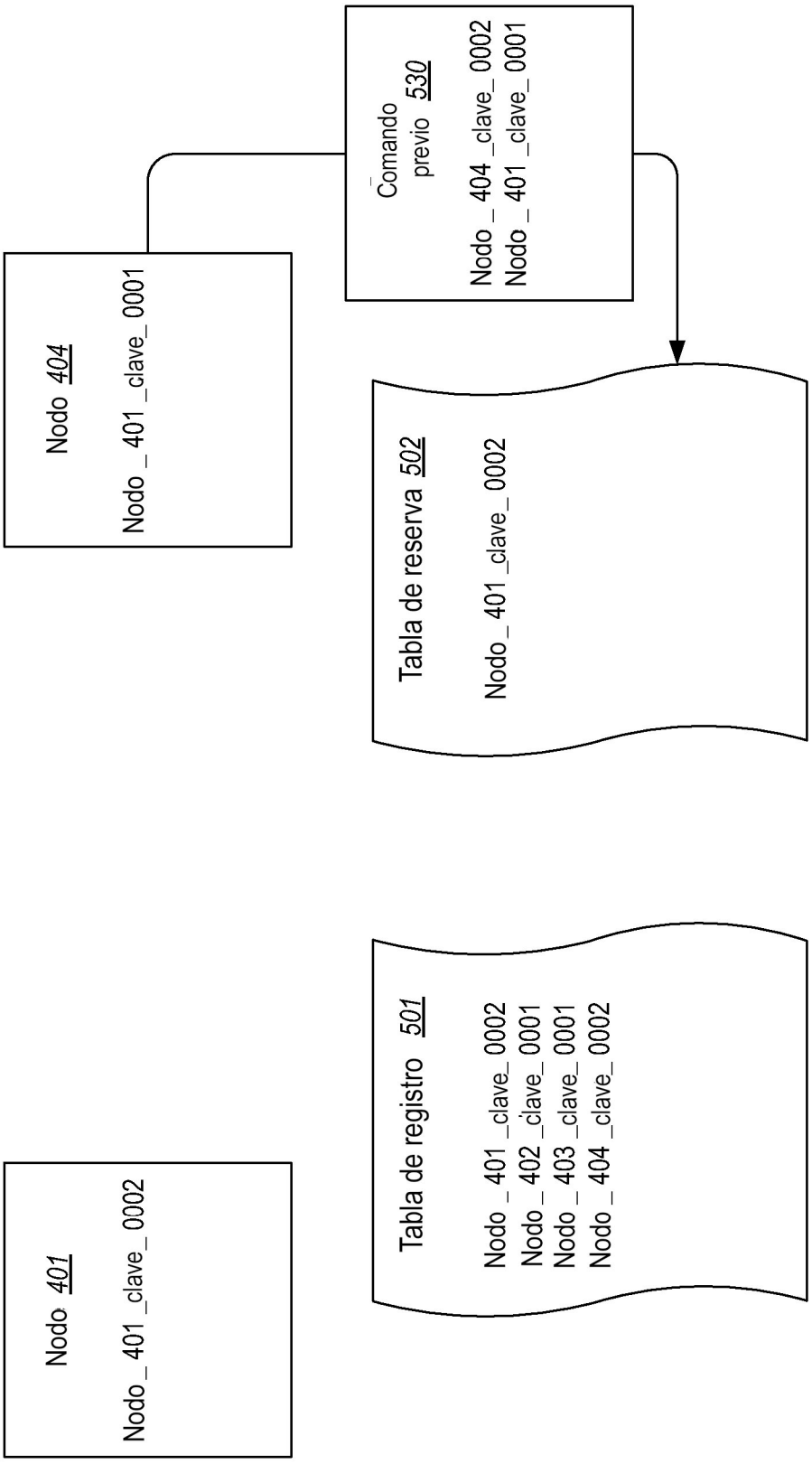
Nodo 401
Nodo _ 401 _clave_ 0002
Nodo 404
Nodo _ 401 _clave_ 0001
Comando previo 530
Nodo _ 404 _clave_ 0002
Nodo _ 401 _clave_ 0001
Tabla de reserva 502
Nodo _ 401 _clave_ 0002
Tabla de registro 501
Nodo _ 401 _clave_ 0002
Nodo _ 402 _clave_ 0001
Nodo _ 403 _clave_ 0001
Nodo _ 404 _clave_ 0002

Figura 5D

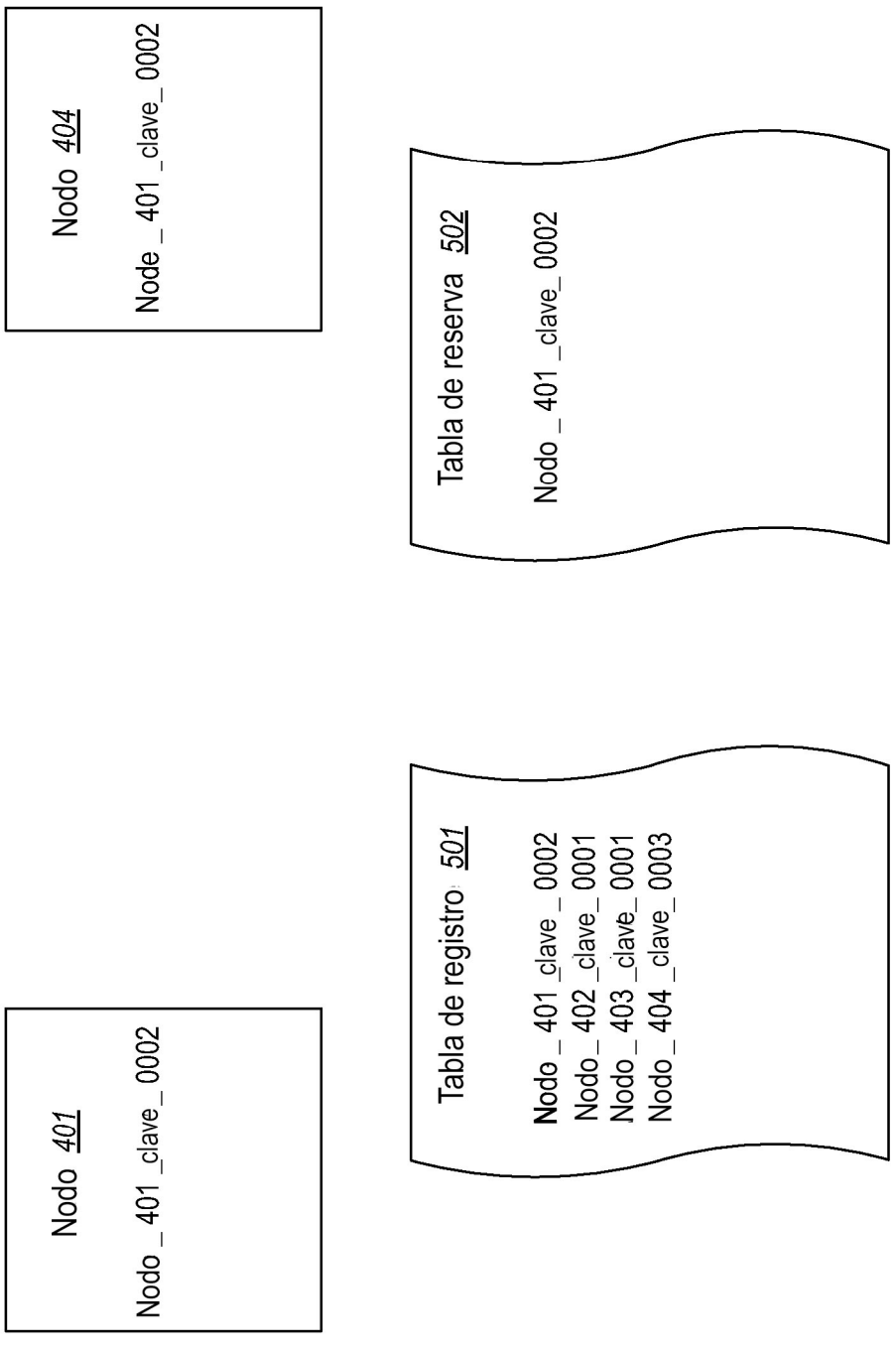
Nodo 404
Node _ 401 _clave_ 0002
Tabla de reserva 502
Nodo _ 401 _clave_ 0002
Nodo 401
Nodo _ 401 _clave_ 0002
Tabla de registro 501
Nodo _ 401 _clave_ 0002
Nodo _ 402 _clave_ 0001
Nodo _ 403 _clave_ 0001
Nodo _ 404 _clave_ 0003

Figura 5E

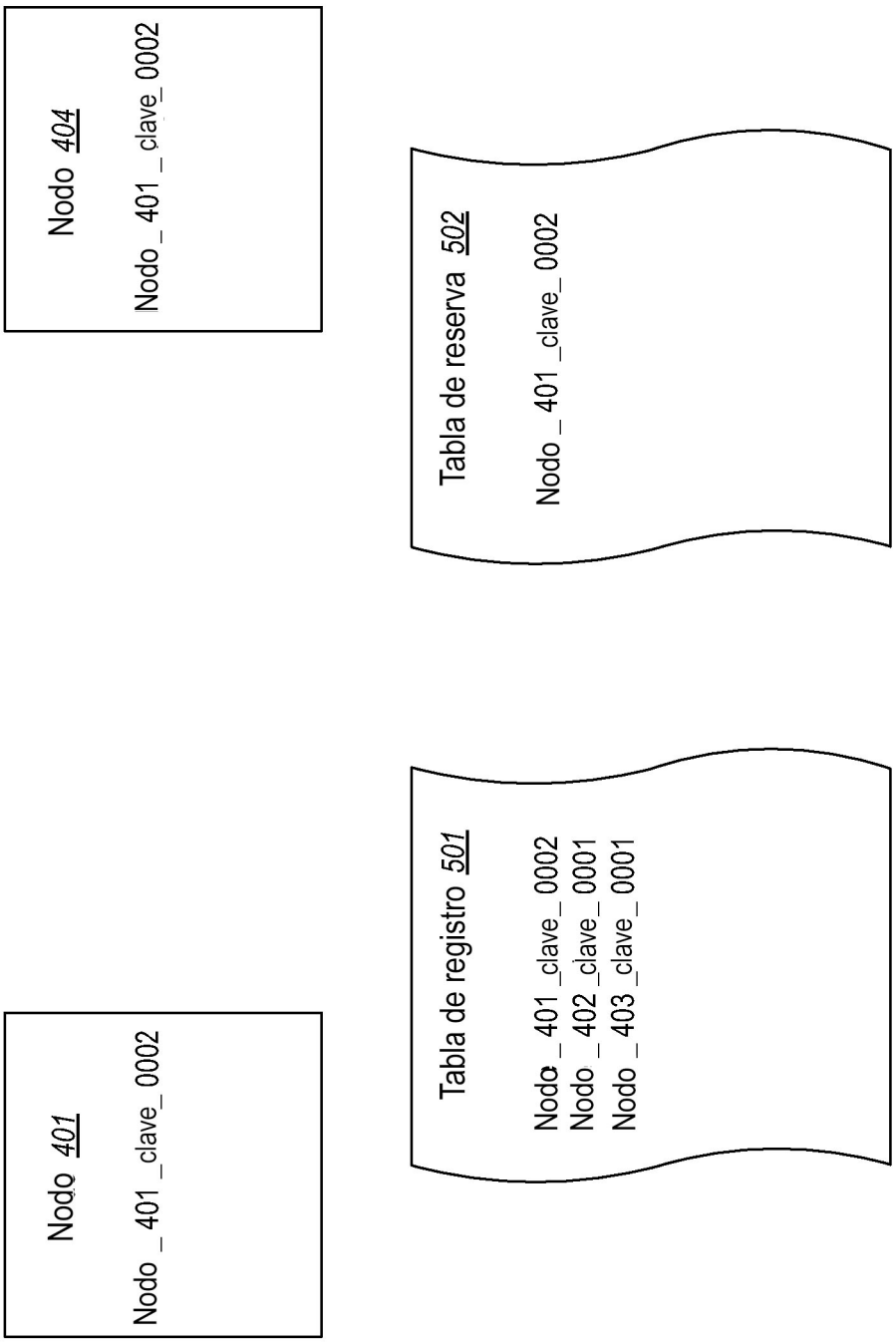
Nodo 404
Nodo _ 401 _ clave_ 0002
Tabla de reserva 502
Nodo _ 401 _clave_ 0002
Nodo 401
Nodo _ 401 _clave_ 0002
Tabla de registro 501
Nodo _ 401 _clave_ 0002
Nodo _ 402 _clave_ 0001
Nodo _ 403 _clave_ 0001

Figura 5F

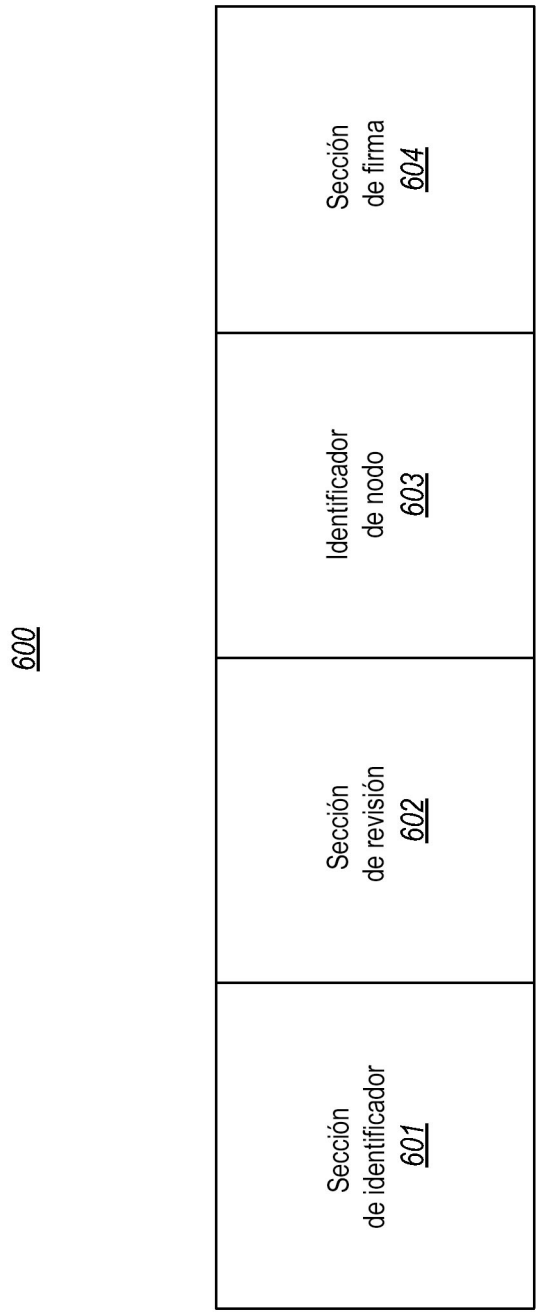
600
Sección de identificador 601
Sección de revisión 602
Identificador de nodo 603
Sección de firma 604

Figura 6

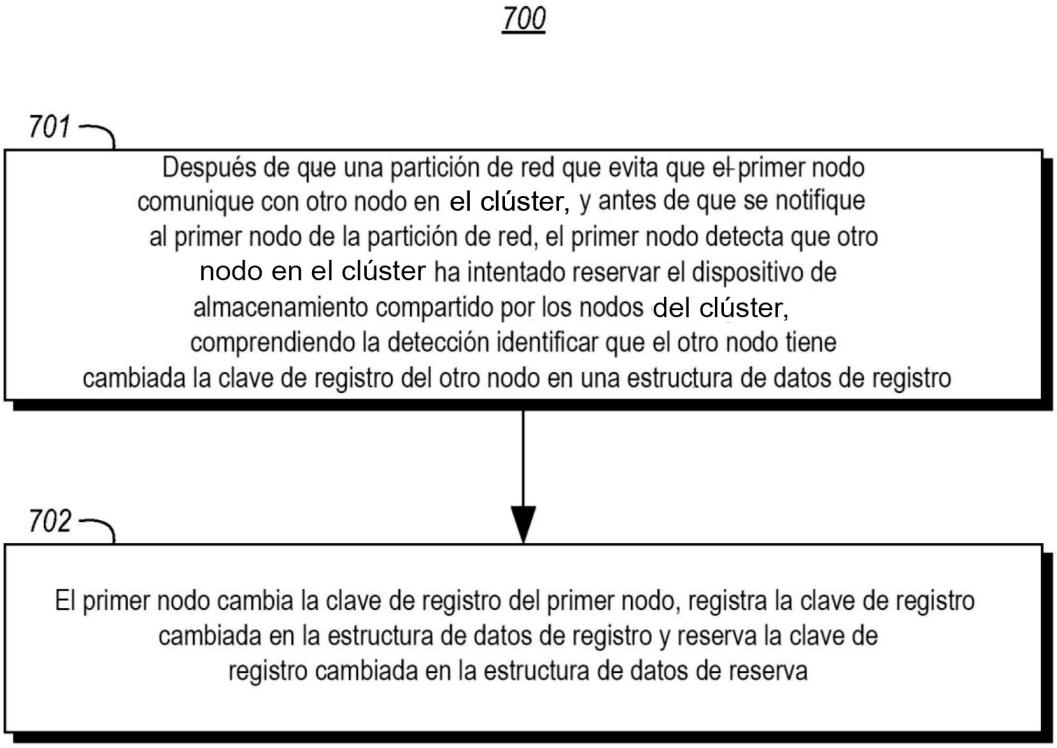
700
701
Después de que una partición de red que evita que el primer nodo comunique con otro nodo en el clúster, y antes de que se notifique al primer nodo de la partición de red, el primer nodo detecta que otro nodo en el clúster ha intentado reservar el dispositivo de almacenamiento compartido por los nodos del clúster, comprendiendo la detección identificar que el otro nodo tiene cambiada la clave de registro del otro nodo en una estructura de datos de registro
702
El primer nodo cambia la clave de registro del primer nodo, registra la clave de registro cambiada en la estructura de datos de registro y reserva la clave de registro cambiada en la estructura de datos de reserva

Figura 7

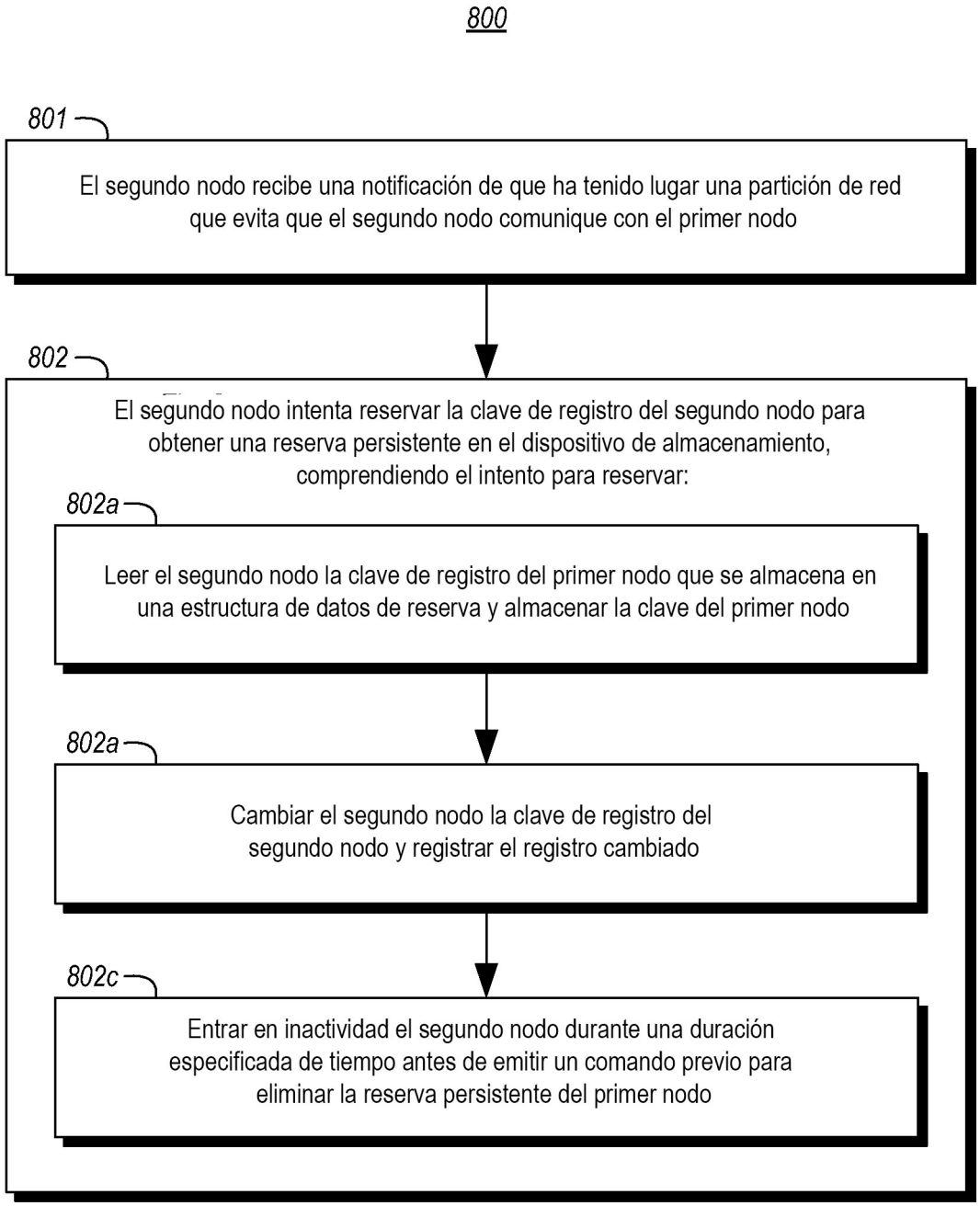
800
801
El segundo nodo recibe una notificación de que ha tenido lugar una partición de red que evita que el segundo nodo comunique con el primer nodo
802
El segundo nodo intenta reservar la clave de registro del segundo nodo para obtener una reserva persistente en el dispositivo de almacenamiento, comprendiendo el intento para reservar:
802a
Leer el segundo nodo la clave de registro del primer nodo que se almacena en una estructura de datos de reserva y almacenar la clave del primer nodo
802a
Cambiar el segundo nodo la clave de registro del segundo nodo y registrar el registro cambiado
802c
Entrar en inactividad el segundo nodo durante una duración especificada de tiempo antes de emitir un comando previo para eliminar la reserva persistente del primer nodo

Figura 8