(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 535 943**

(51) Int. CI.:

G06F 17/30 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **11.05.2012 E 12724447 (3)**

(97) Fecha y número de publicación de la concesión europea: **22.04.2015 EP 2712452**

(54) Título: **Propagación de recuento de referencia**

(30) Prioridad:

13.05.2011 US 201113106927

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
19.05.2015

(73) Titular/es:

SIMPLIVITY CORPORATION (100.0%)
8 Technology Drive
Westborough, MA 01581-1756, US

(72) Inventor/es:

BEAVERSON, ARTHUR J.;
CHITRAPU, KISHORE;
CZERKOWICZ, JOHN MICHAEL y
MANJANATHA, SOWMYA

(74) Agente/Representante:

UNGRÍA LÓPEZ, Javier

DESCRIPCIÓN

Propagación de recuento de referencia

Campo de la invención

La presente invención se refiere al almacenamiento de objetos de datos, y más en particular a un método y aparato para el seguimiento de objetos almacenados en una pluralidad de nodos en una red de iguales manteniendo al mismo tiempo una visión global de las referencias de los objetos.

Antecedentes

A efectos de rendimiento o redundancia, los modernos sistemas de almacenamiento pueden estar formados a partir de componentes discretos, llamados nodos, que están interconectados con una red, tal como una red TCP/IP. Cada nodo es típicamente un ordenador plenamente funcional con CPUs, almacenamiento, memoria, etc. La organización de los nodos en dicho sistema puede ser una red de iguales, lo que significa que todos los nodos son iguales (es decir, no hay autoridad de gestión central y ningún nodo está privilegiado). Como compañeros iguales, los nodos comunican entre sí para resolver el estado. Organizar un sistema de almacenamiento como una red de iguales puede proporcionar una solución de almacenamiento más elástica y escalable, puesto que se puede añadir nodos incrementalmente para rendimiento y/o capacidad, y si un nodo falla, el sistema de almacenamiento sigue funcionando.

Lo que distingue un sistema de almacenamiento de red de iguales de una colección de ordenadores es que los nodos del sistema comunican uno con otro con respecto al almacenamiento de datos subyacente, la sanidad de cada nodo, etc. Específicamente, los nodos del sistema pueden copiar e intercambiar información a efectos de rendimiento e integridad de datos. Esta información puede estar en forma de objetos de datos, o archivos, donde un objeto puede ser una porción de un archivo.

Dado que los objetos se propagan a través del sistema, se requieren estructuras de datos para: a) conocer dónde están los objetos; y b) conocer cuándo los objetos ya no son necesarios. En los sistemas de la técnica anterior donde los objetos tienen recuentos de referencia, es decir, el número de veces que un objeto es referenciado por otro objeto u otra estructura de datos, un objeto puede ser desasignado o borrado de forma segura (por ejemplo, recogida de basura) cuando su recuento de referencias cae a cero.

Sin embargo, hacer el seguimiento de recuentos de referencia relativos a miles de millones de objetos, cuando se pasan millones de objetos por segundo, da lugar a tráfico de red y costos de CPU inaceptables si se usan algoritmos simplistas.

Otro reto es determinar que el recuento de referencias es realmente cero, y luego hallar todas las instancias de objetos para poder borrarlos. Hay una necesidad creciente de protocolos más eficientes y fiables para seguimiento de objetos con el fin de superar estos problemas.

US2011022566 describe un sistema de archivos que tiene un sistema de archivos de firma digital donde los datos, los metadatos y los archivos son considerados como objetos cuyas referencias son mapeadas por las huellas. El objeto está provisto de la huella globalmente única y derivada de contenido. Un objeto raíz dispone de las huellas de todos los objetos de mapeado. Los cambios en el objeto raíz son rastreados para obtener la historia de la actividad del sistema de archivos.

XP040153040 (B. GOLDBERG: “Generational Reference Counting: A Reduced-Communication Distributed Storage Reclamation Scheme”) describe un algoritmo de recuento de referencia generacional (GRC) que implementa un recuento local de referencias que permiten valores negativos con el fin de lograr un sistema libre de error en el caso de mensajes de referencia/desreferencia asíncronos en recuperación de almacenamiento en un almacenamiento de objetos distribuidos.

Resumen de la invención

En un aspecto de la invención, se facilita un sistema y método para el seguimiento de referencias de objetos a través de una pluralidad de nodos de red, donde los objetos están distribuidos entre los nodos almacenando una o más instancias de un objeto (las instancias de un objeto contienen datos idénticos) en uno o más nodos de la red. En esta red, las instancias de un objeto pueden ser todas iguales (de iguales), en contraposición a las instancias entre las que hay alguna relación de jerarquía o maestro/esclavo, es decir, una instancia es primaria o más privilegiada que otra. Estas instancias de iguales pueden ser gestionadas colectivamente por los nodos de red, sin un agente de gestión centralizada. Como se describe más adelante según varias realizaciones de la invención, se facilitan métodos y sistemas para el seguimiento de estas instancias de objetos almacenadas en una pluralidad de nodos de red, seguimiento que permite una determinación global de cuándo un objeto no tiene referencias a través de los nodos en red y puede ser desasignado de forma segura.

Según un aspecto de la invención, cada nodo tiene un almacenamiento local de objetos para efectuar el seguimiento y opcionalmente almacenar objetos en el nodo, y los almacenamientos locales de objetos comparten colectivamente las instancias localmente almacenadas de los objetos a través de la red. Una o más aplicaciones, por ejemplo, un sistema de archivos y/o un sistema de almacenamiento, usan los almacenamientos locales de objetos para almacenar todos los datos persistentes de la aplicación como objetos. La aplicación puede requerir un recuento de referencias para cada objeto almacenado, incluyendo el recuento de referencias el número de referencias al objeto respectivo. Según un aspecto de la invención, el recuento global de referencias para un objeto puede ser rastreado manteniendo en cada nodo un recuento local de referencias LRC a nombres de objeto (en contraposición a instancias de objetos) en el nodo respectivo, incluyendo el valor del LRC una suma de ajustes de recuento de referencias en el nodo local, donde el LRC es independiente de cualquier instancia del objeto almacenado en el almacenamiento local de objetos. Desacoplando el recuento de referencias del recuento de instancias, este método permite valores negativos (enteros con signo) de un LRC, en contraposición a los métodos de la técnica anterior. Además, calculando una suma de los LRCs a partir de una pluralidad de nodos, aquí referida como un recuento global de referencias GRC, el GRC resultante puede ser usado para determinar si, para una aplicación particular que use los nodos en red para almacenar datos persistentes como objetos, es seguro desasignar un objeto (todas sus instancias) porque el objeto ya no está siendo referenciado por la aplicación. Si y solamente si se determina que el GRC es cero, es seguro desasignar las instancias de objetos para un objeto particular.

Según otro aspecto de la invención, la colocación de instancias de objetos en uno o varios nodos se lleva a cabo independientemente del nombre del objeto. En contraposición a los métodos de la técnica anterior que fuerzan la posición de almacenamiento de objetos en base al nombre de objeto, en varias realizaciones de la presente invención la colocación de instancias de objetos en los nodos puede ser determinada por el uso real o previsto de un objeto, por ejemplo, en base al rendimiento de la red o del sistema o la fiabilidad de los datos. Por ejemplo, un almacenamiento local de objetos, conociendo qué sistema(s) de archivo hace(n) referencia a un objeto concreto, puede determinar una colocación preferida del objeto en uno o varios nodos dependiendo del (de los) sistema(s) de archivo que use(n) el objeto. Este sistema es sustancialmente más robusto y eficiente que los sistemas de la técnica anterior que restringen la colocación de objetos de datos en nodos dependiendo de qué sea el objeto, por ejemplo, el nombre del objeto.

Según otro aspecto de la invención, cada nodo local mantiene su propio índice local para mapear nombres de objeto a posiciones físicas en el nodo local donde se almacenan los objetos. Este índice de mapeado puede incluir, por ejemplo, señalizadores a una o más posiciones en el nodo local donde se almacena la instancia de objeto. Las entradas del índice de mapeado local pueden incluir, para un nombre de objeto dado, el LRC y los señalizadores a una o más posiciones de almacenamiento en el nodo local.

Según una realización de la invención, se facilita un medio legible por ordenador conteniendo instrucciones de programa ejecutables para realizar un método incluyendo:

en una pluralidad de nodos en red para compartir objetos almacenados, los objetos tienen nombres globalmente únicos a través de los nodos en red y donde los nombres de objeto no cambian en base a dónde se almacenen los objetos en los nodos;

mantener, en cada nodo, un recuento local de referencias LRC a nombres de objeto, independientemente de cualquier instancia de objeto almacenado en el nodo local, manteniéndose el LRC como un entero con signo,

donde una desreferencia de un nombre de objeto puede generar un valor LRC negativo.

En una realización, el método incluye:

determinar un recuento global de referencias GRC incluyendo una suma de los LRCs.

En una realización el paso de determinación incluye:

identificar instancias de objetos para desasignación cuando se determina que el GRC es cero.

En una realización el paso de determinación incluye:

implementar un protocolo de red de consistencia eventual entre los nodos para determinar si los objetos pueden ser desasignados.

En una realización el paso de determinación incluye:

asignar la propiedad de un rango de identificadores de objeto a uno de los nodos de red,

donde el nodo inicia el paso de determinación para todo o para un subconjunto del rango.

En una realización el paso de método incluye:

mantener un almacenamiento local de objetos en cada nodo, donde los almacenamientos locales de objetos incluyen colectivamente un espacio de nombre de los nombres de objeto globalmente únicos.

En una realización el método incluye:

el almacenamiento local de objetos mantiene un índice de mapeado local de nombre de objeto, LRC y un indicador a cualquier posición física de objeto donde se almacena el objeto en el nodo local.

En una realización el método incluye:

cada objeto tiene una huella de objeto derivada del contenido de objetos como su nombre de objeto.

En una realización el método incluye:

la huella incluye un hash del contenido de objetos.

En una realización el método incluye:

un sistema de archivos que usa los almacenamientos de objetos en uno o una pluralidad de los nodos en un sistema de almacenamiento de red de iguales colectivamente como un método para almacenar todos los datos persistentes del sistema de archivos.

En una realización el método incluye:

* todos los datos de sistema de archivos, los metadatos y los archivos incluyen objetos del (de los) almacenamiento(s) de objetos, teniendo cada objeto una huella de objeto como su nombre de objeto;

* las colecciones de objetos del sistema de archivos también incluyen objetos del (de los) almacenamiento(s) de objetos, incluyendo cada colección un mapeado de una pluralidad de los objetos del sistema de archivos y teniendo su propia huella de objeto derivada del contenido de la colección, donde un cambio en uno o más objetos de la colección cambia la huella de objeto de la colección; y

* un objeto raíz del sistema de archivos que tiene una huella de objeto raíz, de tal manera que cada objeto en el sistema de archivos de espacio de nombres es accesible a través del objeto raíz.

En una realización el método incluye:

seleccionar, en base al rendimiento o la fiabilidad de la red o del sistema, uno o varios nodos como posición(es) para almacenar una o más instancias de un objeto independientemente del nombre de objeto.

En una realización el método incluye:

la compartición de los objetos almacenados incluye comunicar entre nodos con respecto a nombres de objeto, LRCs y posiciones de objetos almacenados.

En una realización el método incluye:

cuando una aplicación desreferencia un nombre de objeto y un ejemplo del objeto no está almacenado en un almacenamiento local de objetos, el almacenamiento local de objetos genera un LRC de uno negativo.

En una realización el método incluye:

un sistema de almacenamiento que usa los almacenamientos de objetos en uno o una pluralidad de los nodos en un sistema de almacenamiento de red de iguales colectivamente como un método para almacenar todos los datos persistentes del sistema de almacenamiento.

Según otra realización de la invención, se facilita un medio legible por ordenador conteniendo instrucciones ejecutables para realizar

un método de seguimiento de referencias a objetos de datos en un sistema de almacenamiento de iguales distribuido que opera en una pluralidad de nodos en red, teniendo los objetos nombres globalmente únicos a través de los nodos en red, teniendo cada nodo un almacenamiento local de objetos para nombrar y almacenar objetos en el nodo local, y compartiendo los almacenamientos locales de objetos los objetos,

incluyendo el método:

determinar, en cada uno de una pluralidad de nodos locales, una o más posiciones para almacenar una instancia de un objeto de datos en el nodo local;

mapear, en cada uno de una pluralidad de nodos locales, la una o más posiciones de la instancia almacenada en el nodo local al nombre de objeto, donde la posición no cambia el nombre de objeto; y

mantener, en cada uno de una pluralidad de nodos locales, un recuento local de referencias LRC a nombres de objeto, incluyendo el valor del LRC una suma de ajustes de recuento de referencias para el objeto en el nodo local, donde el LRC es independiente de si alguna instancia del objeto referenciado está almacenada en el nodo local.

Según otra realización de la invención, se facilita un medio legible por ordenador conteniendo instrucciones de programa ejecutables para realizar un método incluyendo:

para una pluralidad de nodos en red, cada nodo tiene un almacenamiento local de objetos para almacenar objetos en el nodo respectivo, teniendo los objetos nombres globalmente únicos a través de los nodos en red;

los almacenamientos locales de objetos comparten uno o varios objetos; y

una aplicación que usa los almacenamientos locales de objetos para almacenar datos persistentes como objetos, requiriendo la aplicación un recuento de referencias para cada objeto almacenado en los nodos en red incluyendo un número de referencias al objeto respectivo;

mantener un recuento local de referencias LRC a nombres de objeto en cada nodo, incluyendo el valor del LRC una suma de ajustes de recuento de referencias en el nodo local, donde el LRC es independiente de cualquier instancia del objeto referenciado almacenado en el almacenamiento local de objetos para permitir un valor negativo del LRC; y

determinar un recuento global de referencias GRC incluyendo una suma de los LRCs.

En una realización:

la aplicación incluye un sistema de archivos o un sistema de almacenamiento.

Según otra realización de la invención, se facilita un aparato incluyendo:

una pluralidad de nodos en red, teniendo cada nodo un almacenamiento local de objetos para almacenar objetos en el nodo respectivo y teniendo los objetos nombres globalmente únicos a través de los nodos en red;

compartiendo los almacenamientos locales de objetos uno o varios objetos; y

una aplicación que usa los almacenamientos locales de objetos para almacenar datos persistentes como objetos, requiriendo la aplicación un recuento de referencias para cada objeto almacenado en los nodos en red incluyendo un número de referencias al objeto respectivo;

medios para mantener un recuento local de referencias LRC a nombres de objeto en cada nodo, incluyendo el valor del LRC una suma de ajustes de recuento de referencias en el nodo local, donde el LRC es independiente de cualquier instancia del objeto referenciado almacenado en el almacenamiento local de objetos para permitir un valor negativo del LRC; y

medios para determinar un recuento global de referencias GRC incluyendo una suma de los LRCs.

Según otra realización de la invención, se facilita un aparato incluyendo:

una pluralidad de nodos en red, teniendo cada nodo un almacenamiento local de objetos para almacenar objetos en el nodo respectivo y teniendo los objetos nombres globalmente únicos a través de los nodos en red;

compartiendo los almacenamientos locales de objetos uno o varios objetos; y

una aplicación que usa los almacenamientos locales de objetos para almacenar datos persistentes como objetos, requiriendo la aplicación un recuento de referencias para cada objeto almacenado en los nodos en red incluyendo un número de referencias al objeto respectivo;

medios para mantener un recuento local de referencias LRC a nombres de objeto en cada nodo, incluyendo el valor del LRC una suma de ajustes de recuento de referencias en el nodo local, donde el LRC es independiente de

cualquier instancia del objeto referenciado almacenado en el almacenamiento local de objetos para permitir un valor negativo del LRC; y

medios para seleccionar, en base al rendimiento o la fiabilidad de la red o del sistema, uno o más de los almacenamientos locales de objetos como posiciones para almacenar instancias de objetos independientemente del nombre de objeto.

Estos y otros aspectos de la presente invención se describen con más detalle a continuación con respecto a varias realizaciones detalladas.

Breve descripción de los dibujos

Los dibujos acompañantes ilustran una o más realizaciones de la presente invención donde:

La figura 1 es una ilustración esquemática de una pluralidad de nodos en red para llevar a la práctica una realización de la presente invención.

La figura 2 es una ilustración esquemática de un almacenamiento local de objetos en un nodo de la red que comunica con la red mediante un broker de objetos.

La figura 3 es una ilustración esquemática de la compartición de objetos entre dos nodos de red Alicia y Bob.

La figura 4 es una ilustración esquemática de un índice de mapeado local en el nodo Bob.

La figura 5 es una ilustración esquemática de una reconciliación de recuentos globales de referencias entre los nodos Alicia y Bob de la red.

La figura 6 es una ilustración esquemática de una visión lógica del sistema de archivos, que ilustra múltiples referencias independientes a un objeto de la vista de la aplicación.

La figura 7 ilustra un ejemplo de seguimiento de referencias de objetos en tres nodos de la red, donde se permiten recuentos de referencia negativos según la presente invención.

La figura 8 es un diagrama de flujo de comunicaciones entre los dos nodos Alicia y Bob en la red para mantener recuentos locales de referencias.

Y la figura 9 es una ilustración esquemática de comunicaciones de red entre tres nodos Alicia, Bob y Eva para implementar una reconciliación tanto de recuentos de referencia como de recuentos de instancias a través de los nodos de red.

Descripción detallada

En una realización de la invención un sistema de almacenamiento de datos está distribuido a través de una pluralidad de nodos en una red, teniendo cada nodo su propio almacenamiento local de objetos y compartiendo objetos los almacenamientos locales de objetos. Este grupo de nodos en red sigue una convención de denominación universal, donde cada objeto almacenado en la red tiene un nombre globalmente único a través de la red. En esta realización, la pluralidad de nodos operan como una red de iguales, no siendo primario ningún nodo. No hay autoridad de denominación centralizada; en cambio, agentes locales en cada uno de los nodos locales pueden determinar independientemente y ponerse de acuerdo en el nombre de objeto utilizando un algoritmo hash común para determinar el nombre del objeto, sin requerir comunicación entre los nodos. Usando un algoritmo hash fuerte se evitan colisiones.

En esta realización, se prevé que la pluralidad de almacenamientos locales de objetos gestionen miles de millones de objetos, siendo cada objeto pequeño, por ejemplo, de 4, 8, 16, 32, 64, 128 o 256KB (kilobytes) de longitud. Los objetos en el almacenamiento de objetos son opacos, incluyendo una pluralidad arbitraria de bits. El almacenamiento de objetos no tiene conocimiento semántico de los bits; más bien, una aplicación de capa superior, tal como un sistema de archivos, o alternativamente un sistema de almacenamiento de bloques, define el contenido (significado) semántico y las relaciones o referencias entre objetos. Cada almacenamiento local de objetos contiene un índice local para mapear un nombre de objeto a una posición física (por ejemplo, señalizador a una posición física) donde una instancia de objeto puede estar almacenada localmente. Cada almacenamiento local de objetos también mantiene un recuento local de referencias LRC para un nombre de objeto. Sin embargo, en contraposición a la técnica anterior, el recuento local de referencias de un nombre de objeto está desacoplado (es independiente) de las instancias de objetos. El LRC se mantiene como un entero con signo, y consta de una suma parcial de referencias a nombres de objeto en contraposición a un recuento de instancias de objetos. También distingue al LRC su mantenimiento como un entero con signo, permitiendo así valores negativos. Se hacen consiguientemente ajustes en el LRC para cada nueva referencia y desreferencia, es decir, +1 para añadir una referencia, y -1 para

quitar una referencia.

En un ejemplo, una aplicación de sistema de archivos mantiene un recuento global de referencias incluyendo el número de veces que un objeto es referenciado en el sistema de archivos. Aquí, donde el almacenamiento del sistema de archivos de datos persistentes se implementa colectivamente en uno o una pluralidad de almacenamientos locales de objetos en la pluralidad de nodos, el recuento global de referencias de la vista del sistema de archivos incluye el número de referencias a través de la pluralidad de nodos en red al objeto respectivo. En una realización, el sistema de archivos puede ser implementado como un dispositivo físico conocido como un archivador, que incluye tanto el sistema de archivos como un sistema de almacenamiento, y que usa protocolos de transporte de archivos, tal como NAS. En otra realización, además o en lugar de un sistema de archivos, un servidor de almacenamiento de bloques usa los almacenamientos locales de objetos colectivamente en uno o una pluralidad de nodos, para almacenar objetos de datos persistentes. El servidor de bloques es un dispositivo físico que aparece como discos virtuales almacenados en una pluralidad de discos físicos de almacenamiento y usa protocolos de transporte de bloques, tal como iSCSI o FCoE.

Desacoplando el recuento local de referencias de instancias de objetos, el sistema permite colocar objetos (instancias) en alguno o varios nodos locales sin tener en cuenta las propiedades del objeto. En cambio, los nodos locales mantienen, cada uno, un recuento local de referencias a nombre de objeto que incluye una historia (suma) de ajustes de recuento de referencias en el nodo local. Estos recuentos locales de referencias ya no están vinculados al número de instancias de objetos según la aplicación (por ejemplo, el sistema de archivos o servidor de bloques), y así el valor del recuento local de referencias puede ser un valor negativo (lo que no sería posible si estuviesen vinculados al número de instancias de objetos). Desacoplando los recuentos de referencia del número de instancias, los objetos pueden ser colocados en cualquier lugar en la red por razones de mejor rendimiento o fiabilidad de la red o del sistema, o cualquier otra razón, independientemente del recuento global de referencias o el recuento local de referencias. Esto contrasta con los algoritmos de la técnica anterior para colocación de objetos, tal como tabla hash distribuida y otros algoritmos similares, que limitan la colocación en base al nombre de objeto. En contraposición, según la presente invención se puede hacer tantas instancias de un objeto como se desee a través de los nodos, o tan pocas como se desee para cualquier finalidad, por ejemplo, para reducir el tráfico de red, reducir el tiempo de procesado, o para replicación o recuperación de datos.

Según otro aspecto de la invención, surge un problema desacoplando el número de instancias del recuento de referencias, es decir, ya no es posible basarse en el recuento local de referencias para determinar si un objeto puede ser borrado de forma segura. Según una realización de la invención, este problema se resuelve calculando un recuento global de referencias (GRC), incluyendo la suma de todos los recuentos locales de referencias a través de los nodos de red. Este GRC será exactamente igual a la suma de todas las referencias a un objeto a través de todos los nodos de red. Un objeto puede ser desasignado (por ejemplo, borrado) si y solamente si su GRC es cero. Sin embargo, es difícil conocer el GRC puesto que los ajustes de recuento local de referencias no están coordinados con otros nodos de la red, e idealmente, sería como calcular el GRC mientras el sistema está siendo modificado activamente (por ejemplo, se están leyendo y escribiendo objetos en los nodos locales).

Según una realización de la invención, el GRC se determina asignando la propiedad de identificadores de objeto (aquí nombres de objeto) en el espacio de nombre local, a nodos diferentes en la red. No todos los nodos de la red tienen que poseer una porción del espacio de nombre. Sin embargo, todo el espacio de nombre debe estar cubierto por todos o un subconjunto de los nodos. El nodo que posee un rango de identificadores inicia el cálculo de GRC(s) para uno o más objetos en el rango, y la desasignación de objetos resultante (por ejemplo, recogida de basura). Según una realización, se utiliza un protocolo de red basado en consistencia eventual, donde el GRC se calcula mientras el sistema todavía está siendo modificado activamente (por ejemplo, se están leyendo o escribiendo objetos y se están modificando recuentos de referencias de objetos). Como se describe con más detalle más adelante, el nodo propietario proporciona una etiqueta para el uno o más identificadores de objetos para el que se ha de determinar un GRC, los nodos locales usan esta etiqueta para etiquetar los respectivos objetos (permitiendo que continúe la actividad en los nodos), y los nodos intercambian mensajes con etiqueta, donde todos los nodos proporcionan al nodo propietario sus recuentos locales de referencias para los respectivos objetos, permitiendo que el nodo propietario calcule el GRC como una suma de los recuentos locales de referencias.

De manera similar, el nodo propietario puede iniciar y llevar a cabo una reconciliación de instancia, donde el nodo propietario recibe mensajes de todos los nodos locales con relación al número de instancias de un objeto en cada nodo respectivo. El nodo propietario puede calcular entonces un número deseado y la posición de instancias a través de los nodos de red según SLAs (acuerdos de nivel de servicio) y otra métrica, y luego ordenar a los otros nodos que lleven a cabo las adiciones o los borrados deseados de instancias de objetos.

Estos y otros aspectos de la presente invención se describirán con más detalle a continuación con respecto a las realizaciones representadas en los dibujos acompañantes. Estas realizaciones se ofrecen a modo de ejemplos, y varias modificaciones serán fácilmente evidentes a los expertos en la técnica. Así, no se ha previsto que las realizaciones descritas limiten la presente invención. Se incluye cierto material introductorio como antecedentes, proporcionando un contexto para las realizaciones.

La figura 1 es un diagrama esquemático de una red de ordenadores, aquí una pluralidad de nodos en red 12, 14, 16 en una red de iguales 10. Cada nodo incluye un servidor 11a, 11b, 11c, servidores que están acoplados conjuntamente por una red (representada como nube 17) y comunican uno con otro usando protocolos de comunicación estándar, tal como TCP/IP. El servidor 11 puede ser cualquier tipo de servidor incluyendo, aunque sin limitación, servidores basados en Intel (por ejemplo, un servidor IBM 3650).

Cada nodo 12, 14, 16 es referido como un nodo local en la visión de un usuario local 18, e incluye tanto el servidor 11 como el almacenamiento de datos 13 (representado aquí como una pluralidad de discos de almacenamiento). Estos componentes locales están acoplados conjuntamente y comunican mediante un bus de sistema 20. Cada servidor puede incluir un procesador, memoria e interfaz de red, y también puede incluir dispositivos de interfaz de usuario tal como una pantalla, teclado y/o ratón. La interfaz de red acopla cada servidor a la red de nodos.

Globalmente, la pluralidad de nodos de red se denomina una red global 10. El usuario 19 que ve esta red global tiene una visión global de la red. Como se describe más adelante según varias realizaciones de la invención, se facilita software y/o hardware en cada uno de los nodos de red para realizar las varias realizaciones del sistema y método de la invención.

La figura 2 ilustra una realización de un sistema de archivos basado en objetos para uso en la presente invención. Como se describe más específicamente en la Solicitud de Estados Unidos, en tramitación y del mismo cesionario, número de serie 12/823.922 presentada el 25 de Junio de 2010 y titulada Sistema de Archivos, cuya totalidad se incorpora aquí por referencia, se facilita un producto de programa de ordenador y/o circuitería electrónica para nombrar y almacenar archivos en uno o más dispositivos informáticos de almacenamiento. El producto incluye instrucciones ejecutables o circuitería lógica para implementar un sistema de archivos de espacio de nombres 70 utilizando un almacenamiento de objetos 72. En un ejemplo, el sistema de archivos de espacio de nombre es una capa en una pila de almacenamiento entre una capa virtual del sistema de archivos y una capa de abstracción de almacenamiento de bloques. Todos los datos de sistema de archivos incluyen objetos, incluyendo datos, metadatos, y archivos. Cada objeto tiene una huella de objeto como su nombre de objeto, siendo la huella un hash del contenido de objetos. Una colección de objetos del sistema de archivos también incluye un objeto del almacenamiento de objetos, incluyendo cada colección un mapeado de una pluralidad de objetos del sistema de archivos y teniendo su propia huella de objeto derivada del contenido de la colección, donde un cambio en uno o más objetos de la colección cambia la huella de objeto de la colección. Un objeto raíz del sistema de archivos tiene una huella de objeto derivada de todos los objetos del sistema de archivos, de tal manera que cada objeto del sistema de archivos de espacio de nombres sea accesible a través del objeto raíz. Los cambios de seguimiento en el objeto raíz proporcionan una historia de la actividad del sistema de archivos, es decir, saltos del sistema de archivos. Se facilita un objeto de mapa inode incluyendo un mapeado de números inode del sistema de archivos y huellas de objetos del sistema de archivos, que hacen el sistema de archivos compatible con POSIX.

Como se representa en la figura 2, el almacenamiento de objetos de US número de serie 12/823.922 ha sido modificado según la presente invención para funcionamiento en una pluralidad de nodos en red. La figura 2 ilustra un sistema de archivos de espacio de nombres global 70 que utiliza un almacenamiento de objetos 72 para almacenar todos los objetos de datos persistentes del sistema de archivos. Como se describe mejor más adelante, el sistema de archivos de espacio de nombres 70 utiliza ahora una pluralidad de almacenamientos locales de objetos de iguales residentes en una pluralidad de nodos en red 12, 14, 16, como se representa en la figura 1. Para esta finalidad, cada nodo local 72a incluye ahora un índice de mapeado local 73, almacenamiento físico (discos) 74, y un broker de objetos 76, que está en interfaz con la red 17, para enviar/recibir comunicaciones relativas a la asignación y el seguimiento de objetos compartidos entre los nodos locales 12, 14, 16.

La figura 3 ilustra esquemáticamente la operación de un almacenamiento de objetos distribuido según una realización de la invención, donde un sistema de archivos de espacio de nombres comparte una pluralidad de almacenamientos locales de objetos residentes en varios nodos locales de la red. En la realización ilustrada, un primer nodo local 80 se denomina Alicia y un segundo nodo local 82 se denomina Bob. Un sistema de archivos FS_O 81 reside en el nodo Alicia y referencia objetos P, Q y R. Una copia de FS_O puede ser publicada en el nodo Bob simplemente enviando la firma raíz FS_O al nodo Bob (paso 1). Dado que la firma raíz referencia un objeto que a su vez se deriva de todos los objetos en el sistema de archivos, cada objeto en el sistema de archivos de espacio de nombres es accesible a través del objeto raíz. Así, no hay que transferir los objetos referenciados (P, Q, R) propiamente dichos al nodo Bob. Esta publicación del sistema de archivos FS_O en Bob se representa como un triángulo de trazos 83, con el objeto raíz del sistema de archivos designado como un círculo sin trazos 84 en el nodo Bob, para mostrar que el objeto raíz es ahora residente en el nodo Bob.

Si FS_O en el nodo Bob necesita posteriormente el objeto P (paso 2), comprobará su índice local 85 en el nodo Bob y hallará que P no reside en el nodo Bob. Por lo tanto, el broker de objetos del nodo Bob 86 enviará un mensaje de petición 90 a los otros nodos de la red, preguntándoles si tienen el objeto P. En este ejemplo, el nodo Alicia recibe la petición y comprueba su índice local 88, que muestra que el objeto P reside en el almacenamiento local de objetos 89 de Alicia. El nodo Alicia envía un mensaje de respuesta 91 al nodo Bob con una instancia de P. E objeto P se puede almacenar ahora en el almacenamiento local de objetos del nodo Bob.

La figura 4 ilustra otro ejemplo de la compartición de objetos en el almacenamiento de objetos distribuido, y más específicamente cómo los recuentos locales de referencias LRCs, mantenidos en los nodos locales, pueden tener un valor negativo. Aquí de nuevo un sistema de archivos FS_O 81, con objetos P, Q y R, reside en el nodo Alicia 80 (paso 1). Una copia de ellos es transferida al nodo Bob 82 (paso 2), enviando el objeto raíz del sistema de archivos desde el nodo Alicia al nodo Bob. Un usuario local 92 en el nodo Bob proporciona entonces instrucciones (paso 3) para borrar un archivo, el cual es parte del sistema de archivos FS_O, y que resulta que contiene P. En un intento de implementar esta instrucción desreferenciando P (paso 4), el nodo Bob comprueba su almacenamiento local de objetos en busca de referencias al objeto P. Sin embargo, el índice de mapeado local 85 en el nodo Bob muestra un recuento de referencias de cero para el objeto P. Con el fin de implementar la instrucción de borrado del objeto P, el recuento de referencias para P en la entrada del índice local se decrementa uno, de cero (0) a menos uno (-1). Esta capacidad de mantener un valor negativo para el recuento local de referencias LRC preserva los valores de referencia en una visión global (es decir, a través de la red de nodos) sin requerir una búsqueda inmediata para reconciliación de instancias de objetos a través de la red, que generaría excesivo tráfico de red. En cambio, una reconciliación de recuento de referencias a través de los nodos, a saber un cálculo global de recuentos de referencias GRC, puede ser realizado más tarde, cuando se desee, para determinar si es apropiado desasignar todas las instancias del objeto P a través de la red (por ejemplo, para recogida de basura).

La figura 5 ilustra además un ejemplo de una determinación de recuento global de referencias. El recuento global de referencias, GRC, es la suma de todas las referencias a través de los nodos en red a un objeto referenciado por una aplicación dada. El GRC se determina sumando todos los recuentos locales de referencias, LRCs, almacenados en los varios nodos locales. Aquí asumimos que es posible, en una instancia de tiempo dada, conocer todos los LRCs, para calcular GRC. Con este supuesto, la figura 5 ilustra el índice de almacenamiento local de objetos en el nodo Alicia 80 que incluye una entrada para el objeto P con un recuento local de referencias de uno positivo (+1), el objeto Q con un recuento local de referencias de uno positivo (+1), y el objeto Z con un recuento local de referencias de tres positivo (+3). El nodo Bob 82 tiene un índice de almacenamiento local de objetos con entradas para el objeto X con un recuento local de referencias de uno positivo (+1), el objeto Y con un recuento local de referencias de uno positivo (+1), y el objeto Z con un recuento local de referencias de uno positivo (+1). El GRC para el objeto Z es entonces la suma de los recuentos locales de referencias para Z en el nodo Alicia, que es tres positivo (+3), y en el nodo Bob, que es uno positivo (+1), de tal manera que el GRC para el objeto Z es cuatro positivo (+4). Esto indica que todavía existen referencias en los nodos de red al objeto Z, y Z no deberá ser desasignado. Solamente si el GRC para el objeto Z es cero, es seguro desasignar el objeto Z.

La figura 6 ilustra el concepto de referencias independientes a un objeto en un sistema de archivos, y cómo un sistema de archivos 100 que utiliza un almacenamiento de objetos 102 mantiene una visión lógica del sistema de archivos de los objetos (es decir, el conocimiento del contenido del objeto) mientras que el almacenamiento de objetos 102 mantiene una visión física de las posiciones de objeto (donde están almacenados los objetos). En la figura 6 se ilustran dos referencias independientes en una estructura de árbol; en la parte superior o raíz del árbol un signo de diagonal (/) indica un sistema de archivos concreto. Debajo del nodo superior del árbol hay dos bifurcaciones a cada uno del archivo FOO y el archivo BAR. Debajo del archivo FOO hay bifurcaciones a objetos A y B. Debajo del objeto BAR hay bifurcaciones a los objetos B y Q. En este ejemplo hay dos referencias independientes al objeto B, una a través del archivo FOO y la otra a través del archivo BAR. El almacenamiento de objetos incluye así un recuento de referencias de dos (2) para el objeto B (B_2). Los objetos restantes A, Q, FOO, BAR y / tienen un recuento de referencias de uno.

La figura 7 ilustra además cómo el desacople de referencias de objetos e instancias de objetos según la presente invención se puede permitir mientras todavía se permite una determinación de un recuento global de referencias a través de la red de nodos. En la presente invención, un recuento local de referencias solo no determina si se puede borrar una instancia. En cambio, como se ha descrito previamente, solamente la visión global obtenida sumando todos los recuentos locales de referencias es la que determina si es seguro desasignar todas las instancias de un objeto (porque ya no hay referencias en el sistema de archivos a dicho objeto). Por ejemplo, la figura 7 ilustra nodos de red Alicia 80, Bob 82 y Eva 94. Un archivo FOO incluye un objeto P que se almacena en el almacenamiento local de objetos 103 en el nodo Alicia, y el índice en el nodo Alicia tiene un recuento local de referencias de uno positivo (+1) para el objeto P. El nodo Bob 82 no tiene una copia del objeto P almacenada en su almacenamiento local de objetos 104, y el recuento local de referencias para objeto P es uno negativo (-1). El nodo Eva 94 no tiene instancia del objeto P en su almacenamiento local de objetos 105, y el recuento local de referencias para el objeto P es uno positivo (+1). Así, en el nodo Bob y el nodo Eva los recuentos de referencia para el objeto P no tienen relación a si hay una instancia almacenada en el nodo Bob o el nodo Eva respectivo. Sin embargo, cuando se calcula el recuento global de referencias GRC para el objeto P, incluyendo la suma de los tres recuentos locales de referencias para el objeto P, la suma calculada es uno positivo (+1), lo que indica que todavía hay una referencia al objeto P y P no deberá ser desasignado.

La figura 8 es un gráfico de flujo que ilustra un ejemplo del mantenimiento de recuentos locales de referencias a través de la pluralidad de nodos. Como se ha descrito previamente, para cualquier objeto O, el recuento global de referencias de O para el sistema de archivos P corresponde exactamente a la suma de todas las referencias al objeto O a través de los nodos en red, suponiendo que todos los recuentos de referencia pendientes se hayan propagado completamente por todo el sistema de archivos A. En la figura 8, el nodo local Alicia y el nodo local Bob

tienen su propio almacenamiento local de objetos y broker de objetos. El sistema de archivos P en el nodo Alicia incluye objetos A y B. El sistema de archivos P tiene dos referencias al objeto A en el nodo Alicia, y el recuento local de referencias de objeto un en el nodo Alicia es dos positivo (A_2). El recuento local de referencias del objeto B es uno positivo (B_1) en el nodo Alicia. El sistema de archivos Q en el nodo Bob tiene el objeto referenciado A una vez, y así el recuento local de referencias para el objeto A en el nodo B es uno positivo (A_1).

El broker de objetos en el nodo Bob desea ahora replicar el objeto A al nodo Alicia. El broker de objetos en el nodo Alicia recibe una petición enviada por el nodo Bob e incrementa el recuento local de referencias para el objeto A en el nodo Alicia a tres positivo (A_3).

Más tarde, el sistema de archivos P en el nodo Alicia desea mover una copia del sistema de archivos P al nodo Bob. Cuando el sistema de archivos P en el nodo Bob lee el objeto B, no halla el objeto B en el almacenamiento local de objetos en el nodo Bob. El broker de objetos en el nodo Bob envía entonces una petición de lectura para el objeto B al nodo Alicia. El broker de objetos en el nodo Alicia lee B del almacenamiento local de objetos en el nodo Alicia, envía el objeto B al nodo Bob, y el objeto B es suministrado entonces al sistema de archivos P en el nodo Bob.

Más tarde, el sistema de archivos Q en el nodo Bob desreferencia su primera referencia a A, disminuyendo el recuento local de referencias para el objeto A a cero (A_0). Posteriormente, el sistema de archivos Q en el nodo Bob emite una segunda desreferencia del objeto A, reduciendo el recuento local de referencias para el objeto A en el nodo Bob a menos uno (A_1). En la técnica anterior no se permiten recuentos de referencia negativos y el nodo Bob tendría que comunicar con los otros nodos y hallar otro nodo para aceptar el recuento de referencias negativo del objeto A. Sin embargo, según la presente invención, no hay necesidad de realizar una reconciliación inmediata y se evita este tráfico de red. En cambio, la reconciliación se puede posponer hasta que se desee, como se describe mejor más adelante.

Según otro aspecto de la invención, a los nodos locales se les da la propiedad de un rango específico de identificadores de objetos, o claves (aquí nombres de objeto) en el espacio de claves total. El nodo propietario es responsable de iniciar una propagación de recuentos de referencia. No todos los nodos del sistema de almacenamiento tienen que poseer una porción del espacio de claves. Sin embargo, todo el espacio de claves debe estar cubierto por algún conjunto de nodos. El propietario (nodo) de la clave es responsable de calcular el recuento global de referencias GRC del objeto correspondiente. En una realización, el método incluye llevar la cuenta tanto de los recuentos locales de referencias como de los recuentos locales de instancias. En otras realizaciones, se puede llevar la cuenta de uno u otro.

En un ejemplo, un proceso de llevar la cuenta puede incluir los pasos siguientes:

a. Nodo propietario (etiquetado)

El propietario crea una etiqueta única que es un identificador para indicar un cálculo de GRC en un punto de tiempo concreto;

El propietario decide el rango de claves RK para el que se ha de calcular el GRC;

El propietario envía (RK, T) a todos los nodos en el sistema de almacenamiento.

b. Nodo(s) receptor(es) (acumulación)

Para cada (RK, T) recibido, el receptor debe etiquetar todas las claves que sean locales con T;

Cualesquiera modificaciones adicionales a una clave en RK después del poner de etiqueta T, deben ser distintas;

Para cada clave en RK, el receptor envía su recuento local de referencias, y el número de instancias locales de nuevo al nodo propietario junto con la etiqueta T.

c. Nodo propietario (reconciliación)

El propietario suma los recuentos locales de referencias recibidos de los otros nodos, junto con su propio recuento local de referencias para determinar el recuento global de referencias. Si el recuento global de referencias es cero, el propietario envía entonces un mensaje de borrado con la etiqueta T a todos los otros nodos, y borra cualquier instancia local de la clave en el nodo propietario;

Los nodos receptores, y el nodo propietario, no pueden borrar realmente el objeto si había actividad después de la puesta de T en el objeto;

El propietario también revisa el número de instancias del objeto en varios nodos de la red, y determina si varias instancias se pueden añadir o borrar para mejorar el rendimiento o análogos.

d. Nodo(s) receptor(es) (acción)

El receptor lleva a cabo la instrucción de borrar, y envía de nuevo un reconocimiento al propietario con T;

El receptor lleva a cabo las instrucciones relativas a la creación de instancia si así se indica, y envía un reconocimiento de nuevo al propietario con T.

Este método de reconciliación proporciona “consistencia eventual”. El orden en el que los mensajes entre los nodos son recibidos o procesados no es importante, a condición de que todos vuelvan al nodo propietario, eventualmente. La etiqueta actúa como una barrera que permite al nodo receptor continuar la actividad en el RK puesto, mientras los nodos participan en el proceso de reconciliación. En una realización, la etiqueta es una GUID (ID globalmente única).

Un ejemplo de este proceso de etiquetado y llevar la cuenta se ilustra en la figura 9. Aquí, el nodo Alicia posee una porción del espacio de claves y decide trabajar en una porción secundaria de dicho rango. Alicia genera una etiqueta única T y envía una petición de acumulación relativa tanto a los recuentos locales de referencias como a los recuentos locales de instancias para el rango KR deseado, junto con la etiqueta T.

Los nodos Bob y Eva reciben eventualmente la petición de Alicia del rango de clave KR, etiqueta T. La petición puede llegar a los nodos Bob y Eva en cualquier orden. Cada nodo receptor debe etiquetar el rango de clave. Si, por ejemplo, un objeto X en el rango de claves KR en el nodo Bob ha de ser modificado, Bob debe crear una barrera, por ejemplo, ahora hay dos objetos X: una etiqueta de objeto X y el otro objeto X_{other}. Cada nodo receptor responde entonces al nodo Alicia por cada clave K en el rango R con lo siguiente: el recuento local de referencias (refs), y el número de instancias (señalizadores a instancia física, Iones) en el nodo local.

El nodo propietario Alicia tiene ahora suficiente información para calcular el recuento global de referencias y el recuento global de instancias para cada clave K en el rango R bajo la etiqueta T. Alicia acumula las respuestas de todos los nodos. Si el recuento global de referencias GRC para un objeto es cero, Alicia puede enviar instrucciones a todos los nodos para desasignar el objeto. El receptor responde con un mensaje de reconocimiento. Alicia puede aplicar varios métodos (por ejemplo, acuerdos de nivel de servicio) para determinar si se deberá crear o borrar instancias en uno o varios de los otros nodos. Solamente si el recuento global de referencias es cero, se pueden borrar todas las instancias.

Otras realizaciones

La presente invención se puede usar para implementar un sistema de archivos y/o un índice para un sistema de archivos, como se describe en US número de serie 12/823.922, en tramitación y del mismo cesionario, titulada Sistema de Archivos, de A. J. Beaverson y P. Bowden, y US número de serie 12/823.452 titulada Indexación Escalable de P. Bowden y A. J. Beaverson, ambas presentadas el 25 de Junio de 2010, y ambas reivindican prioridad por la Provisional de Estados Unidos número 61/269.633 presentada el 26 de Junio de 2009. Aquí se reivindica prioridad por cada una de estas solicitudes y las descripciones completas de cada una se incorporan por ello por referencia en su totalidad.

Las realizaciones de la invención se pueden implementar en circuitería electrónica digital o en hardware de ordenador, microprogramas, software, o en combinaciones de los mismos. Las realizaciones de la invención se pueden implementar como un producto de programa de ordenador, es decir, un programa de ordenador realizado de forma tangible en un medio legible por ordenador, por ejemplo, en un dispositivo de almacenamiento legible por máquina, para ejecución por, o para controlar la operación de, aparatos de procesado de datos, por ejemplo, un procesador programable, un ordenador, o múltiples ordenadores. Un programa de ordenador puede estar escrito en cualquier forma de lenguaje de programación, incluyendo lenguajes compilados o interpretados, y se puede desplegar en cualquier forma, incluyendo como un programa autónomo o como un módulo, componente, subrutina u otra unidad adecuada para uso en un entorno de cálculo. Un programa de ordenador puede ser desplegado para ser ejecutado en un ordenador o en múltiples ordenadores en un lugar o distribuidos a través de múltiples lugares e interconectados por una red de comunicaciones.

Los pasos de método de las realizaciones de la invención pueden ser realizados por uno o más procesadores programables que ejecuten un programa de ordenador para realizar funciones de la invención operando en datos introducidos y generando una salida. Los pasos de método también pueden ser realizados por, y el aparato de la invención puede ser implementado como, circuitería lógica de finalidad especial, por ejemplo, una FPGA (puerta serie de campo programable) o un ASIC (circuito integrado específico de aplicación).

Los procesadores adecuados para la ejecución de un programa de ordenador incluyen, a modo de ejemplo, microprocesadores tanto generales como de finalidad especial, y cualquiera o más procesadores de cualquier tipo de ordenador digital. Por lo general, un procesador recibirá instrucciones y datos de una memoria de lectura solamente o una memoria de acceso aleatorio o ambas. Los elementos esenciales de un ordenador son un

procesador para ejecutar instrucciones y uno o más dispositivos de memoria para almacenar instrucciones y datos. Por lo general, un ordenador también incluirá, o estará acoplado operativamente para recibir datos o transferir datos, o ambos, a uno o más dispositivos de almacenamiento masivo para almacenar datos, por ejemplo, discos magnéticos, magnetoópticos u ópticos. Los soportes de información adecuados para realizar instrucciones de programa de ordenador y datos incluyen todas las formas de memoria no volátil, incluyendo a modo de ejemplo dispositivos de memoria de semiconductores, por ejemplo, EPROM, EEPROM, y dispositivos de memoria flash; discos magnéticos, por ejemplo, discos duros internos o discos extraíbles; discos magnetoópticos; y discos CD ROM y DVD-ROM. El procesador y la memoria pueden ser complementados por circuitería lógica de finalidad especial o incorporados en ella.

Se ha de entender que la descripción anterior tiene la finalidad de ilustrar y no de limitar el alcance de la invención.

REIVINDICACIONES

1. Un medio legible por ordenador conteniendo instrucciones de programa ejecutables para realizar un método incluyendo:

en una pluralidad de nodos en red donde cada nodo tiene su propio almacenamiento local de objetos, almacenando objetos los almacenamientos de objetos y compartiendo uno o más objetos, teniendo los objetos nombres globalmente únicos a través de los nodos en red y donde los nombres de objeto no cambian en base a donde los objetos están almacenados en los nodos;

mantener, en cada nodo, un recuento local de referencias LRC a nombres de objeto, independientemente de cualquier instancia de objeto almacenada en el almacenamiento local de objetos, manteniéndose el LRC como un entero con signo, donde se hacen ajustes en el LRC para cada nueva referencia y desreferencia locales y una desreferencia de un nombre de objeto puede generar un valor LRC negativo;

donde la propiedad de nombres de objeto es asignada a diferentes nodos, asignándose un rango de nombres de objeto a uno de los nodos de red y la propiedad asignada a nodo inicia un paso de determinación de recuento global de referencias GRC para todo o para un subconjunto del rango, incluyendo el paso de determinación de GRC:

proporcionando el nodo propietario una etiqueta única para un objeto para el que se ha de determinar un GRC;

intercambiando los nodos mensajes con la etiqueta única, por lo que los otros nodos proporcionan al nodo propietario su LRC para el objeto;

calculando el nodo propietario el GRC como una suma de los LRCs recibidos de los otros nodos y el LRC del nodo propietario para el objeto; y

si el GRC resultante es cero, el nodo propietario envía un mensaje de borrado con la etiqueta única a los otros nodos.

2. El medio de la reivindicación 1, donde el método incluye:

el propietario y otros nodos borran el objeto cuando el recuento GRC es cero.

3. El medio de la reivindicación 2, donde el paso de determinación incluye:

los nodos operan como una red de nodos de iguales.

4. El medio de la reivindicación 3, donde el GRC se calcula mientras se están leyendo o escribiendo objetos y recuentos de referencias de objetos locales están siendo modificados.

5. El medio de la reivindicación 1, donde los almacenamientos locales de objetos incluyen colectivamente un espacio de nombre de los nombres de objeto globalmente únicos.

6. El medio de la reivindicación 1, donde el método incluye:

el almacenamiento local de objetos mantiene un índice de mapeado local del nombre de objeto, LRC y un indicador a cualquier posición física de objeto donde se almacena el objeto en el almacenamiento local de objetos.

7. El medio de la reivindicación 1, donde el método incluye:

cada objeto tiene una huella de objeto derivada del contenido de objetos como su nombre de objeto.

8. El medio de la reivindicación 7, donde el método incluye:

la huella incluye un hash del contenido de objetos.

9. El medio de la reivindicación 1, donde un sistema de archivos usa los almacenamientos locales de objetos colectivamente como un método para almacenar todos los datos persistentes del sistema de archivos.

10. El medio de la reivindicación 9, donde el método incluye:

todos los datos de sistema de archivos, metadatos y archivos incluyen objetos del (de los) almacenamiento(s) de objetos, teniendo cada objeto una huella de objeto como su nombre de objeto;

las colecciones de objetos del sistema de archivos también incluyen objetos del (de los) almacenamiento(s) de

objetos, incluyendo cada colección un mapeado de una pluralidad de los objetos del sistema de archivos y teniendo su propia huella de objeto derivada del contenido de la colección, donde un cambio en uno o más objetos de la colección cambia la huella de objeto de la colección; y

teniendo un objeto raíz del sistema de archivos una huella de objeto raíz derivada de todos los objetos del sistema de archivos, de tal manera que cada objeto en el sistema de archivos de espacio de nombres sea accesible a través del objeto raíz.

11. El medio de la reivindicación 1, donde el método incluye:

seleccionar, en base al rendimiento o la fiabilidad de la red o del sistema, uno o varios nodos como posición(es) para almacenar una o más instancias de un objeto independientemente del nombre de objeto.

12. El medio de la reivindicación 1, donde el método incluye:

la compartición de los objetos almacenados incluye comunicar entre nodos con respecto a nombres de objeto, LRCs y posiciones de objetos almacenados.

13. El medio de la reivindicación 1, donde el método incluye:

cuando una aplicación desreferencia un nombre de objeto y una copia del objeto no está almacenada en un nodo local, el nodo local genera un LRC de uno negativo.

14. El medio de la reivindicación 1, donde un sistema de almacenamiento usa los almacenamientos locales de objetos colectivamente como un método para almacenar todos los datos persistentes del sistema de almacenamiento.

15. Un aparato incluyendo:

una pluralidad de nodos en red, teniendo cada nodo un almacenamiento local de objetos para almacenar objetos en el nodo respectivo y teniendo los objetos nombres globalmente únicos a través de los nodos en red, donde los nombres de objeto no cambian en base a donde los objetos están almacenados en los nodos; compartiendo los almacenamientos locales de objetos uno o varios objetos;

una aplicación que usa los almacenamientos locales de objetos para almacenar datos persistentes como objetos,

requiriendo la aplicación un recuento de referencias para cada objeto almacenado en los nodos en red incluyendo un número de recorridos independientes a través de la pluralidad de nodos al objeto respectivo;

medios para mantener un recuento local de referencias LRC a nombres de objeto en cada nodo, independientemente de cualquier instancia de objeto almacenada en el almacenamiento local de objetos, manteniéndose el LRC como un entero con signo;

donde se hacen ajustes al LRC para cada nueva referencia y desreferencia locales, donde una desreferencia de un nombre de objeto puede generar un valor LRC negativo; y

medios para determinar un recuento global de referencias GRC para un objeto, donde la propiedad de nombres de objeto es asignada a nodos diferentes, asignándose un rango de nombres de objeto a uno de los nodos de red, y la propiedad asignada a nodo está configurada para hacer que los medios de determinación inicien un paso de determinación de GRC para todo o un subconjunto del rango, incluyendo los medios de determinación GRC y los nodos configurados para implementar pasos de determinación de GRC:

proporcionando el nodo propietario una etiqueta única para un objeto para el que se ha de determinar un GRC;

intercambiando los nodos mensajes con la etiqueta única, por lo que los otros nodos proporcionan al nodo propietario su LRC para el objeto;

calculando el nodo propietario el GRC como una suma de los LRCs recibidos de los otros nodos y el LRC del nodo propietario para el objeto; y

si el GRC resultante es cero, el nodo propietario envía un mensaje de borrado con la etiqueta única a los otros nodos.

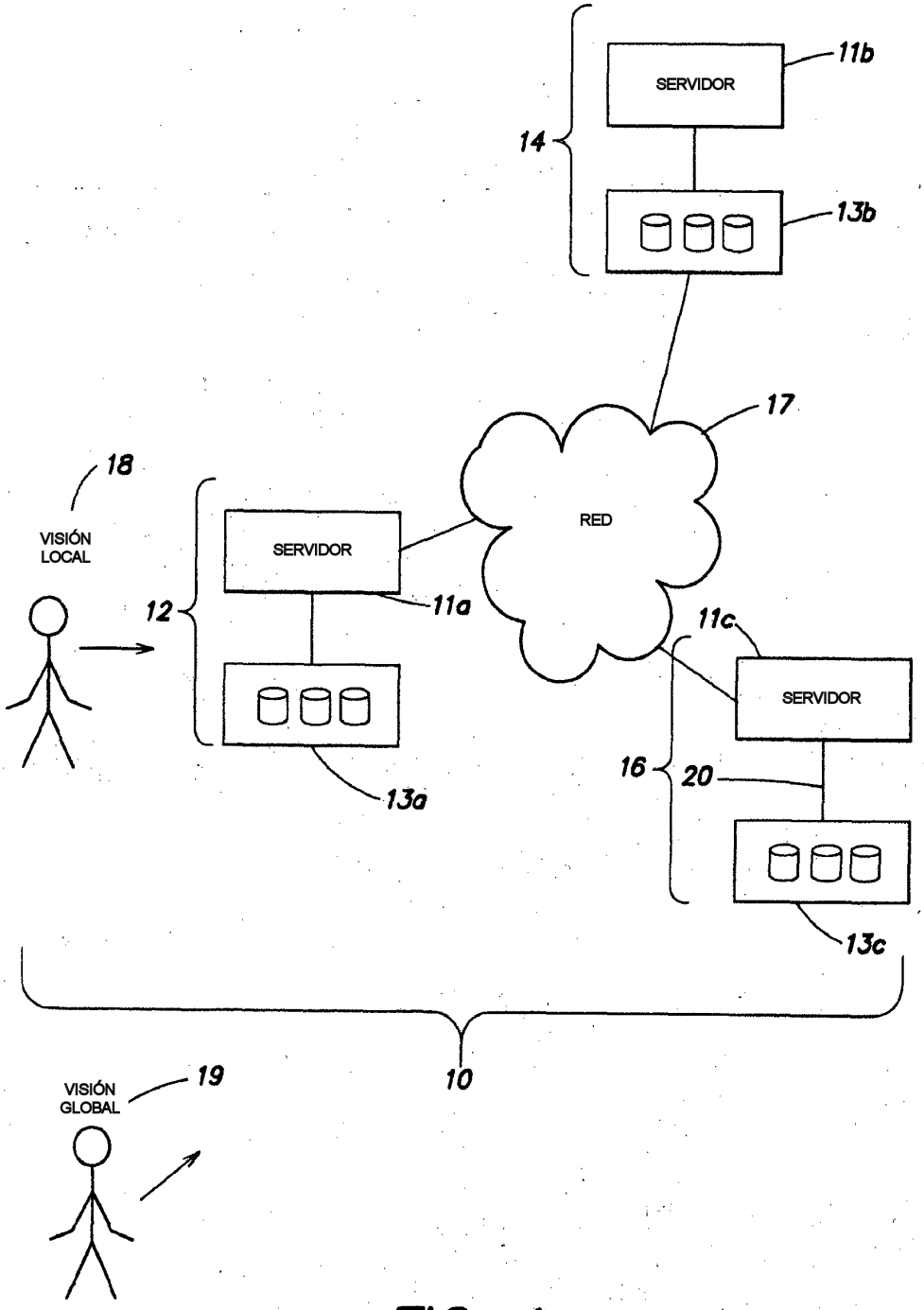
SERVIDOR
11b
14
13b
17
RED
18
VISIÓN LOCAL
SERVIDOR
12
11a
11c
SERVIDOR
16
20
13a
13c
10
VISIÓN GLOBAL
19

FIG. 1

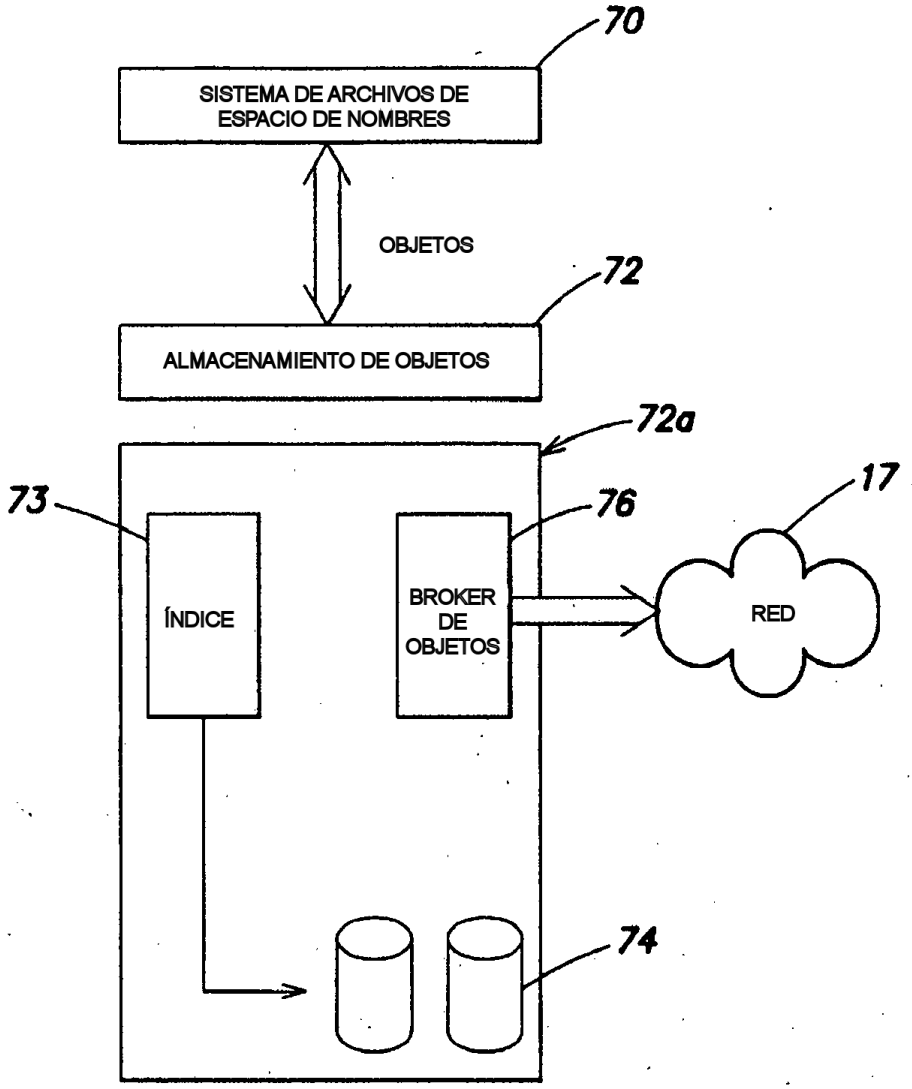
70
SISTEMA DE ARCHIVOS DE ESPACIO DE NOMBRES
OBJETOS
72
ALMACENAMIENTO DE OBJETOS
72a
73
ÍNDICE
76
BROKER DE OBJETOS
17
RED
74

FIG. 2

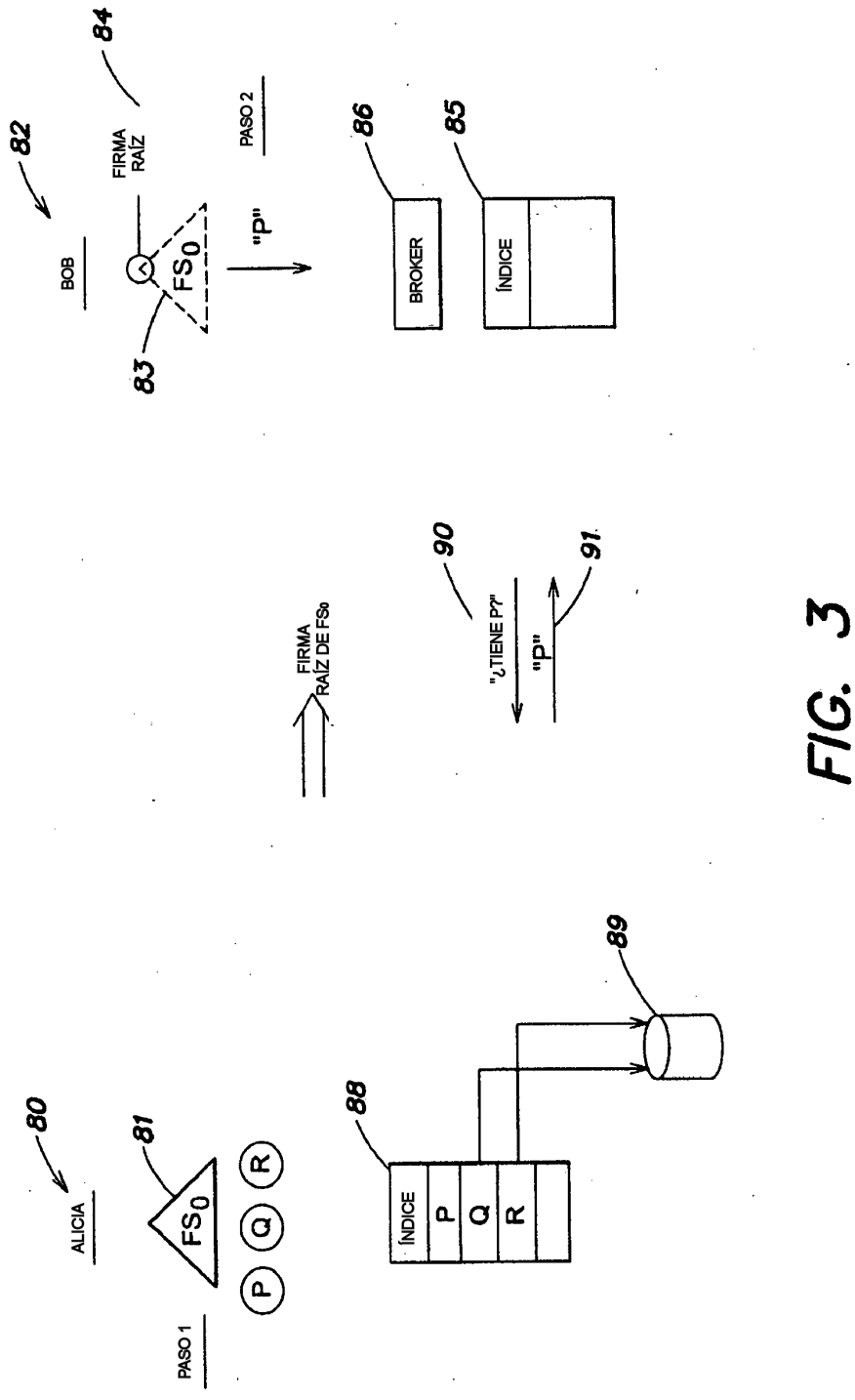
82
BOB
84
FIRMA RAÍZ
FS_0
83
PASO 2
"P"
86
BROKER
85
ÍNDICE
90
"¿TIENE P?"
91
"P"
FIRMA RAÍZ DE FS_0
80
ALICIA
81
FS_0
P
Q
R
PASO 1
88
ÍNDICE
P
Q
R
89

FIG. 3

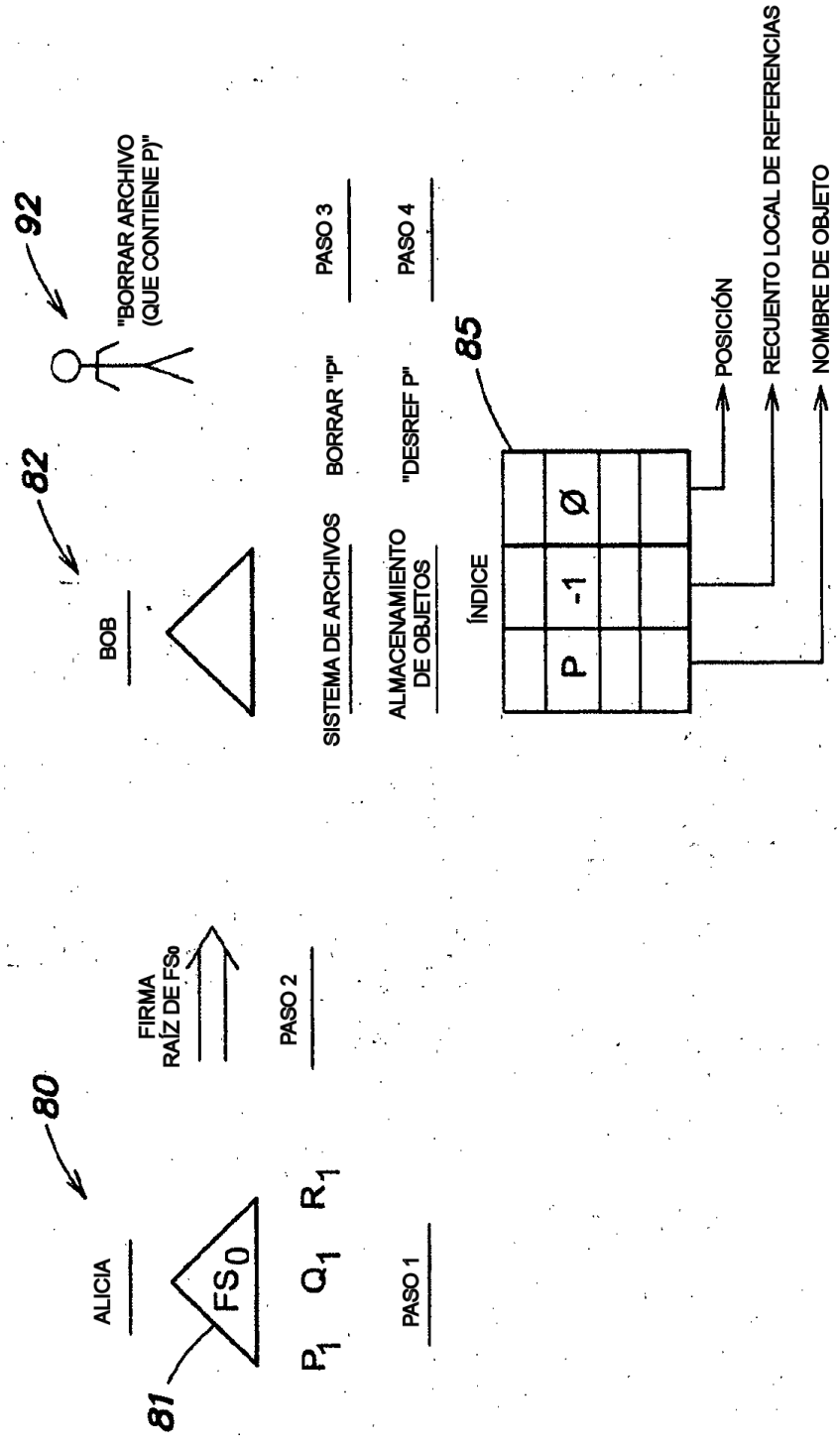
80
ALICIA
81
FS_0
P_1 Q_1 R_1
PASO 1
FIRMA
RAÍZ DE FS_0
PASO 2
82
BOB
SISTEMA DE ARCHIVOS
BORRAR "P"
PASO 3
ALMACENAMIENTO
DE OBJETOS
"DESREF P"
PASO 4
92
"BORRAR ARCHIVO
(QUE CONTIENE P)"
85
ÍNDICE
P
-1
Ø
POSICIÓN
RECUENTO LOCAL DE REFERENCIAS
NOMBRE DE OBJETO

FIG. 4

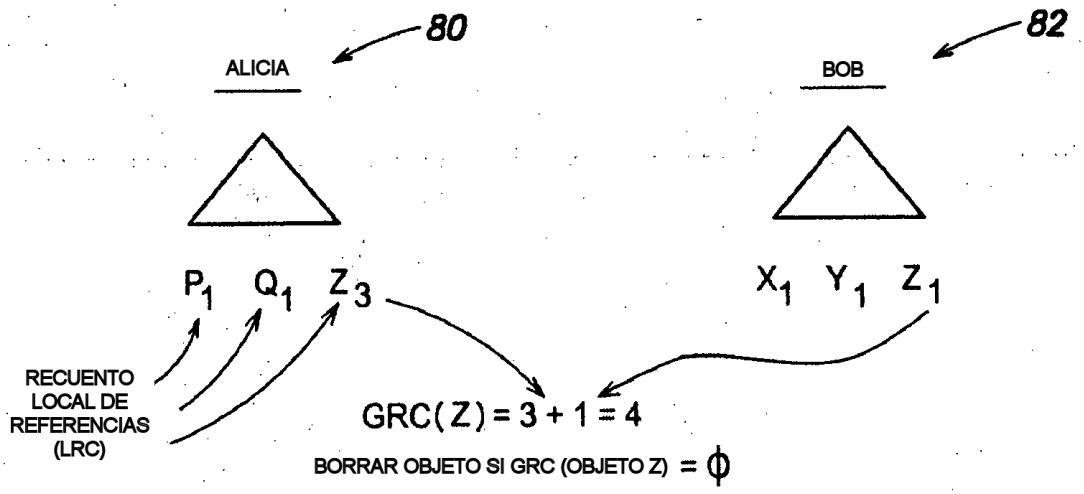
80
ALICIA
82
BOB
P1 Q1 Z3
X1 Y1 Z1
RECUENTO LOCAL DE REFERENCIAS (LRC)
GRC(Z) = 3 + 1 = 4
BORRAR OBJETO SI GRC (OBJETO Z) = ϕ

FIG. 5

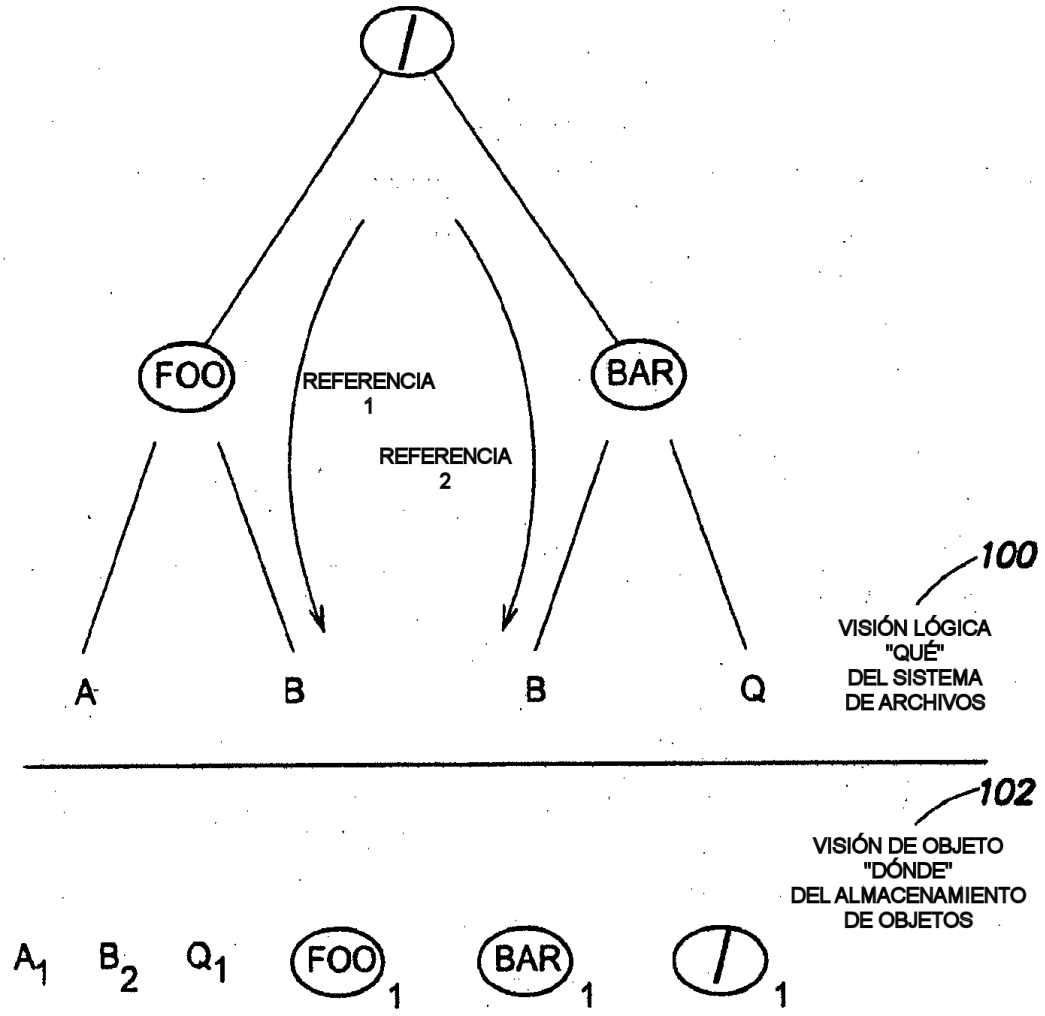
/
FOO
BAR
REFERENCIA 1
REFERENCIA 2
A
B
B
Q
100
VISIÓN LÓGICA "QUÉ" DEL SISTEMA DE ARCHIVOS
102
VISIÓN DE OBJETO "DÓNDE" DEL ALMACENAMIENTO DE OBJETOS
A1
B2
Q1
FOO1
BAR1
/1

FIG. 6

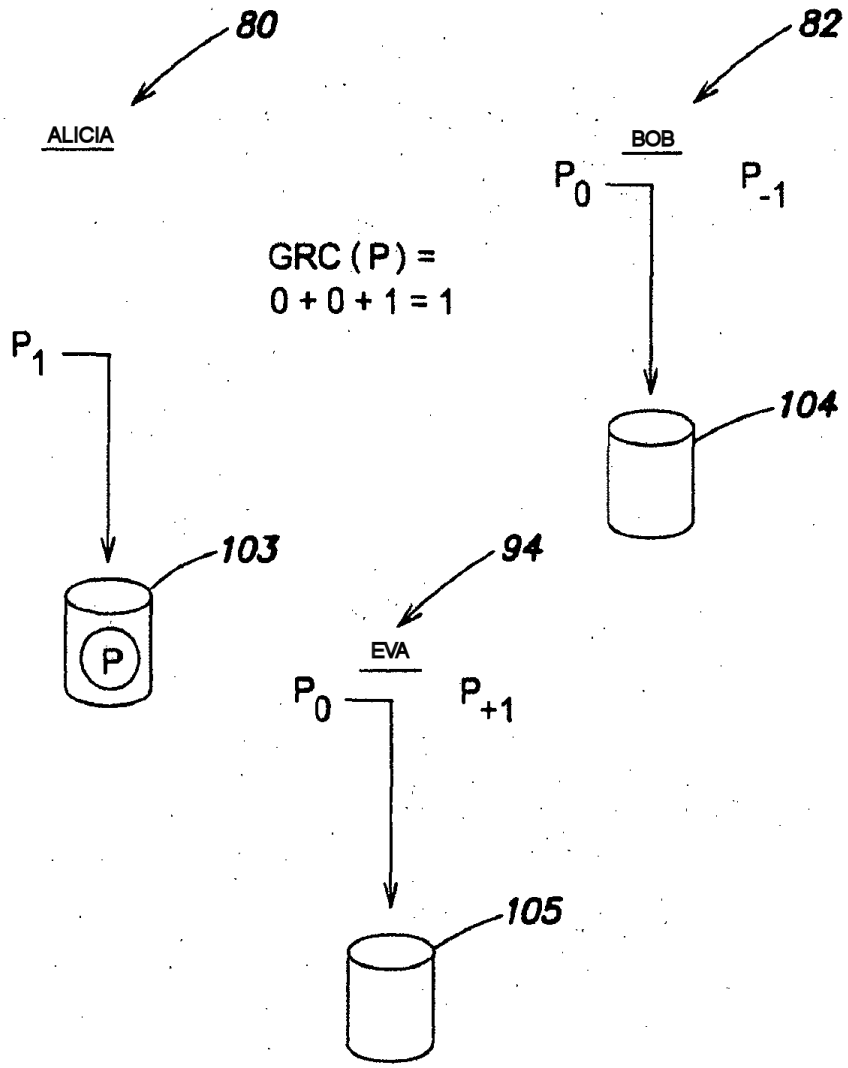

FIG. 7

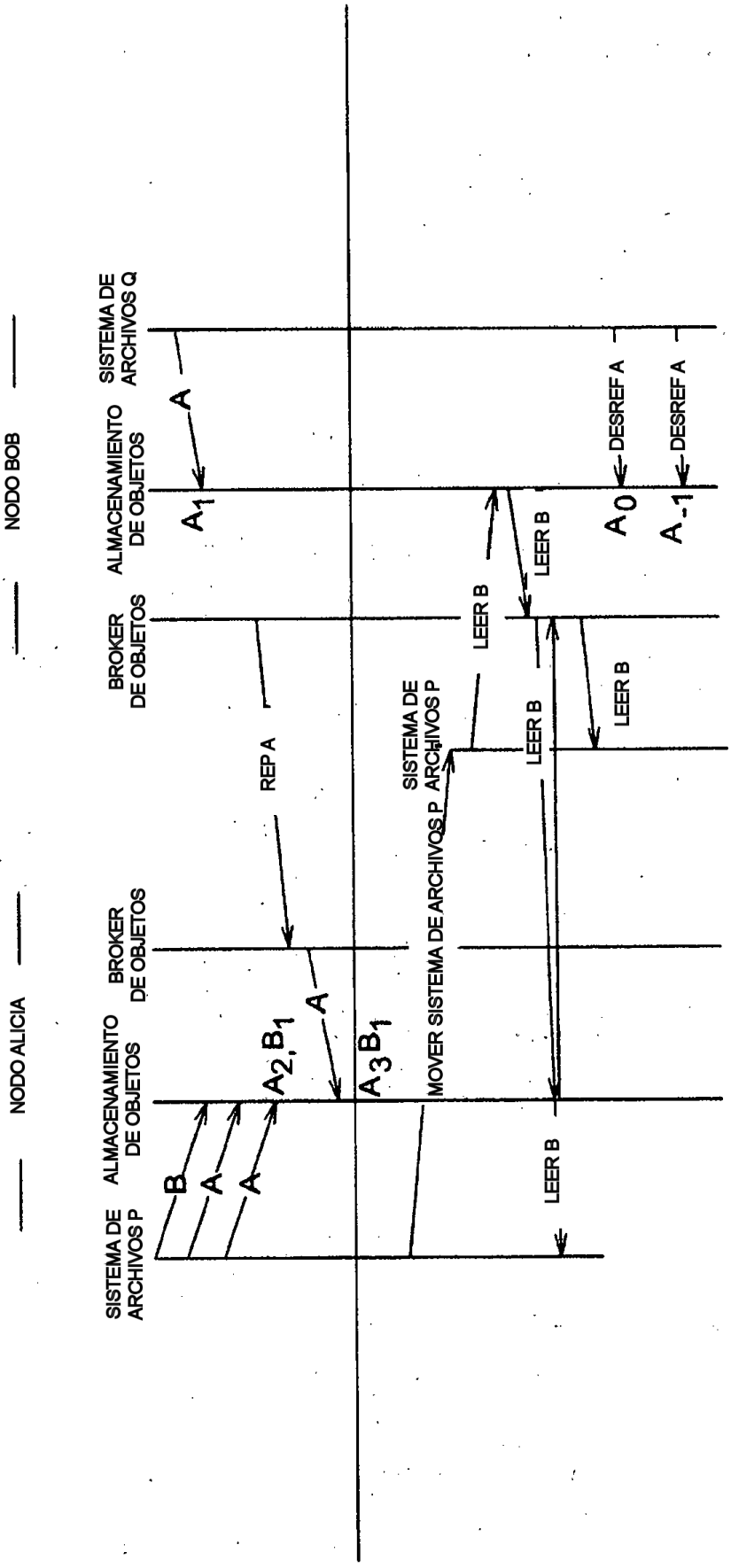
NODO ALICIA
NODO BOB
SISTEMA DE ARCHIVOS P
ALMACENAMIENTO DE OBJETOS
BROKER DE OBJETOS
BROKER DE OBJETOS
ALMACENAMIENTO DE OBJETOS
SISTEMA DE ARCHIVOS Q
B
A
A
A2, B1
A
A3 B1
REP A
MOVER SISTEMA DE ARCHIVOS P
SISTEMA DE ARCHIVOS P
LEER B
LEER B
LEER B
LEER B
LEER B
A1
A0
A-1
DESREF A
DESREF A

FIG. 8

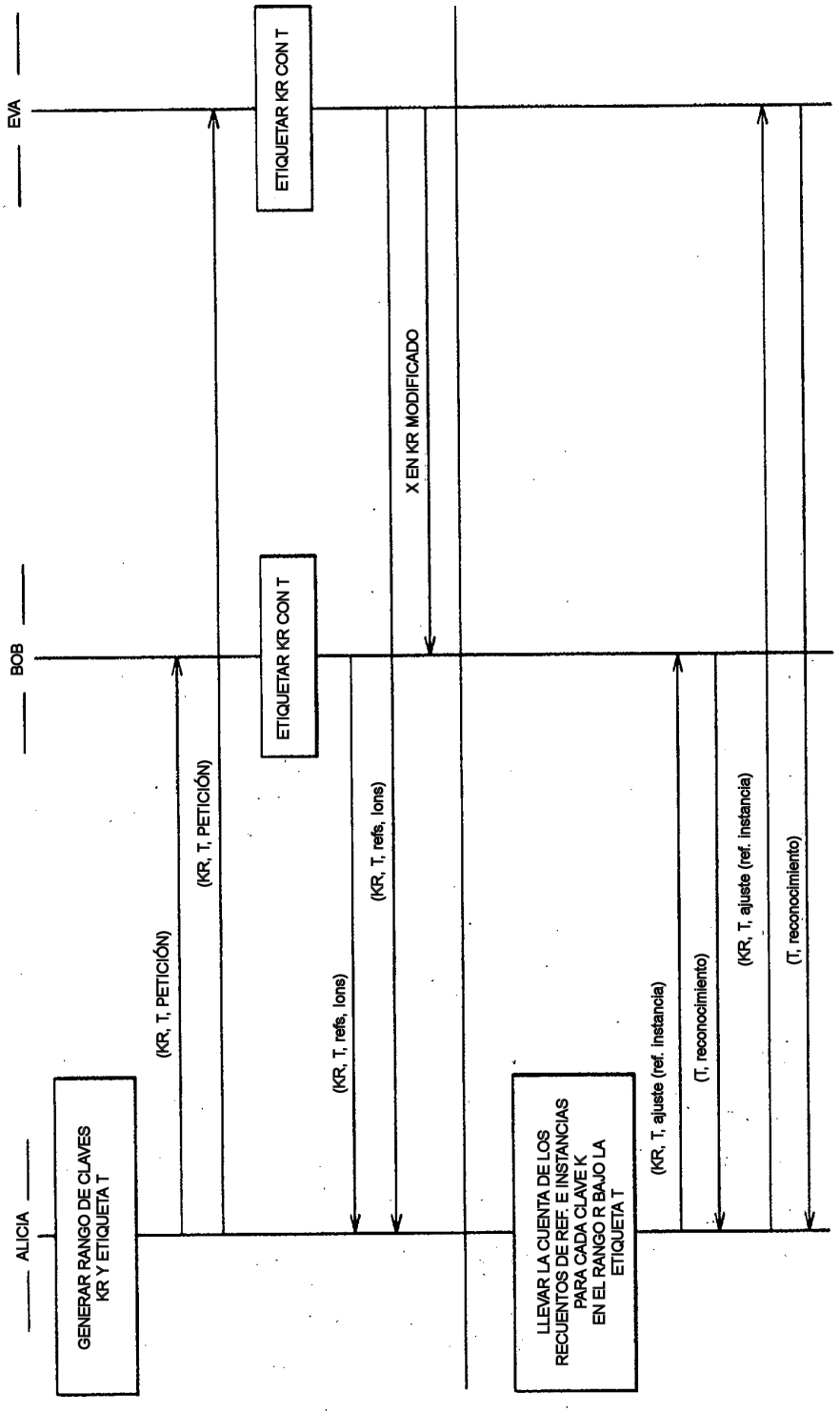
ALICIA
BOB
EVA
GENERAR RANGO DE CLAVES KR Y ETIQUETA T
(KR, T, PETICIÓN)
(KR, T, PETICIÓN)
ETIQUETAR KR CON T
ETIQUETAR KR CON T
X EN KR MODIFICADO
(KR, T, refs, Ions)
(KR, T, refs, Ions)
LLEVAR LA CUENTA DE LOS RECUENTOS DE REF. E INSTANCIAS PARA CADA CLAVE K EN EL RANGO R BAJO LA ETIQUETA T
(KR, T, ajuste (ref. instancia)
(T, reconocimiento)
(KR, T, ajuste (ref. instancia)
(T, reconocimiento)

FIG. 9