(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 720 482**

(51) Int. CI.:

G06F 15/16 (2006.01)
G06F 9/50 (2006.01)
G06F 3/06 (2006.01)
G06F 17/30 (2006.01)
H04L 29/08 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **13.12.2012 PCT/US2012/069595**

(87) Fecha y número de publicación internacional: **20.06.2013 WO13090640**

(96) Fecha de presentación y número de la solicitud europea: **13.12.2012 E 12856978 (7)**

(97) Fecha y número de publicación de la concesión europea: **23.01.2019 EP 2791813**

(54) Título: **Equilibrio de carga en sistemas de almacenamiento en grupos**

(30) Prioridad:

13.12.2011 US 201113324497

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
22.07.2019

(73) Titular/es:

MICROSOFT TECHNOLOGY LICENSING, LLC (100.0%)
One Microsoft Way
Redmond, WA 98052, US

(72) Inventor/es:

JAIN, NAVENDU y
YOU, GAE-WON

(74) Agente/Representante:

ELZABURU, S.L.P

DESCRIPCIÓN

Equilibrio de carga en sistemas de almacenamiento en grupos

Antecedentes

Los centros de datos empresariales y en la nube pueden incluir decenas de miles de servidores, que proporcionan petabytes de almacenamiento a un gran número de usuarios y aplicaciones. Además, a medida que los servicios en la nube continúan creciendo rápidamente, se están construyendo sistemas de almacenamiento en la nube a gran escala para servir datos a miles de millones de usuarios en todo el mundo. El objetivo principal de estos sistemas es proporcionar sistemas de rendimiento escalable y de almacenamiento de datos de alta disponibilidad, mientras se minimizan los gastos operativos, particularmente el coste del ancho de banda de datos móviles en el sistema de almacenamiento. Sin embargo, las diversas cargas de trabajo de entrada/salida (I/O) pueden causar un desequilibrio significativo de datos en los servidores, dando como resultado puntos calientes, o servidores/unidades de disco que son sobrecargados con un gran número de solicitudes I/O. Tales puntos calientes pueden causar grandes retrasos a los usuarios finales. Como resultado, estos servicios mezclan típicamente terabytes de datos por día para equilibrar carga entre los grupos. Además, el mismo desafío ha sido planteado y tratado activamente en el contexto de la creación de bases de datos en las nubes. Sin embargo, esta solución o bien no aborda la reconfiguración dinámica de la carga o bien asume que los nodos fuente y objetivo de la migración dinámica de datos son conocidos por un sistema de gestión de bases de datos relacional de objetos.

Además, las técnicas anteriores han tenido como objetivo resolver estos desafíos individualmente. Para abordar el desequilibrio de carga, muchas técnicas realizan la colocación dinámica de objetos de datos individuales, o distribuyen objetos aleatoriamente en todo el grupo, p.ej., basándose en el resumen criptográfico. Sin embargo, la redistribución adaptativa de objetos implica el conocimiento de patrones de carga para miles de millones de objetos. La optimización de los costes de configuración para estos patrones requiere solucionadores fuera de línea, por ejemplo solucionadores de mochila o basados en programación lineal, para tomar decisiones de migración. Sin embargo, como tal optimización es inherentemente costosa, estos enfoques son adecuados a pequeña escala y menos efectivos cuando los sistemas crecen a gran escala. Mientras tanto, los enfoques de la efectividad del negocio para conseguir la escalabilidad, por ejemplo, utilizando un recocido simulado o un algoritmo evolutivo, sufren de altos costes de reconfiguración. El documento US 2005/0080810 A1 (Matsuura) describe una unidad de observación de carga de un servidor de gestión de sistema de archivo distribuido que observa el estado de la carga de cada disco. Cuando la carga de un disco específico sobrepasa un nivel predeterminado, una unidad de control de datos mueve datos almacenados en ese disco a un disco arbitrario. Otras realizaciones describen la migración de datos de datos que tienen una carga que sobrepasa un cierto nivel a un disco de destino elegido de acuerdo con un rendimiento, su capacidad de repuesto, o un nivel de servicio de un cliente que requiere que el disco de destino esté dentro de una distancia de red máxima permitida del cliente.

Compendio

A continuación se presenta un compendio simplificado de la innovación con el fin de proporcionar una comprensión básica de algunos aspectos descritos en la presente memoria. Este compendio no es un resumen extensivo del tema reivindicado. No pretende identificar ningún elemento clave ni críticos de la materia reivindicada ni delinear el alcance de la innovación objeto. Su único propósito es presentar algunos conceptos de la materia reivindicada en una forma simplificada como un preludio a la descripción más detallada que es presentada más adelante.

Una realización proporciona un método para equilibrar carga en un sistema de almacenamiento en grupos. El método incluye identificar un nodo fuente dentro del sistema de almacenamiento en grupos desde el cual mover un número de objetos de datos, en donde el nodo fuente incluye un nodo con una carga total que sobrepasa un valor de umbral. El método también incluye seleccionar los objetos de datos para moverlos desde el nodo fuente, en donde los objetos de datos son elegidos de tal manera que la carga total del nodo fuente ya no sobrepasa el valor de umbral. El método incluye además determinar un nodo objetivo dentro del sistema de almacenamiento en grupos basándose en una proximidad al nodo fuente y a la carga total del nodo objetivo y mover los objetos de datos desde el nodo fuente al nodo objetivo.

Otra realización proporciona un sistema para equilibrar carga en un sistema de almacenamiento en grupos. El sistema incluye un nodo fuente, en donde el nodo fuente es un nodo dentro del sistema de almacenamiento en grupos con una primera carga total que sobrepasa un valor de umbral. El sistema también incluye un nodo objetivo, en donde el nodo objetivo es un nodo dentro del sistema de almacenamiento en grupos con una segunda carga total que no sobrepasa el valor de umbral. El nodo objetivo es elegido basándose en una proximidad al nodo fuente. El sistema incluye además un sistema de equilibrio de carga configurado para seleccionar un número de objetos de datos para moverlos desde el nodo fuente al nodo objetivo, en donde los objetos de datos son elegidos de tal manera que la primera carga total del nodo fuente ya no sobrepasa el valor de umbral, y mover los objetos de datos desde el nodo fuente al nodo objetivo.

Otra realización proporciona uno o más medios legibles por ordenador, no transitorios para almacenar instrucciones legibles por ordenador. Las instrucciones legibles por ordenador proporcionan un sistema de equilibro de carga cuando son ejecutadas por uno o más dispositivos de procesamiento. Las instrucciones legibles por ordenador incluyen un código configurado para identificar un nodo fuente desde el cual mover un número de objetos de datos, en donde el nodo

fuente es un punto caliente. Las instrucciones legibles por ordenador también incluyen un código configurado para seleccionar los objetos de datos para moverlos desde el nodo fuente, en donde los objetos de datos son elegidos de tal manera que el nodo fuente ya no es un punto caliente. Además, las instrucciones legibles por ordenador incluyen un código configurado para determinar un nodo objetivo, en donde el nodo objetivo no es un punto caliente y está ubicado dentro de un intervalo predeterminado del nodo fuente, y mover los objetos de datos desde el nodo fuente al nodo objetivo.

Este compendio es proporcionado para introducir una selección de conceptos en una forma simplificada; estos conceptos se han descrito adicionalmente a continuación en la Descripción Detallada. Este Compendio no pretende identificar características clave o características esenciales de la materia reivindicada, ni pretende ser utilizado para limitar el alcance de la materia reivindicada.

Breve descripción de los dibujos

La Figura 1 es un sistema informático que puede ser utilizado para equilibrar carga en sistema de almacenamiento en grupos;

La Figura 2 es una realización de un sistema de almacenamiento en grupos para el cual el equilibrio de la carga puede ser realizado con el fin de conseguir una alta escalabilidad y reducir los costes de reconfiguración;

La Figura 3 es un diagrama de bloques de un método para equilibrar carga en sistemas de almacenamiento en grupos;

La Figura 4 es una ilustración de un enfoque de equilibrio de carga de “dividir y vencer” para grupos disjuntos de puntos calientes superpuestos;

La Figura 5 es un diagrama de bloques que muestra un medio legible por ordenador, no transitorio, tangible que almacena códigos adaptados para realizar equilibrio de carga en sistemas de almacenamiento en grupos.

Los mismos números se han utilizado a lo largo de la descripción y las figuras para hacer referencia a componentes y características similares. Los números en las serie 100 hacen referencia a características originalmente encontradas en la Figura 1, los números en la serie 200 hacen referencia a características originalmente encontradas en la Figura 2, los números en la serie 300 hacen referencia a las características originalmente encontradas en la Figura 3, y así sucesivamente.

Descripción detallada

Diferentes realizaciones descritas en esta memoria exponen un método y sistema para equilibrar carga en sistemas de almacenamiento en grupos. El equilibrio de carga puede ser logrado migrando objetos de datos específicos desde un nodo fuente a cualquier número de nodos objetivo. El nodo fuente puede ser un “punto caliente”, en donde un punto caliente es un nodo con una carga agregada, p. ej., un número total de operaciones de entrada/salida (I/O) por segundo o bytes transferidos por segundo para operaciones de lectura y escritura, que sobrepasan un nivel de umbral predeterminado. Además, los objetos de datos específicos que han de ser movidos desde el nodo fuente al nodo objetivo pueden ser elegidos basándose en el número de operaciones de I/O, tamaño, o solicitudes, entre otros criterios, para cada objeto de datos. Los objetos de datos pueden estar dispuestos en orden de acuerdo con un valor de carga para cada objeto de datos, en donde el valor de carga especifica el número de solicitudes de I/O para un objeto de datos específico. El objeto de datos con el valor de carga más alto puede ser hecho migrar desde el nodo fuente al primer nodo objetivo. La migración de objetos de datos puede continuar en orden descendente, es decir, desde el valor de carga más alto al valor de carga más bajo, hasta que la carga agregada para el nodo fuente caiga por debajo del valor de umbral predeterminado.

El sistema y método descritos en la presente memoria pueden realizar la migración de los objetos de datos de acuerdo con “operaciones de movimiento”, que implican la migración de objetos de datos entre dos particiones. Las dos particiones pueden ser fragmentos de datos almacenados dentro de dos nodos separados, en donde un fragmento puede ser un fragmento de información de un cierto tamaño predeterminado. Las operaciones de movimiento pueden ser realizadas de tal manera que los objetos de datos son movidos desde un nodo fuente al nodo objetivo apropiado más cercano. Además, en diferentes realizaciones, se puede suponer que las “operaciones de intercambio” ya han sido realizadas antes de las operaciones de movimiento. Las operaciones de intercambio pueden implicar el cambio de una función de una réplica principal y de una réplica secundaria de una partición, en donde la réplica principal y la réplica secundaria pueden estar ubicadas en dos nodos separados dentro del sistema de almacenamiento en grupos.

Como una cuestión preliminar, algunas de las figuras describen conceptos en el contexto de uno o más componentes estructurales, a los que se hace referencia de diferentes formas como funcionalidad, módulos, características, elementos, etc. Los distintos componentes mostrados en las figuras pueden ser implementados de cualquier forma, por ejemplo, por software, hardware (p. ej., componentes de lógica discretos, etc.), firmware, y así sucesivamente, o cualquier combinación de estas implementaciones. En una realización, los diferentes componentes pueden reflejar la utilización de componentes correspondientes en una implementación real. En otras realizaciones, cualquier componente individual ilustrado en las figuras puede ser implementado por un número de componentes reales. La representación de cualesquiera dos o más componentes separados en las figuras puede reflejar diferentes funciones realizadas por un solo

componente real. La Figura 1 proporciona detalles sobre un sistema que puede ser utilizado para implementar las funciones mostradas en las figuras.

Otras figuras describen los conceptos en forma de diagrama de flujo. En esta forma, se han descrito ciertas operaciones como que constituyen distintos bloques realizados en un cierto orden. Tales implementaciones son ejemplares y no limitativas. Ciertos bloques descritos en la presente memoria pueden ser agrupados juntos y realizados en una sola operación, ciertos bloques pueden ser descompuestos en bloques de componentes plurales, y ciertos bloques pueden ser realizados en un orden que difiere del que se ha ilustrado en la presente memoria, que incluye una forma paralela de realizar los bloques. Los bloques mostrados en los diagramas de flujo pueden ser implementados mediante software, hardware, firmware, procesamiento manual, y similares, o cualquier combinación de estas implementaciones. Como se ha utilizado en la presente memoria, el hardware puede incluir sistemas informáticos, componentes lógicos discretos, tales como circuitos integrados de aplicación específica (ASIC), y similares, así como cualesquiera combinaciones de los mismos.

En cuanto a la terminología, la frase “configurado para” abarca cualquier forma en la que se pueda construir cualquier tipo de funcionalidad para realizar una operación identificada. La funcionalidad puede estar configurada para realizar una operación utilizando, por ejemplo, software, hardware, firmware y similares, o cualesquiera combinaciones de los mismos.

El término “lógica” abarca cualquier funcionalidad para realizar una tarea. Por ejemplo, cada operación ilustrada en los diagramas de flujo corresponde a una lógica para realizar esa operación. Una operación puede ser realizada utilizando, por ejemplo, software, hardware, firmware, etc., o cualquier combinación de los mismos.

Como se ha utilizado en esta memoria, los términos “componente”, “sistema”, “cliente” y similares están destinados a referirse a una entidad relacionada con el ordenador, ya sea hardware, software (p. ej., en ejecución), y/o firmware, o una combinación de los mismos. Por ejemplo, un componente puede ser un proceso que se ejecuta en un procesador, un objeto, un ejecutable, un programa, una función, una biblioteca, una subrutina, y/o un ordenador o una combinación de software y hardware.

A modo de ilustración, tanto una aplicación que se ejecuta en un servidor como el servidor pueden ser un componente. Uno o más componentes pueden residir dentro de un proceso y un componente puede estar localizado en un ordenador y/o estar distribuido entre dos o más ordenadores. El término “procesador” se entiende generalmente que se refiere a un componente de hardware, tal como una unidad de procesamiento de un sistema informático.

Además, la materia reivindicada puede ser implementada como un método, aparato, o artículo de fabricación utilizando programación estándar y/o técnicas de ingeniería para producir software, hardware, o cualquier combinación de los mismos para controlar un ordenador para implementar la materia descrita. El término “artículo de fabricación” como se ha utilizado en la presente memoria pretende abarcar un programa informático accesible desde cualquier dispositivo legible por ordenador no transitorio, o medio.

El medio de almacenamiento legible por ordenador no transitorio puede incluir pero no está limitado a dispositivos de almacenamiento magnético (p. ej., disco duro, disco flexible, y cintas magnéticas, entre otros), discos ópticos (por ejemplo, disco compacto (CD), y disco versátil digital (DVD), entre otros), tarjetas con circuito integrado, y dispositivos de memoria flash (p. ej., tarjeta, pincho, y unidad de llave, entre otros). Por el contrario, el medio legible por ordenador generalmente (es decir, no necesariamente los medios de almacenamiento) puede incluir adicionalmente medios de comunicación tales como medios de transmisión para señales inalámbricas y similares.

La Figura 1 es un sistema informático 100 que puede ser utilizado para equilibrar carga en sistemas de almacenamiento en grupos. El sistema informático 100 puede estar incluido dentro de un dispositivo informático, tal como un ordenador de sobremesa, un ordenador portátil, o un dispositivo móvil, entre otros. Además, el sistema informático 100 puede ser implementado dentro del servidor, en donde el servidor puede incluir un servidor Web o un servidor en la nube, entre otros. El sistema informático 100 puede incluir un procesador 102 que está adaptado para ejecutar instrucciones almacenadas, así como un dispositivo de memoria 104 que almacena instrucciones que son ejecutables por el procesador 102. El procesador 102 puede ser un procesador de un solo núcleo, un procesador de múltiples núcleos, un grupo informático, o cualquier número de otras configuraciones. El dispositivo de memoria 104 puede incluir memoria de acceso aleatorio (RAM), memoria de solo lectura (ROM), memoria flash, o cualesquiera otros sistemas de memoria adecuados. Las instrucciones almacenadas ejecutadas por el procesador 102 pueden implementar un método para equilibrar carga en sistema de almacenamiento en grupos moviendo objetos de datos específicos desde un nodo fuente a cualquier número de nodos objetivo. El procesador 102 puede estar conectado a través de un bus 106 a uno o más dispositivos de entrada y salida.

El sistema informático 100 puede incluir un dispositivo de almacenamiento 108 adaptado para almacenar un módulo 110 de equilibrio de carga. En diferentes realizaciones, el módulo 110 de equilibrio de carga puede interactuar o coordinarse con otros módulos alojados en el dispositivo de almacenamiento 108 o componentes del sistema de control dentro del sistema informático 100. Además, en diferentes realizaciones, el sistema informático 100 puede ser un sistema distribuido, y el módulo 110 de equilibrio de carga puede ser un programa distribuido.

El dispositivo de almacenamiento 108 puede incluir un disco duro, una unidad óptica, una unidad USB, una agrupación

de unidades, o cualesquiera combinaciones de las mismas. Un controlador de interfaz de red (NIC) 112 puede ser incluido para conectar el sistema informático 100 a través del bus 106 a una red 114. A través de la red 114, el sistema informático 100 puede está acoplado de forma comunicativa a múltiples nodos 116. En diferentes realizaciones, los nodos 116 pueden ser nodos de almacenamiento, tales como ordenadores personales (PC) de productos básicos, o servidores. Además, los nodos 116 pueden ser utilizados para proporcionar acceso de lectura y escritura a los datos. Además, los nodos 116 también pueden estar interconectados entre sí a través de la red 114. En algunas realizaciones, el sistema informático 100 puede estar conectado a decenas de miles de los nodos 116 organizados en una topología de árbol en expansión, en donde la topología de árbol en expansión es un diseño de red que incluye enlaces redundantes entre los nodos 116 para proporcionar trayectorias de copia de seguridad automáticas si un enlace activo entre los nodos 116 falla. En algunas realizaciones, el motor de vigilancia 117 puede ser ejecutado en los nodos 116 y puede proporciona la utilización de la carga y otras medidas para el módulo 110 de equilibrio de carga.

En algunas realizaciones, una interfaz hombre-máquina (HMI) 118 puede conectar el sistema informático 100 a cualquiera de un número de dispositivos 120 de interfaz de usuario, tales como una pantalla táctil, un altavoz, o un dispositivo señalador, entre otros. El sistema informático 100 también puede estar enlazado a través del bus 106 a una interfaz 122 de I/O adaptada para conectar el sistema informático 100 a cualquier número de dispositivo 124 de I/O. Los dispositivos 124 de I/O pueden incluir, por ejemplo, un dispositivo de visualización, un disco duro externo, un dispositivo de Bus de Serie Universal (USB), o una impresora, entre otros. Además, el sistema informático 100 puede no incluir todos los componentes descritos en la presente memoria en cada caso, y se puede incluir cualquier número de componentes adicionales dentro del sistema informático 100.

La Figura 2 es una realización de un sistema 200 de almacenamiento en grupos para el que el equilibrio de carga puede ser realizado con el fin de conseguir un alta escalabilidad y reducir los costes de reconfiguración. En diferentes realizaciones, el sistema 200 de almacenamiento en grupos puede ser replicado, basándose en grupos, incluyendo el sistema de almacenamiento de objetos de datos cualquier número de nodos, tales como los nodos 202, 204, 206, y 208. Los nodos 202, 204, 206, y 208 pueden ser nodos de almacenamiento, tales como PC de productos básicos, o servidores. Además, el sistema 200 de almacenamiento en grupos puede incluir decenas de miles de los nodos 202, 204, 206, y 208 interconectados en una red de centro de datos, en donde la red de centro de datos puede estar organizada en una topología de árbol de expansión, como se ha tratado con respecto a la Figura 1. En diferentes realizaciones, la conexión de red entre los nodos 202, 204, 206, y 208 dentro del sistema 200 de almacenamiento en grupos puede ser implementada utilizando cualquier número de conmutadores o router de red 210, 212, 214, 216, 218, y 220. Además, la conexión entre los nodos 202, 204, 206, y 208 puede ser una conexión inalámbrica o con cable, dependiendo de la aplicación específica.

Debido a que los nodos 202, 204, 206, y 208 pueden estar organizados en una topología de árbol en expansión, la distancia de red entre múltiples nodos puede variar. Esto puede dar como resultado diferencias en la conectividad de ancho de banda entre diferentes nodos. Por ejemplo, dos nodos en la misma red o grupo pueden tener una conectividad de ancho de banda superior que dos nodos en diferentes redes o grupos. Así, la proximidad de un nodo a otro nodo afecta a los costes de migración de datos de mover objetos de datos entre los dos nodos.

Los datos pueden ser almacenados dentro de cada uno de los nodos 202, 204, 206, y 208 en unidades o particiones, o fragmentos. Una partición puede ser una unidad de datos de algún tamaño fijo con un cierto número de réplicas, en donde las replicas puede ser copias idénticas de la partición. Además, para cada partición, puede haber una réplica principal y cualquier número de réplicas secundarias. En diferentes realizaciones, las réplicas pueden habilitar la tolerancia de fallos asegurando que el fallo de una réplica de una partición no resulte en la pérdida de datos, ya que los datos también existen en todas las otras réplicas de la partición. Además, cada réplica de una partición particular puede ser colocada en un campo de fallo separado, p. ej., un nodo o red separado, con el fin de aumentar la tolerancia de fallos de la partición.

En diferentes realizaciones, el sistema 200 de almacenamiento en grupos puede estar acoplado de forma comunicativa a cualquier número de dispositivos de cliente 222, en donde los dispositivos de cliente 222 pueden incluir teléfonos móviles, tabletas, ordenadores de sobremesa, ordenadores portátiles, lectores electrónicos, televisiones, o reproductores multimedia, entre otros. Los dispositivos de cliente 222 pueden ser utilizados para iniciar las solicitudes de I/O, en donde las solicitudes de I/O pueden incluir solicitudes para realizar operaciones de lectura o escritura, o transacciones, dentro del sistema 200 de almacenamiento en grupos. Por lo tanto, en general, los dispositivos de cliente 222 pueden incluir cualquier tipo de dispositivo informático que sea capaz de iniciar tales solicitudes de I/O.

Además, las solicitudes de I/O de los dispositivos de cliente 222 pueden ser enviadas inicialmente a un servicio de metadatos 224, como se ha indicado por la flecha 226. El servicio de metadatos 224 puede determinar que objetos de datos son afectados por una solicitud de I/O particular recibida desde un dispositivo de cliente 222. Además, el servicio de metadatos 224 puede asignar cada objeto de datos a su partición constituyente dentro del sistema 200 de almacenamiento en grupos y cada partición a su réplica principal actual, como se ha indicado por la flecha 228. Además, en algunas realizaciones, el servicio de metadatos puede sondear periódicamente cada nodo 202, 204, 206, y 208 dentro del sistema 200 de almacenamiento en grupos con el fin de rastrear la disponibilidad de cada nodo, como se ha indicado por la flecha 228. El servicio de metadatos también puede utilizar asignaciones para mantener la coherencia de lectura y escritura entre las réplicas de cada partición dentro de los nodos 202, 204, 206, y 208.

Después de que el servicio de metadatos 224 haya asignado cada objeto de datos a su partición constituyente y a la réplica principal correspondiente dentro del sistema 200 de almacenamiento de datos, el servicio de metadatos 224 puede devolver la información de asignación que pertenece a los objetos de datos relevantes al dispositivo de cliente 222 que ha iniciado la solicitud de I/O particular, como se ha indicado por la flecha 228. El dispositivo de cliente 222 puede entonces ejecutar las operaciones de lectura y escritura especificadas por la solicitud de I/O particular accediendo a los nodos 202, 204, 206, y 208 apropiados dentro del sistema 200 de almacenamiento en grupos, como se ha indicado por la flecha 230.

De acuerdo con algunas realizaciones, para una solicitud de escritura, el dispositivo de cliente 222 puede enviar un mensaje el sistema 200 de almacenamiento en grupos especificando una identificación (ID) de partición particular en relación a la partición deseada a la que se han de escribir los datos, así como el tamaño de los datos. El dispositivo de cliente 222 puede enviar entonces los datos reales para la solicitud de escritura a la partición deseada dentro del sistema 200 de almacenamiento en grupos. Dentro de la partición deseada, la réplica principal puede determinar el orden apropiado de la solicitud de escritura y enviar la solicitud de escritura a las réplicas secundarias. Si la solicitud de escritura es válida, la réplica principal puede enviar un mensaje de “aceptar” de vuelta al dispositivo de cliente 222 con el fin de confirmar la ejecución de la solicitud de escritura.

La Figura 3 es un diagrama de bloques de un método 300 para equilibrar carga en sistemas de almacenamiento en grupos. De acuerdo con el método 300, se puede alcanzar el equilibrio de carga a través de la migración de objetos de datos entre dos nodos. Además, se puede alcanzar la migración de datos mediante operaciones de movimiento, lo que da como resultado la migración de objetos de datos entre múltiples nodos. Además, se puede suponer que las operaciones de intercambio ya han sido realizadas antes del comienzo del método 300. En algunas realizaciones, las operaciones de intercambio pueden ser utilizadas para equilibrar carga dentro de un sistema de almacenamiento en grupos cambiando una función de una réplica principal y una réplica secundaria de una partición, en donde la réplica principal y la réplica secundaria pueden estar ubicadas en dos nodos separados dentro del sistema de almacenamiento en grupos.

El método 300 puede ser utilizado para minimizar una carga total de un punto caliente, o nodo fuente, en un sistema de almacenamiento en grupos migrando un cierto número de objetos de datos lejos del punto caliente. El objetivo del método 300 puede ser minimizar el coste de ancho de banda, o el coste de la migración de datos, de mover un objeto de datos particular desde un nodo a otro nodo. En algunas realizaciones, este objetivo puede ser expresado como sigue:

$$\textit{Minimizar} \qquad \Sigma_i \ \Sigma_j \ X_{ij} \ B_{ij}, \qquad \text{Ec. 1}$$

en donde X_{ij} es igual a 1 si el objeto i-th se mueve al nodo j-th o de otra manera 0, y B_{ij} es el coste de ancho de banda de mover el objeto i-th al nodo j-th. Además, el método 300 puede intentar equilibrar la carga en el punto caliente en el sistema de almacenamiento en grupos de acuerdo con varias restricciones, o condiciones. Tal procedimiento de equilibrio puede ser utilizado para aliviar la sobrecarga en el punto caliente. En diferentes realizaciones, las condiciones pueden ser expresadas como sigue:

$$\forall j: \sum_i X_{ij} L_i - \sum_{\nu \in S_j} \sum_{j'} X_{\nu j'} L_\nu + L_{*j} \leq C_j, \qquad \text{Ec. 2}$$

$$\forall i: \sum_{j \in R(i,r)} X_{ij} = 1, \qquad \text{Ec. 3}$$

$$\forall i, \forall j \notin R(i,r): X_{ij} \leq 0, \qquad \text{Ec. 4}$$

$$\forall p, \forall q: \sum_{i \in G_p} \sum_{j \in F_q} X_{ij} \leq 1, \text{ y} \qquad \text{Ec. 5}$$

$$\forall i, \forall q: \sum_{j \in F_q} X_{ij} \leq I_{iq}, \qquad \text{Ec. 6}$$

en donde C_j es la capacidad de carga del nodo j-th, S_j es el conjunto de objetos de carga seleccionado para ser movido desde el nodo j-th, L_i es la carga del objeto i-th, L_{*j} es la carga total del objeto j-th, R(i, r) es el número de nodos dentro de un radio r del nodo al que pertenece el objeto i-th, G_p es el conjunto de datos seleccionado con el índice de partición p a mover desde el nodo fuente sobrecargado, F_q es el conjunto de nodos con el mismo índice de campo de fallo q, e I_{iq} es igual a 0 si el campo q-th contiene la misma partición que el objeto i-th o de otra manera 1. Como se ha expresado por la Ec. 2, la primera restricción es que, para cada nodo j-th en el sistema de almacenamiento en grupos, la carga total ha de permanecer por debajo de un valor de umbral predeterminado, es decir, la capacidad de carga C_j. Como se ha expresado por las Ec. 3 y 4, la segunda restricción es que el nodo objetivo ha de estar al menos dentro de una cierta distancia, o radio r, del nodo fuente. Se puede comprender que, como se emplea en la presente memoria, el radio r puede ser considerado un proxy para el número de router, conmutadores, y similares, que los objetos de datos deben pasar a través para alcanzar un nodo objetivo. Esto puede reflejarse a menudo en la proximidad física de dos nodos, pero tal proximidad física no es requerida.

Además, como se ha expresado por las Ec. 5 y 6, la tercera restricción asegura la tolerancia de fallos impidiendo que los objetos de datos con la misma partición se coloquen en el mismo campo de fallo, p. ej., en el mismo nodo o red. En particular, G_p indica el grupo de todas las réplicas que pertenecen a un índice de partición p en el conjunto de objetos de datos seleccionados. Así, la Ec. 5 impone una restricción de tal manera que los objetos de datos candidatos que tienen el mismo índice de partición no pueden ser colocados el mismo campo de fallos, mientras las Ec. 6 impone una restricción de tal manera que un objeto de datos candidato no puede ser colocado en un campo de fallos que ya tiene una réplica de la misma partición. Las restricciones de tolerancia de fallos impuestas por las Ec. 5 y 6 pueden ser particularmente útiles cuando un conjunto de objetos de datos candidatos es fusionado desde los nodos con regiones de búsqueda superpuestas. Para un solo nodo, sin embargo, estas restricciones se mantienen de forma trivial, ya que como máximo una copia de una partición de réplica puede estar alojada en el nodo. Además, las restricciones impuestas por las Ec. 2-6 pueden ser utilizadas para el desarrollo y la implementación de las operaciones del método 300, que son tratadas en detalle a continuación.

El método 300 comienza en el bloqueo 302 con la identificación de un nodo fuente dentro de un sistema de almacenamiento en grupos desde el cual mover un número de objetos de datos. El método 300 puede ser ejecutado de acuerdo con un caso de un sistema de equilibrio de carga, en donde el sistema de equilibrio de carga está configurado para equilibrar la carga de la solicitud de I/O en nodos dentro de un sistema de almacenamiento en grupos particular. Además, el sistema de equilibrio de carga puede ser un sistema middleware de gestión de datos en una plataforma informática en la nube que permite la interacción entre múltiples nodos, o servidores, contenidos dentro de un sistema de almacenamiento en grupos particular.

En diferentes realizaciones, el nodo fuente puede ser un punto caliente, es decir, un nodo con una carga agrupada, o total que sobrepasa un valor de umbral predeterminado. La carga total para un nodo particular puede ser el número total de operaciones de I/O por segundo o el número total de bytes transferidos por segundo para las operaciones de lectura y escritura para el nodo. Además, el valor de umbral para determinar si un nodo es un punto caliente puede ser determinado de tal manera que un tiempo de respuesta del nodo puede no aumentar por encima de un valor específico o tasa de rendimiento medido en términos de si el número de transacciones procesado en una ventana de tiempo dada sobrepasa un valor específico.

Además, en algunas realizaciones, un servicio de metadatos puede ser utilizado para identificar el nodo fuente a partir del cual mover los objetos de datos. El servicio de metadatos puede identificar el nodo fuente en respuesta a un influjo de un número predeterminado de solicitudes de I/O para el nodo fuente particular desde cualquiera de un número de dispositivos de cliente. El número predeterminados de solicitudes de I/O puede ser elegido de tal manera que, si el número predeterminado de solicitudes de I/O es recibido para un nodo particular, el nodo puede ser considerado un punto caliente. Adicionalmente, en diferentes realizaciones, el servicio de metadatos puede ayudar a ejecutar el método 300 asignando solicitudes de I/O entrantes desde un dispositivo de cliente a objetos de datos específicos dentro de cualquiera de los nodos dentro del sistema de almacenamiento en grupos, que incluye un nodo fuente o un nodo objetivo, o ambos.

En el bloque 304, se pueden seleccionar los objetos de datos que han de ser movidos desde el nodo fuente. En diferentes realizaciones, los objetos de datos pueden ser seleccionados de tal manera que, una vez que los objetos de datos han sido movidos lejos del nodo fuente, la carga total del nodo fuente ya no sobrepasará un nivel de umbral. Así, puede ser deseable seleccionar objetos de datos basándose en un valor de carga de cada objeto de datos individual. El valor de carga puede ser igual al número de solicitudes de I/O para un objeto de datos particular dentro del nodo fuente. En algunas realizaciones, se puede asignar una variable de decisión entera a cada objeto de datos dentro del nodo fuente. Si la variable de decisión entera es igual a 1, indicando que el objeto de datos particular tiene un valor de carga alto, el objeto de datos puede ser movido lejos del nodo fuente. Si la variable de decisión entera es igual a 0, indicando que el objeto de datos particular tiene un valor de carga bajo, el objeto de datos puede permanecer dentro del nodo fuente. Además, en diferentes realizaciones, se puede asignar una variable de decisión real a cada objeto de datos dentro del nodo fuente. La variable de decisión real puede ser igual a cualquier número real entre 0 y 1, inclusive, en donde 0 indica que el objeto de datos tiene un valor de carga de 0. Además, los objetos de datos pueden estar dispuestos en orden descendente de acuerdo con su variable de decisión real, y los objetos de datos con las variables de decisión real más altas, que indican que tienen el número más alto de solicitudes de I/O, puede ser seleccionados para ser movidos en primer lugar. Además, en algunas realizaciones, los objetos de datos que han de ser movidos desde el nodo fuente pueden ser seleccionados aleatoriamente.

En el bloque 306, se puede determinar un nodo objetivo dentro del sistema de almacenamiento en grupos basándose en la carga total del nodo objetivo y en la proximidad del nodo objetivo al nodo fuente. Por ejemplo, el nodo objetivo puede ser cualquier nodo dentro del sistema de almacenamiento en grupos que tiene una carga total que no sobrepasa un valor de umbral predeterminado, es decir, que no es un punto caliente, y que está dentro de una cierta distancia predeterminada, o radio, desde el nodo fuente. En algunos casos, se pueden determinar múltiples nodos objetivo para un caso particular del método 300, y los objetos de datos pueden ser movidos de forma selectiva a cualquiera de los nodos objetivo. Los nodos objetivo que están más cerca del nodo fuente pueden ser preferidos, ya que los costes de la migración de datos aumentan cuando la distancia entre el nodo objetivo y el nodo fuente aumenta. Así, se puede especificar un radio inicial desde el nodo fuente para el área en el que identifica un nodo objetivo apropiado. Después, si no se puede encontrar un nodo objetivo apropiado dentro del radio inicial, el radio puede ser aumentado de forma creciente hasta que se encuentre un nodo objetivo apropiado. Además, como se ha tratado anteriormente, el radio, o

distancia radial, puede ser considerado un proxy para el número de router, conmutadores, y similares, que los objetos de datos deben hacer pasar a su través para alcanzar el nodo objetivo. Esto puede reflejarse en la proximidad física de dos nodos, pero no se requiere tal proximidad física.

En el bloque 308, los objetos de datos pueden ser movidos desde el nodo fuente al nodo objetivo. En algunas realizaciones, la eficiencia del método 300 puede ser aumentada moviendo de forma selectiva objetos de datos específicos a cualquiera de un número de nodos objetivo. Además, los objetos de datos pueden ser movidos, o migrados, de acuerdo con la ruta posible más corta desde el nodo fuente al nodo objetivo apropiado, ya que los costes de la migración de datos aumentan cuando la distancia de desplazamiento de un objeto de datos aumenta.

Además, el método 300 no pretende indicar que las operaciones del método 300 han de ser ejecutadas en cualquier orden particular o que todas las operaciones han de estar presentes en cada caso. Además, las operaciones pueden ser añadidas al método 300 de acuerdo con la aplicación específica. Por ejemplo, el método 300 puede ser utilizado para identificar en paralelo un número de conjuntos de nodos fuente no superpuestos con cargas totales que sobrepasan el valor de umbral. Los nodos fuente dentro de un conjunto particular de nodos fuente pueden estar dentro de una distancia radial específica unos de otros, es decir, pueden tener radios de búsqueda superpuestos. Se puede seleccionar un número de objetos de datos para ser movido desde cada uno de los conjuntos de nodos fuente, se puede determinar cualquier número de nodos objetivo apropiados dentro de un radio de búsqueda específico desde cada uno de los conjunto de nodos fuente, y los objetos de datos pueden ser movidos simultáneamente desde cada uno de los conjuntos de nodos fuente a los nodos objetivo. Múltiples casos del sistema de equilibrio de carga pueden ser ejecutados en paralelo con el fin de mover simultáneamente los objetos de datos desde cada uno de los conjuntos de nodos fuente a los nodos objetivo elegidos. En diferentes realizaciones, tal método para realizar la migración de datos simultáneamente para un número de conjuntos de nodo fuente disjuntos, o no superpuestos, puede ser denominado como un enfoque de equilibrio de carga de “dividir y vencer”.

Además, en diferentes realizaciones, el movimiento de los objetos de datos puede dar como resultado interferencias con aplicaciones que se ejecutan en primer plano en la red durante la reconfiguración de los objetos de datos y/o dentro de los nodos fuente y objetivo correspondientes. Esto puede ser aliviado especificando un presupuesto de ancho de banda de recursos, o separando el tráfico en primer plano y el tráfico en un plano posterior a través de la utilización de protocolos de transporte.

La Figura 4 es una ilustración 400 de un enfoque de equilibrio de carga de dividir y vencer para grupos disjuntos de puntos calientes superpuestos. La ilustración 400 puede representar un número de nodos fuente dentro de un sistema de almacenamiento en grupos. El sistema de almacenamiento en grupos puede incluir un primer punto caliente 402, un segundo punto caliente 404, un tercer punto caliente 406, un cuarto punto caliente 408, y un quinto punto caliente 410. Además, los puntos calientes 402, 404, 406, 408, y 410 pueden estar incluidos cada uno dentro de un “vecindario” 412, 414, 416, 418, y 420, respectivamente, de un radio predeterminado, tal como el radio 422. El radio predeterminado puede ser un radio estándar utilizado para todos los puntos calientes 402, 404, 406, 408, y 410, o puede ser un radio único, específico para cada punto caliente individual 402, 404, 406, 408, o 410.

Los puntos calientes con vecindarios solapados pueden ser fusionados en los grupos 424 y 426. En diferentes realizaciones, los vecindarios 412 y 414 para el primer punto caliente 402 y el segundo punto caliente 404, respectivamente, pueden ser fusionados en el grupo 424. Además, los vecindarios 416, 418, y 420 para el tercer punto caliente 406, el cuarto punto caliente 408, y el quinto punto caliente 410 pueden ser fusionados en el grupo 426. Después, el sistema de equilibrio de carga puede ejecutar el método 300 para los grupos 424 y 426 en paralelo. Además, el equilibrio de carga para los puntos calientes dentro de cada grupo 424 o 426 puede ser ejecutado simultáneamente de acuerdo con el método 300. En diferentes realizaciones, este enfoque de equilibrio de carga de dividir y vencer puede dar como resultado una disminución en los costes de computación para el método 300 sin pérdida de precisión.

La Figura 5 es un diagrama de bloques que muestra un medio 500 legible por ordenador, no transitorio, tangible que almacena un código adaptado para realizar el equilibrio de carga en sistema de almacenamiento en grupos. El medio 500 legible por ordenador, no transitorio, tangible puede ser accedido por un procesador 502 sobre un bus informático 504. Además, el medio 500 legible por ordenador, no transitorio, tangible puede incluir un código configurado para dirigir el procesador 502 para realizar las operaciones del método actual. Los diferentes componentes de software tratados en la presente memoria pueden ser almacenados en el medio 500 legible por ordenador, no transitorio, tangible, como se ha indicado en la Figura 5. Por ejemplo, se puede utilizar un módulo 506 de equilibrio de carga para equilibrar carga en nodos específicos dentro de un sistema de almacenamiento en grupos migrando objetos de datos entre nodos. Además, el medio 500 legible por ordenador, no transitorio, tangible puede incluir componentes de software adicionales no mostrados en la Figura 5.

Aunque la materia ha sido descrita en un lenguaje específico para características estructurales y/o actos metodológicos, se ha de comprender que la materia definida en las reivindicaciones adjuntas no está limitada necesariamente a las características o actos específicos descritos anteriormente. Más bien, las características y actos específicos descritos anteriormente son divulgados como formas ejemplares de implementación de las reivindicaciones.

REIVINDICACIONES

1. Un método (300) para equilibrar carga en un sistema (200) de almacenamiento en grupo, que comprende:

identificar (302) una pluralidad de nodos fuente dentro del sistema (200) de almacenamiento en grupos a partir de los cuales mover una pluralidad de objetos de datos, en donde cada nodo fuente comprende un nodo con una carga total que sobrepasa un valor de umbral;

dividir la pluralidad de nodos fuente en una pluralidad de conjuntos de nodos fuente, en donde cada uno de la pluralidad de conjuntos de nodos fuente comprende un nodo fuente o un pequeño subconjunto de nodos fuente con radios de búsqueda superpuestos, un radio de búsqueda de nodo fuente correspondiente a una distancia de red desde el nodo fuente para buscar otro nodo;

seleccionar (304) la pluralidad de objetos de datos a mover desde cada uno de la pluralidad de conjuntos de nodos fuente, en donde la pluralidad de objetos de datos son elegidos en cada nodo de tal manera que la carga total de cada nodo fuente ya no sobrepase el valor de umbral;

determinar (306) una pluralidad de nodos objetivo dentro de un radio de búsqueda específico desde cada uno de la pluralidad de conjuntos de nodos fuente basándose en la carga total de cada nodo objetivo;

mover simultáneamente (308) la pluralidad de objetos de datos desde cada uno de la pluralidad de conjuntos de nodos fuente a la pluralidad de nodos objetivo, que comprende minimizar las interferencias con las aplicaciones que se ejecutan en primer plano en la red durante la reconfiguración de los objetos de datos especificando un presupuesto de ancho de banda de recursos o separando el tráfico en primer plano del tráfico en un plano posterior a través de la utilización de protocolos de transporte.

2. El método (300) de la reivindicación 1, que comprende mover cualquiera de la pluralidad de objetos de datos desde el nodo fuente a cualquiera de una pluralidad de nodos objetivo basándose en la carga total en cada uno de la pluralidad de nodos objetivo y en la proximidad de cada uno de la pluralidad de nodos objetivo al nodo fuente.

3. El método (300) de la reivindicación 1, en donde cada nodo fuente de la pluralidad de nodos fuente es un punto caliente.

4. El método (300) de la reivindicación 1, que comprende seleccionar la pluralidad de objetos de datos a mover desde el nodo fuente basándose en un valor de carga, en donde se seleccionan la pluralidad de objetos de datos comenzando con un objeto de datos con un valor de carga más alto y continuando en un orden descendente.

5. El método (300) de la reivindicación 4, en el que el valor de carga comprende un número de solicitudes de entrada/salida para un objeto de datos particular dentro del nodo fuente.

6. El método (300) de la reivindicación 1, en donde las operaciones de intercambio que comprenden cambiar una función de una réplica principal y de una réplica secundaria de una partición del objeto de datos que ha de ser movido, en donde la réplica principal y la réplica secundaria pueden estar ubicadas en dos nodos separados dentro del sistema de almacenamiento en grupos, son realizadas antes de mover el objeto de datos.

7. El método (300) de la reivindicación 1, que comprende seleccionar la pluralidad de objetos de datos a mover desde el nodo fuente basándose en una variable de decisión real asignada a cada objeto de datos dentro del nodo fuente, en donde la variable de decisión real comprende un número real entre cero y uno, o igual a cero o uno, y en donde la pluralidad de objetos de datos son seleccionados para ser movidos comenzando con un objeto de datos con una variable de decisión real más alta y continuando en un orden descendente.

8. Un medio (500) legible por ordenador que comprende un código configurado, cuando es ejecutado en un procesador, para dirigir el procesador (502) para realizar el método (300) de cualquiera de las reivindicaciones precedentes.

9. Un sistema para equilibrar carga en un sistema (200) de almacenamiento en grupos, que comprende:

una pluralidad de nodos fuente (402, 404, 406, 408, 410), en donde cada nodo fuente (402, 404, 406, 408, 410) comprende un nodo dentro del sistema (200) de almacenamiento en grupos con una primera carga total que sobrepasa un valor de umbral, en donde la pluralidad de nodos fuente (402, 404, 406, 408, 410) está dividida en una pluralidad de conjuntos de nodos fuente (424, 426), en donde cada uno de la pluralidad de conjuntos de nodos fuente (424, 426) comprende un nodo fuente o un pequeño subconjunto de nodos fuente con radios (422) de búsqueda superpuestos, correspondiendo un radio de búsqueda de un nodo fuente a una distancia de red desde el nodo fuente para buscar otro nodo;

una pluralidad de nodos objetivo, en donde cada uno de la pluralidad de nodos objetivo comprende un nodo dentro del sistema (200) de almacenamiento en grupos con una segunda carga total que no sobrepasa el valor de umbral, y en donde cada uno de la pluralidad de nodos objetivo está dentro de un radio de búsqueda específico desde cada uno de la pluralidad de conjuntos de nodos fuente (424, 426); y

un sistema de equilibrio de carga configurado para:

seleccionar una pluralidad de objetos de datos a mover desde cada uno de la pluralidad de nodos fuente (402, 404, 406, 408, 410) a un nodo objetivo asociado, en donde la pluralidad de objetos de datos son elegidos de tal manera que la primera carga total del nodo fuente ya no sobrepasa el valor de umbral; y

mover simultáneamente la pluralidad de objetos de datos desde cada uno de la pluralidad de conjuntos de nodos fuente (424, 426) a la pluralidad de nodos objetivo, que comprende minimizar la interferencia con las aplicaciones que se ejecutan en primer plano en la red durante la reconfiguración de los objetos de datos especificando un presupuesto de ancho de banda de recursos o separando el tráfico en primer plano y el tráfico en un plano posterior a través de la utilización de protocolos de transporte.

10. El sistema de la reivindicación 9, en donde la primera carga total y la segunda carga total comprenden un número total de operaciones de entrada/salida por segundo o un número total de bytes trasferidos por segundo para operaciones de lectura y escritura.

11. El sistema de la reivindicación 9, que comprende ejecutar múltiples casos del sistema de equilibrio de carga en paralelo con el fin de mover simultáneamente una pluralidad de objetos de datos desde cada uno de la pluralidad de nodos fuente (402, 404, 406, 408, 410), en donde la pluralidad de nodos fuente (402, 404, 406, 408, 410) comprende nodos no superpuestos que están al menos a una distancia de red predeterminada entre sí.

12. El sistema de la reivindicación 9, en donde cada nodo fuente de la pluralidad de nodos fuente (402, 404, 406, 408, 410) es un punto caliente.

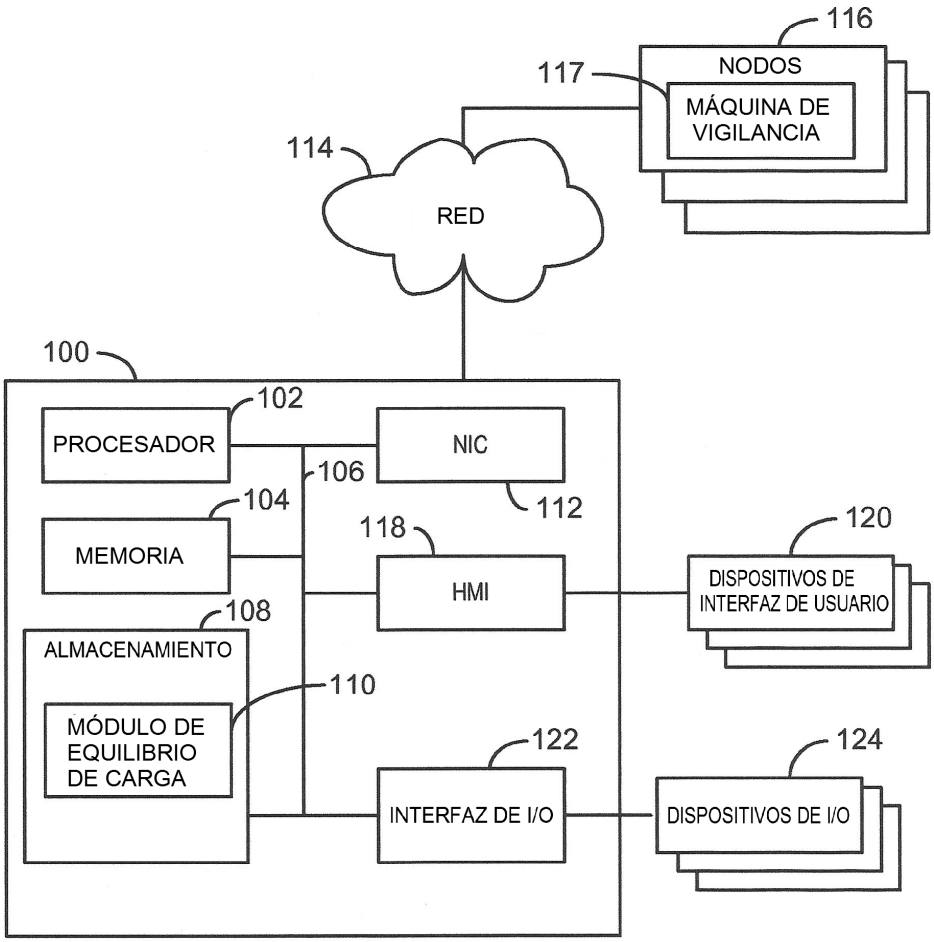

116
NODOS
117
MÁQUINA DE VIGILANCIA
114
RED
100
102
PROCESADOR
NIC
106
112
104
MEMORIA
118
HMI
120
DISPOSITIVOS DE INTERFAZ DE USUARIO
108
ALMACENAMIENTO
110
MÓDULO DE EQUILIBRIO DE CARGA
122
INTERFAZ DE I/O
124
DISPOSITIVOS DE I/O

FIG. 1

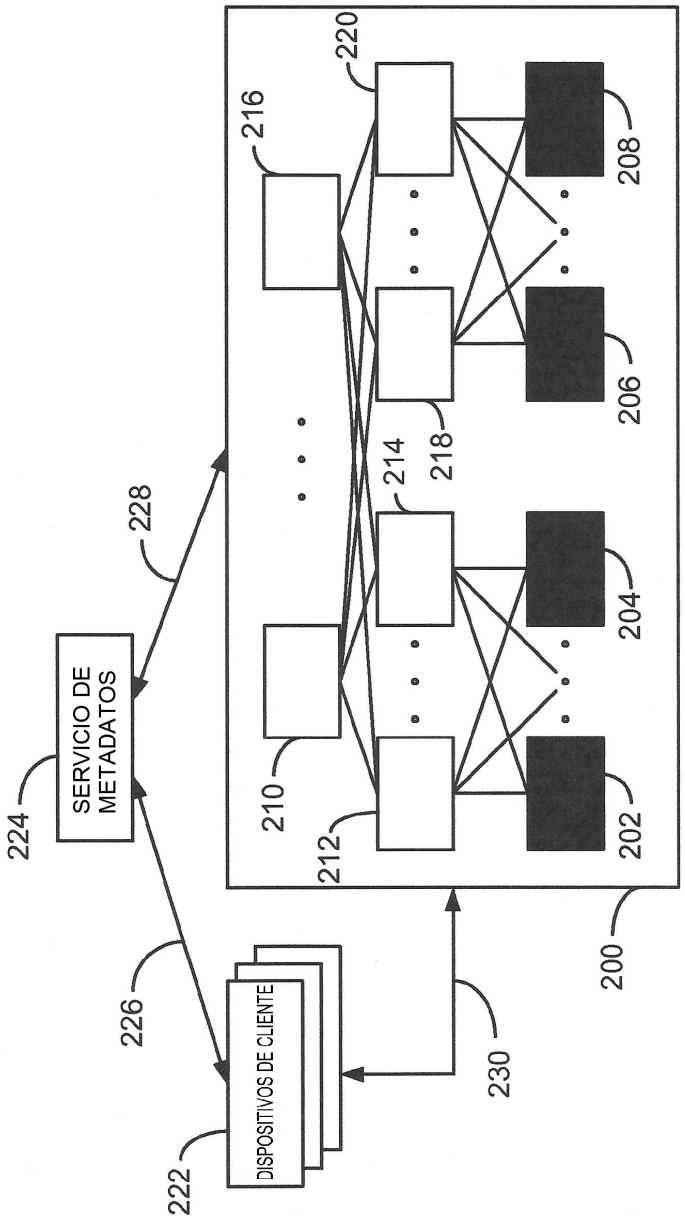
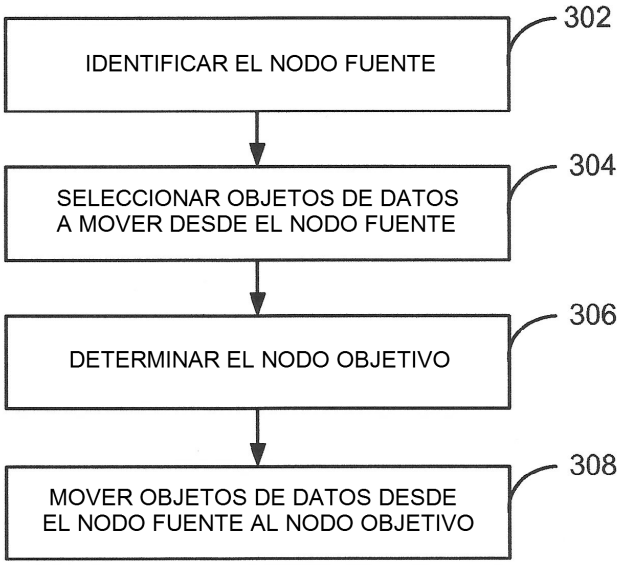

SERVICIO DE METADATOS
DISPOSITIVOS DE CLIENTE
200
202
204
206
208
210
212
214
216
218
220
222
224
226
228
230

FIG. 2


IDENTIFICAR EL NODO FUENTE
302
SELECCIONAR OBJETOS DE DATOS A MOVER DESDE EL NODO FUENTE
304
DETERMINAR EL NODO OBJETIVO
306
MOVER OBJETOS DE DATOS DESDE EL NODO FUENTE AL NODO OBJETIVO
308

300

FIG. 3

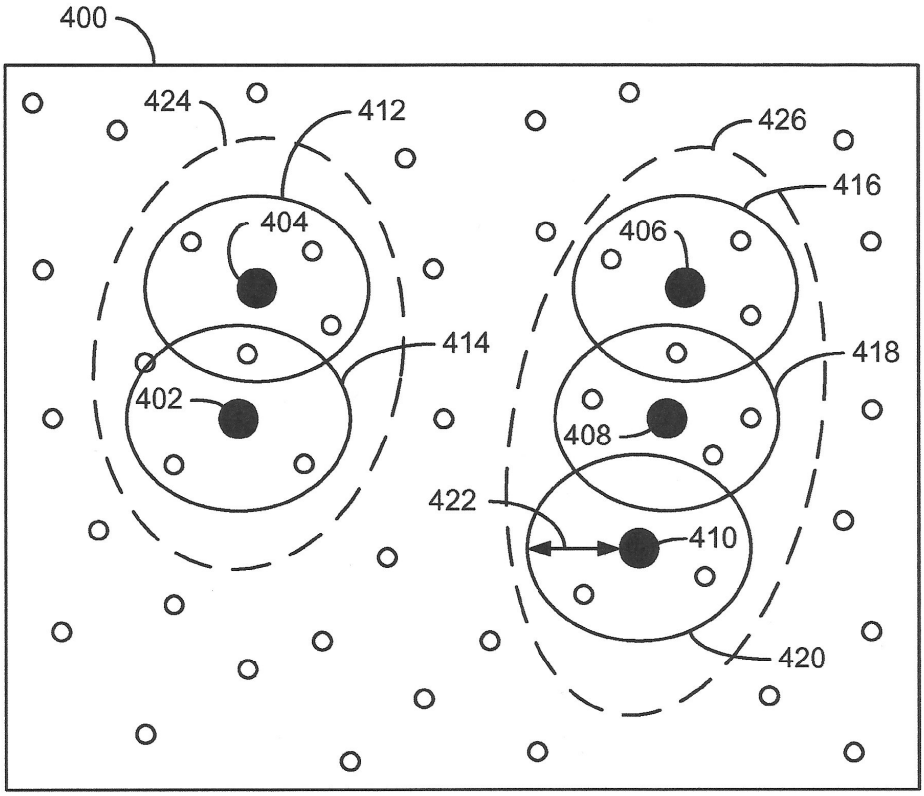


FIG. 4

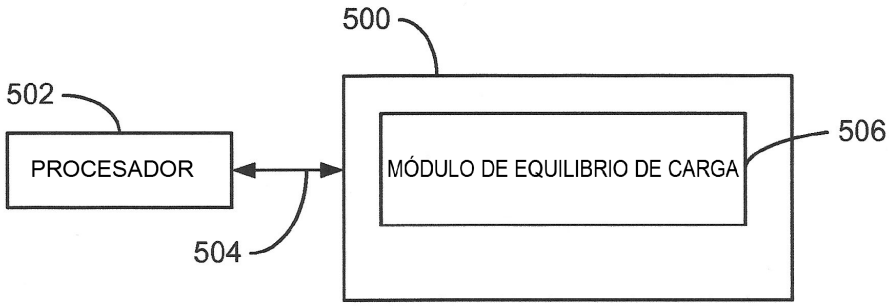


FIG. 5