(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 709 503**

(51) Int. CI.:

H04L 12/715 (2013.01)
H04L 12/721 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **31.08.2015 PCT/FR2015/052298**

(87) Fecha y número de publicación internacional: **10.03.2016 WO16034797**

(96) Fecha de presentación y número de la solicitud europea: **31.08.2015 E 15771674 (7)**

(97) Fecha y número de publicación de la concesión europea: **07.11.2018 EP 3189636**

(54) Título: **Procedimiento de supervisión y de alerta de configuración de encaminamiento en un grupo que comprende enlaces de comunicación estáticos, y programa informático que implementa este procedimiento**

(30) Prioridad:

03.09.2014 FR 1458250

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
16.04.2019

(73) Titular/es:

BULL SAS (100.0%)
Rue Jean Jaurès
78340 Les Clayes-sous-Bois, FR

(72) Inventor/es:

FICET, JEAN-VINCENT;
DUGUE, SÉBASTIEN y
GERPHAGNON, JEAN-OLIVIER

(74) Agente/Representante:

ELZABURU, S.L.P

Aviso: En el plazo de nueve meses a contar desde la fecha de publicación en el Boletín Europeo de Patentes, de la mención de concesión de la patente europea, cualquier persona podrá oponerse ante la Oficina Europea de Patentes a la patente concedida. La oposición deberá formularse por escrito y estar motivada; sólo se considerará como formulada una vez que se haya realizado el pago de la tasa de oposición (art. 99.1 del Convenio sobre Concesión de Patentes Europeas).

DESCRIPCIÓN

Procedimiento de supervisión y de alerta de configuración de encaminamiento en un grupo que comprende enlaces de comunicación estáticos, y programa informático que implementa este procedimiento

La presente invención hace referencia al encaminamiento en un grupo, es decir, a la determinación de rutas de comunicación entre un conjunto de nodos del grupo y, más particularmente, a un procedimiento de supervisión y de alerta de configuración de encaminamiento en un grupo que comprende enlaces de comunicación estáticos, así como un programa informático que implementa este procedimiento.

El cálculo de alto rendimiento, denominado también HPC (acrónimo de cálculo de alto rendimiento, *High Performance Computing*, en terminología anglosajona) se está desarrollando tanto para la investigación académica como para la industria, especialmente en campos de la técnica tales como la aeronáutica, la energía, la climatología y las ciencias biológicas. En particular, la modelización y la simulación ayudan a reducir los costes de desarrollo y aceleran el lanzamiento de productos innovadores, más fiables y de menor consumo de energía. Para los investigadores, el cálculo de alto rendimiento se ha convertido en un medio indispensable de investigación.

Estos cálculos son implementados, en general, en sistemas de procesamiento de datos denominados grupos (clusters, en inglés). Un grupo comprende habitualmente un conjunto de nodos interconectados. Ciertos nodos son utilizados para realizar tareas de cálculo (nodos de cálculo), otros para almacenar datos (nodos de almacenamiento) y uno o varios administran el grupo (nodos de administración). Cada nodo es, por ejemplo, un servidor que implementa un sistema operativo tal como Linux (Linux es una marca comercial). El enlace entre los nodos se realiza, por ejemplo, con la ayuda de enlaces de comunicación Ethernet o InfiniBand (Ethernet e Infiniband son marcas comerciales). Un ejemplo de la técnica anterior se encuentra en el documento WO2011/144848. La figura 1 muestra esquemáticamente un ejemplo de una topología 100 de un grupo, de tipo *fat-tree*. Este último comprende un conjunto de nodos referenciados de manera genérica como 105. Los nodos que pertenecen al conjunto 110 son en el presente documento nodos de cálculo, mientras que los nodos del conjunto 115 son nodos de servicio (nodos de almacenamiento y nodos de administración). Los nodos de cálculo pueden ser agrupados en subconjuntos 120 denominados bloques de cálculo, denominándose el conjunto 115 bloque de servicio.

Los nodos están conectados entre sí por conmutadores (denominados *switch* en terminología anglosajona), por ejemplo, de manera jerárquica. En el ejemplo mostrado en la figura 1, los nodos están conectados a conmutadores 125 de primer nivel que están conectados a su vez a conmutadores 130 de segundo nivel que, a su vez, están conectados a conmutadores 135 de tercer nivel.

Tal como se muestra en la figura 2, cada nodo comprende en general uno o varios microprocesadores, memorias locales, así como una interfaz de comunicación. De manera más precisa, el nodo 200 comprende en el presente documento un bus de comunicación 202 al que están conectados:

- unidades centrales de tratamiento o microprocesadores 204 (o CPU, acrónimo de *Central Processing Unit* en terminología anglosajona);

- componentes de la memoria de acceso aleatorio 206 (RAM, acrónimo de *Random Access Memory*, en terminología anglosajona), que comprende registros adaptados para registrar variables y parámetros creados y modificados durante la ejecución de programas (tal como se muestra, cada componente de la memoria de acceso aleatorio puede estar asociado a un microprocesador); e,

- interfaces de comunicación 208, adaptadas para transmitir y recibir datos.

El nodo 200 dispone además de medios de almacenamiento interno 212, tales como discos duros, que pueden comprender en particular el código ejecutable de programas.

El bus de comunicación permite la comunicación y la interoperabilidad entre los diversos elementos incluidos en el nodo 200 o conectados al mismo. Los microprocesadores 204 controlan y gestionan la ejecución de las instrucciones o partes del código de software del programa o de los programas. Durante el encendido, el programa o los programas almacenados en una memoria no volátil, por ejemplo, un disco duro, son transferidos en la memoria de acceso aleatorio 206.

En este caso, se observa que los rendimientos de un grupo están directamente ligados a la elección de las rutas que permiten la transferencia de datos entre los nodos, establecidas a través de los enlaces de comunicación. De manera general, enlaces físicos de comunicación están establecidos entre los nodos y los conmutadores durante la configuración hardware de un grupo, estando determinadas las propias rutas de comunicación en una fase de inicialización a partir de una definición de los enlaces que deben ser establecidos entre los nodos. Según la tecnología de comunicación implementada, la configuración de las vías puede ser estática o dinámica.

A modo de ilustración, la tecnología InfiniBand permite, en un grupo, una configuración estática de las rutas. Esta configuración utiliza tablas estáticas de encaminamiento (o LFT, acrónimo de *Linear Forwarding Table* en

terminología anglosajona) en cada conmutador. Cuando se implementa esta tecnología, se puede utilizar un algoritmo de encaminamiento, tal como los algoritmos conocidos como FTree, MINHOP, UPDN y LASH.

La elección del algoritmo que debe ser utilizado habitualmente la realiza un administrador según, entre otras cosas, la topología del grupo. Puede ser, por ejemplo, el algoritmo FTree. Sin embargo, si el algoritmo elegido no permite realizar el encaminamiento, el administrador del grupo (habitualmente encargado del encaminamiento) elige en general, de manera automática, otro algoritmo, por ejemplo, el algoritmo MINHOP (en general menos eficiente que el elegido inicialmente).

A modo de ilustración y de manera simplificada, el algoritmo FTree determina las rutas para que se distribuyan lo más posible a través de los enlaces de comunicación existentes. Para estos propósitos, cuando se encamina una red de comunicación totalmente conectada según una arquitectura de *fat-tree*, cada nodo de la red se considera de igual importancia. Asimismo, cuando se establece una ruta entre dos nodos de un mismo enlace, el número de rutas que utilizan este enlace, denominado la carga del enlace, se incrementa en uno. Cuando el algoritmo de encaminamiento intenta establecer una nueva ruta y surgen varias posibilidades, compara los niveles de carga asociados con los enlaces en los que se basan estas posibilidades y elige el que tiene el nivel más bajo de carga.

Durante la utilización del grupo, si un enlace o elemento, tal como un nodo o un conmutador falla, se realiza un nuevo encaminamiento.

Puesto que la calidad de encaminamiento afecta directamente a los rendimientos de un grupo, existe la necesidad de supervisar una configuración de encaminamiento en un grupo que comprende enlaces de comunicación estáticos y, cuando sea apropiado, alertar a un administrador de un potencial problema de encaminamiento.

La invención permite resolver al menos uno de los problemas expuestos anteriormente.

La invención tiene por objeto, por lo tanto, un procedimiento informático para la supervisión de al menos un parámetro de encaminamiento de un grupo que comprende una serie de nodos y una serie de conmutadores, conectando varios enlaces de comunicación estáticos nodos y conmutadores de dichas series de nodos y de conmutadores, comprendiendo cada conmutador una serie de puertos de salida, cuyo procedimiento se caracteriza por que comprende las etapas siguientes,

- selección de al menos un conmutador de dicha serie de conmutadores,

- cálculo de un número de rutas por puerto para cada puerto de cada conmutador seleccionado, siendo definidas las rutas durante una etapa de encaminamiento para vincular cada uno de los nodos a otro;

- cálculo de un número promedio de rutas por puerto para al menos un conmutador seleccionado;

- comparación de cada número de rutas por puerto calculado con el número promedio de rutas por puerto calculado; y

- en respuesta a dicha comparación, notificación de un desequilibrio potencial de encaminamiento de dicho grupo.

Por lo tanto, el procedimiento según la invención hace posible detectar y notificar un problema potencial de encaminamiento que puede provocar una reducción de los rendimientos del grupo.

Según un modo de realización particular, cada conmutador comprende una tabla de encaminamiento para identificar un puerto según una característica de un paquete de datos recibido, comprendiendo además el procedimiento una etapa de obtención de una tabla de encaminamiento para cada conmutador seleccionado, estando basada dicha etapa de cálculo de un número de rutas por puerto para cada puerto de cada conmutador seleccionado en una tabla de encaminamiento obtenida para cada conmutador seleccionado.

Según un modo de realización particular, dicho etapa de comparación de cada número de rutas por puerto con el número promedio de rutas por puerto comprende una etapa de cálculo de la diferencia de cada número de rutas por puerto con el número promedio de rutas por puerto y una etapa de comparación de la diferencia con al menos un umbral.

Según un modo de realización particular, dicha etapa de cálculo de la diferencia de cada número de rutas por puerto con el número promedio de rutas por puerto es una etapa de cálculo del valor absoluto de la diferencia de cada número de rutas por puerto con el número promedio de rutas por puerto.

Según un modo de realización particular, la diferencia entre cada número de rutas por puerto y el número promedio de rutas por puerto se considera cero si es negativa.

Según un modo de realización particular, dicha comparación de la diferencia con al menos un umbral comprende una etapa de selección de un primer o un segundo umbral dependiendo de si la diferencia es positiva o negativa.

Según un modo de realización particular, el procedimiento comprende además las etapas siguientes:

- identificación de un algoritmo seleccionado para encaminar dicho grupo;

- identificación de un algoritmo utilizado para encaminar dicho grupo; y

- si los identificadores de dicho algoritmo seleccionado y de dicho algoritmo utilizado para encaminar dicho grupo son diferentes, notificación de un problema potencial de optimización de encaminamiento de dicho grupo.

Según un modo de realización particular, el procedimiento realiza un bucle sobre sí mismo para identificar un problema potencial de encaminamiento de dicho grupo.

Según un modo de realización particular, el procedimiento es implementado en un administrador de grupo de tipo InfiniBand.

La invención hace referencia asimismo a un programa informático que comprende instrucciones adaptadas a la implementación de cada una de los etapas del procedimiento descrito anteriormente, cuando dicho programa es ejecutado en un ordenador, así como un medio de almacenamiento de información, amovible o no, que puede ser parcial o totalmente legible por un ordenador o un microprocesador, que comprende instrucciones de código de un programa informático para la ejecución de cada una de las etapas del procedimiento anterior.

Las ventajas proporcionadas por este programa informático y este medio de almacenamiento de información son similares a las mencionadas anteriormente.

Otras ventajas, objetivos y características de la presente invención resultan evidentes a partir de la descripción detallada que sigue, realizada a modo de ejemplo no limitativo, haciendo referencia a los dibujos adjuntos en los que:

- la figura 1 muestra un ejemplo de topología de un grupo;

- la figura 2 muestra un ejemplo de arquitectura de un nodo de un grupo;

- la figura 3, que comprende las figuras 3a y 3b, muestra un ejemplo de enlaces entre varios conmutadores cuando el encaminamiento se considera equilibrado y cuando no lo es, respectivamente;

- la figura 4 muestra algunas etapas de un ejemplo de algoritmo para estimar un valor representativo del equilibrado del encaminamiento de un grupo y, cuando corresponda, alertar a un administrador de un potencial desequilibrio; y

- la figura 5 muestra ciertas etapas de un ejemplo de algoritmo para alertar a un administrador de si existe un riesgo de encaminamiento que no permite el funcionamiento de un grupo en sus mejores condiciones.

Se ha observado que la calidad de encaminamiento de un grupo se puede estimar según la homogeneidad de la distribución de las rutas en los enlaces entre los diferentes elementos de este grupo. En otras palabras, una distribución uniforme de las rutas en los enlaces conduce a una distribución uniforme de los datos relativos a los intercambios de datos a través del grupo, permitiendo evitar la utilización excesiva de algunos enlaces y la infrautilización de otros enlaces, lo que conduce habitualmente a congestiones y al aumento en los retardos de latencia.

La distribución de las rutas a través de los enlaces es, por lo tanto, una indicación de calidad de encaminamiento y, de manera más general, una indicación de calidad de utilización de los elementos del grupo.

Por otra parte, se recuerda que el encaminamiento de datos en una red de tipo InfiniBand, en forma de paquetes de datos, se basa en un mecanismo de retransmisión local según una indicación de destino denominada LID (acrónimo de identificador local, *Local IDentifier*, en terminología anglosajona). De este modo, cada conmutador InfiniBand (que comprende varios puertos de entrada y varios puertos de salida) decide a qué puerto de salida se debe reenviar un paquete en función de la indicación de destino (LID) del paquete considerado

Para este propósito, cada conmutador utiliza una tabla de encaminamiento local llamada LFT (acrónimo de *Linear Forwarding Table*, en terminología anglosajona). Dichas tablas, construidas durante una etapa de encaminamiento por un módulo de encaminamiento y transmitidas a cada conmutador, establecen un enlace entre una indicación de destino (LID) de un paquete de datos y un puerto de salida del conmutador al que está asociada la tabla de encaminamiento considerada. En otras palabras, una etapa de encaminamiento local de un conmutador InfiniBand consiste en un conjunto de pares que comprenden una indicación de encaminamiento (LID) y una referencia de puerto de salida. Un ejemplo de dicha tabla de encaminamiento se proporciona en el anexo (Tabla 1).

Cuando un conmutador InfiniBand recibe un paquete de datos, se analiza el encabezamiento de este último para determinar una indicación de destino (LID). A continuación, se consulta la tabla de encaminamiento para determinar la referencia del puerto de salida asociado a esta indicación de destino. El paquete es transmitido a continuación a su destino a través del puerto de salida cuya referencia ha sido determinada a partir de la indicación de destino. De este modo, por ejemplo, si un paquete que tiene, en su encabezamiento, una indicación de destino

igual a 28, llega a un conmutador InfiniBand cuya tabla de encaminamiento corresponde a la tabla 1 que figura en el anexo, este paquete es retransmitido a través del puerto de salida que tiene la referencia 25.

Tal como se indicó anteriormente, el encaminamiento de un grupo InfiniBand es estático, lo que significa que las rutas son predecibles y no cambian hasta que el grupo se apaga o hasta que un dispositivo o bien un enlace sufre un fallo que conduce a un cambio de ruta del grupo. Habitualmente lo realiza un módulo de encaminamiento administrador InfiniBand, denominado *InfiniBand Subnet Manager*, que se encarga de calcular las tablas de encaminamiento de los conmutadores y transmitirlas a los mismos.

El administrador InfiniBand Manager propone, en general, varios algoritmos de encaminamiento, que presentan cada uno características particulares. La elección de un algoritmo en general la realiza un administrador según la topología del grupo. De este modo, por ejemplo, la elección del administrador se puede hacer en el algoritmo conocido bajo el nombre de FTree, que es particularmente eficiente para los grupos que tienen una topología de *fat-tree*. Si el administrador InfiniBand falla en el encaminamiento del grupo con el algoritmo seleccionado, se selecciona automáticamente un algoritmo menos eficiente, por ejemplo, el algoritmo conocido bajo el nombre de MinHop.

Cuando existen varios enlaces de comunicación entre dos conmutadores, el encaminamiento se considera equilibrado entre estos dos conmutadores si existe un mismo número de rutas por enlace. A la inversa, el encaminamiento no se considera equilibrado si no existe el mismo número de rutas por enlace o si el número de rutas por enlace varía por encima de un umbral predeterminado o se calcula dinámicamente.

La figura 3, que comprende las figuras 3a y 3b muestran un ejemplo de varios conmutadores cuando el encaminamiento se considera equilibrado y cuando no lo es, respectivamente.

Tal como se muestra en la figura 3a, el grupo 100 comprende en el presente documento un primer conjunto de nodos de referencias 105-11 a 105-1n, un segundo conjunto de nodos de referencias 105-p1 a 105-pm, así como cuatro conmutadores 110-1 a 110-4. Cada nodo del primer conjunto está conectado al conmutador 110-1 mediante un enlace específico. Asimismo, cada nodo del segundo conjunto está conectado al conmutador 110-2 mediante un enlace específico. El conmutador 110-1 está conectado al conmutador 110-3 mediante dos enlaces 115-1 y 115-2 y al conmutador 110-4 mediante dos enlaces 115-3 y 115-4. De manera simétrica, el conmutador 110-2 está conectado al conmutador 110-3 mediante dos enlaces 115-5 y 115-6 y al conmutador 110-4 mediante dos enlaces 115-7 y 115-8.

Tal como se indica en la figura 3a, cada uno de los enlaces 115-1 a 115-8 comprende 8 rutas. Por consiguiente, dado que el número de rutas por enlace es el mismo para todos los enlaces, el encaminamiento se considera equilibrado.

La figura 3b representa un grupo 100’ similar al grupo 100 mostrado en la figura 3a. En el presente documento, comprende dos conjuntos de nodos (105'-11 a 105'-1n y 105'-p1 a 105'-pm), así como cuatro conmutadores (110'-1 a 110'-4). Cada nodo de los conjuntos primero y segundo está conectado a los conmutadores 110'-1 y 110'-2, respectivamente, mediante un enlace específico. El conmutador 110'-1 está conectado a los conmutadores 110'-3 y 110'-4 mediante los dos enlaces 115'-1 y 115'-2 y los dos enlaces 115'-3 y 115'-4, respectivamente. De manera simétrica, el conmutador 110'-2 está conectado a los conmutadores 110'-3 y 110'-4 mediante los dos enlaces 115'-5 y 115'-6 y los dos enlaces 115'-7 y 115'-8, respectivamente.

Tal como se indica en la figura 3b, los enlaces 115'-1 y 115'-2 comprenden 4 y 12 rutas, respectivamente, mientras que cada uno de los enlaces 115'-3 a 115'-8 comprende 8 rutas. Dado que el número de rutas por enlace no es el mismo para todos los enlaces, el encaminamiento no se considera equilibrado.

Por lo tanto, la calidad de encaminamiento de un grupo se determina en este caso según la distribución de las rutas por puerto de conmutador, para cada conmutador, para todos los conmutadores de un grupo de conmutadores (por ejemplo, todos los conmutadores conectados directamente a uno o a varios elementos dados) o para todos los conmutadores del grupo.

Para estos propósitos, según un modo de realización particular, la tabla de encaminamiento de cada conmutador se analiza para calcular el número de rutas por puerto, para cada puerto del conmutador considerado. Este análisis puede ser realizado por el administrador InfiniBand.

Con fines ilustrativos, la tabla de encaminamiento que figura en el anexo (tabla 1) puede ser utilizada para calcular el número de rutas asignadas a cada puerto de salida del conmutador correspondiente. La tabla 2 que figura en el anexo indica el número de rutas para cada puerto. Tal como se indica, se asignan cinco rutas al puerto 21, se asignan tres rutas al puerto 19 y una sola ruta a los otros puertos.

Por lo tanto, conociendo la tabla de encaminamiento de todos los conmutadores del grupo, el administrador InfiniBand u otro módulo de supervisión puede calcular el número de rutas asignadas a cada puerto de salida de cada conmutador y deducir una información de equilibrado del encaminamiento del grupo.

La figura 4 muestra algunas etapas de un ejemplo de algoritmo para estimar un valor representativo del equilibrado del encaminamiento de un grupo y, en su caso, alertar a un administrador de un potencial desequilibrio. Este algoritmo puede ser implementado, por ejemplo, en un administrador InfiniBand.

Tal como se muestra, una primera etapa (etapa 400) tiene por objeto la selección de conmutadores Ci (variando i de 1 al número NbC de conmutadores seleccionados). La selección puede estar dirigida a un solo conmutador, a todos los conmutadores de un grupo (por ejemplo, todos los conmutadores conectados directamente a uno o a varios elementos dados) o a todos los conmutadores del grupo. La selección puede ser realizada de manera automática según criterios predeterminados, de manera manual por un administrador o de manera mixta (siendo propuesta una preselección, por ejemplo, a un administrador que puede validarla y, en su caso, modificarla).

En una etapa siguiente, se obtiene la tabla de encaminamiento (TRi) asociada a cada conmutador seleccionado (etapa 405). En este caso, se observa que, si el algoritmo es implementado en un administrador InfiniBand, la obtención de estas tablas se realiza de manera particularmente fácil, puesto que es el administrador InfiniBand el que las determina. Si el algoritmo se implementa en otro módulo, las tablas de encaminamiento se pueden obtener por medio de solicitudes de los conmutadores correspondientes, del módulo o de los módulos que las han creado o de cualquier otro dispositivo que las esté almacenando.

Para cada conmutador seleccionado, el número de rutas por puerto de salida NbRij (variando j desde 1 hasta el número NbPi de puertos de salida del conmutador Ci) se calcula a partir de la tabla de encaminamiento asociada con este conmutador (etapa 410). Tal como se indicó anteriormente, el número de rutas por enlace se puede determinar contando el número de indicaciones de destino (LID) diferentes asociadas con el mismo puerto de salida de un conmutador.

A continuación, se calcula el número promedio de rutas por puerto NbR para el conjunto de conmutadores seleccionados (etapa 415). Este número se calcula habitualmente según la relación siguiente:

$$NbR = \frac{\sum_{i=1}^{NbC} \sum_{j=1}^{NbPi} NbRij}{\sum_{i=1}^{NbC} NbPi}$$

En este caso, se observa que la granularidad del valor representativo del equilibrado del encaminamiento del grupo está determinada por el conjunto de conmutadores seleccionados.

Tal como se indica en la figura 4 mediante la flecha en línea de puntos, el algoritmo se puede repetir para varios conjuntos de conmutadores (y, por lo tanto, para cada conmutador considerado de manera aislada).

En la siguiente etapa, el número de rutas NbRij del puerto j de cada conmutador i se compara con el número promedio NbR de rutas por puerto (etapa 420).

Si la diferencia entre el número de rutas NbRij del puerto j del conmutador i y el número promedio NbR de rutas por puerto excede un umbral θ, se considera que el encaminamiento del grupo está desequilibrado. En este caso, se envía una notificación a un administrador (etapa 425), por ejemplo, en forma de un mensaje electrónico. La notificación puede incluir una indicación relativa al conmutador o conmutadores en cuestión y, si corresponde, al puerto o puertos de este último, para los cuales se ha detectado un problema potencial.

El umbral θ por encima del cual se considera que el encaminamiento del grupo está desequilibrado es, en teoría, próximo a cero. Sin embargo, en la práctica, el administrador lo determina preferentemente según la topología del grupo y su carga (relacionada, en particular, con el número y la naturaleza de las aplicaciones que se ejecutan).

En este caso, se observa que, si resulta que un problema de carga está esencialmente vinculado a una sobrecarga de un enlace (cuando el número de rutas NbRij del puerto j del conmutador i es mayor que el número promedio NbR de rutas por puerto), lo que puede provocar congestión y/o un aumento de los retardos de latencia, una infra carga de un enlace (cuando el número de rutas NbRij del puerto j del conmutador i es menor que el número promedio de NbR de rutas por puerto) indica que existe una sobrecarga en otra parte del grupo, pudiendo no ser detectada esta sobrecarga (debido a una distribución de la sobrecarga en varios enlaces).

Naturalmente, es posible no detectar más que las sobrecargas (NbRij – NbR > θ), incluso las infra cargas (NbR – NbRij > θ), incluso las sobrecargas y las infra cargas con diferentes valores (NbRij NbR > θ_1 o NbR – NbRij > θ_2).

Según otro modo de realización, el algoritmo de encaminamiento utilizado se tiene en cuenta para alertar a un administrador de un problema potencial de encaminamiento.

La figura 5 muestra ciertas etapas de un ejemplo de algoritmo para alertar a un administrador si existe un riesgo de encaminamiento que no permite el aprovechamiento de un grupo en sus mejores condiciones.

En particular, el algoritmo mostrado en la figura 5 permite alertar a un administrador si el algoritmo de encaminamiento utilizando no es el algoritmo de encaminamiento seleccionado (o validado) por un administrador y/o si existe un potencial desequilibrio de encaminamiento en el grupo. Este algoritmo puede ser implementado especialmente en un administrador InfiniBand.

Tal como se muestra, una primera etapa tiene por objeto en el presente documento identificar el algoritmo de encaminamiento seleccionado o validado por un administrador (ARS) para encaminar el grupo (etapa 500). Se puede obtener un identificador de este algoritmo seleccionado o validado a partir de los parámetros de configuración del módulo de administración del grupo, por ejemplo, un administrador InfiniBand. Dicho identificador puede ser, por ejemplo, el identificador «FTREE» que designa el algoritmo de encaminamiento conocido como FTree.

A continuación, antes o en paralelo, se identifica el algoritmo de encaminamiento realmente utilizado (ARU) para encaminar el grupo (etapa 505). De nuevo, un identificador de este algoritmo utilizado se puede obtener a partir de los parámetros de configuración del módulo de administración del grupo, por ejemplo, un administrador InfiniBand. Dicho identificador puede ser, por ejemplo, el identificador «MINHOP» que designa el algoritmo de encaminamiento conocido bajo el nombre de MINHOP.

El identificador del algoritmo de encaminamiento utilizado se compara a continuación con el algoritmo de encaminamiento seleccionado (etapa 510). Si estos identificadores son diferentes, se envía una notificación a un administrador (etapa 515), por ejemplo, en forma de un mensaje electrónico, que indica que el algoritmo de encaminamiento utilizado no es el algoritmo de encaminamiento seleccionado y que menciona, además, preferentemente, una referencia del algoritmo de encaminamiento utilizado (y, de manera opcional, una referencia del algoritmo de encaminamiento seleccionado).

Por el contrario, si el identificador del algoritmo de encaminamiento utilizado es el mismo que el del algoritmo de encaminamiento seleccionado (etapa 510), se realizan etapas para estimar un valor representativo del equilibrado del encaminamiento de un grupo y, en su caso, alertar a un administrador de un posible desequilibrio. Estas etapas, de referencias 400' a 420', pueden ser similares a las etapas 400 a 420 descritas haciendo referencia a la figura 4.

Tal como se describió anteriormente, la etapa 400' tiene por objeto la selección automática, semiautomática o manual de uno o varios conmutadores Ci (variando i de 1 al número NbC de conmutadores seleccionados). La tabla de encaminamiento (TRi) asociada con cada conmutador seleccionado se obtiene durante la siguiente etapa (etapa 405'), observándose en este caso de nuevo que, si el algoritmo se implementa en un administrador InfiniBand, la obtención de estas tablas resulta particularmente fácil, puesto que es el administrador InfiniBand quien las determina. Si el algoritmo se implementa en otro módulo, las tablas de encaminamiento se pueden obtener por medio de solicitudes de los conmutadores correspondientes, del módulo o de los módulos que las hayan creado o de cualquier otro dispositivo que las esté almacenando.

Para cada conmutador seleccionado, el número de rutas por puerto NbRij (variando j de 1 al número NbPi de puertos del conmutador Ci) se calcula en este caso a partir de la tabla de encaminamiento asociada con este conmutador (etapa 410'). Tal como se indicó anteriormente, el número de rutas por enlace se puede determinar contando el número de indicaciones de destino (LID) diferentes asociadas con el mismo puerto de un conmutador.

El número promedio de rutas por puerto NbR se calcula para el conjunto de conmutadores seleccionados (etapa 415'). De nueva, este número se calcula habitualmente según la relación siguiente:

$$NbR = \frac{\sum_{i=1}^{NbC} \sum_{j=1}^{NbPi} NbRij}{\sum_{i=1}^{NbC} NbPi}$$

estando determinada la granularidad del valor representativo del equilibrado del encaminamiento del grupo por el conjunto de conmutadores seleccionados.

En una etapa siguiente, el número de rutas NbRij para cada puerto j de cada conmutador i se compara con el número promedio NbR de rutas por puerto (etapa 420')

Si la diferencia entre el número de rutas NbRij del puerto j del conmutador i y el número promedio de rutas por puerto de NbR supera un umbral θ, se considera que el encaminamiento del grupo está desequilibrado. En este caso, se envía una notificación a un administrador (etapa 515), por ejemplo, en forma de mensaje electrónico. De nuevo, la notificación puede comprender una indicación relativa al conmutador o los conmutadores y, si corresponde, al puerto o puertos de este o estos últimos, para el cual se ha detectado un problema potencial.

En este caso, incluso, es posible no detectar más que las sobrecargas (NbRij – NbR > θ), las infra cargas (NbR – NbRij > θ), o las sobrecargas e infra cargas con diferentes umbrales (NbRij – NbR > θ_1 o NbR – NbRij > θ_2)

Tal como se muestra en la figura 5, el algoritmo realiza un bucle preferentemente sobre sí mismo hasta que se termina. De este modo, permite alertar a un administrador si existe un riesgo de encaminamiento que no permite el aprovechamiento de un grupo en sus mejores condiciones durante la utilización de este último.

Cabe señalar que los algoritmos descritos haciendo referencia a las figuras 4 y 5 pueden ser implementados, por ejemplo, en un dispositivo similar al descrito haciendo referencia a la figura 2, en forma de un programa informático.

Naturalmente, para satisfacer necesidades específicas, un experto en el campo de la invención puede aplicar modificaciones en la descripción anterior.

ANEXO

Tabla 1

LID	Referencia de puerto
2	19
4	21
5	19
7	21
8	21
10	22
11	23
12	10
15	19
17	21
18	7
20	21
25	9
28	25
31	32

Tabla 2

Referencia de puerto	Nb rutas
7	1
9	1
10	1
19	3
21	5
22	1
23	1
25	1
32	1

REIVINDICACIONES

1. Procedimiento para ordenador de supervisión de al menos un parámetro de encaminamiento de un grupo que comprende una serie de nodos y una serie de conmutadores, conectando varios enlaces de comunicación estática nodos y conmutadores de dicha serie de nodos y de conmutadores, comprendiendo cada conmutador una serie de puertos de salida, estando caracterizado este procedimiento por que comprende las etapas siguientes,

- selección (400) de al menos un conmutador de dicha serie de conmutadores;

- cálculo (410) de un número de rutas por puerto para cada puerto de cada conmutador seleccionado, estando definidas rutas durante una etapa de encaminamiento para conectar cada uno de los nodos a otro;

- cálculo (415) de un número promedio de rutas por puerto para al menos un conmutador seleccionado;

- comparación (420) de cada número de rutas por puerto calculado con el número promedio de rutas por puerto calculado; y

- en respuesta a dicha comparación, notificación (425) de un potencial desequilibrio de encaminamiento de dicho grupo.

2. Procedimiento según la reivindicación 1, según el que cada conmutador comprende una tabla de encaminamiento para identificar un puerto según una característica de un paquete de datos recibido, comprendiendo además el procedimiento una etapa de obtención (405) de una tabla de encaminamiento para cada conmutador seleccionado, estando basada dicha etapa de cálculo de un número de rutas por puerto para cada puerto de cada conmutador seleccionado en una tabla de encaminamiento obtenida para cada conmutador seleccionado.

3. Procedimiento según la reivindicación 1 o la reivindicación 2, según el que dicha etapa de comparación de cada número de rutas por puerto con el número promedio de rutas por puerto comprende una etapa de cálculo de la diferencia entre cada número de rutas por puerto y el número promedio de rutas por puerto, y una etapa de comparación de la diferencia con al menos un umbral.

4. Procedimiento según la reivindicación 3, según el que dicha etapa de cálculo de la diferencia entre cada número de rutas por puerto con el número promedio de rutas por puerto es una etapa de cálculo del valor absoluto de la diferencia de cada número de rutas por puerto y el número medio de rutas por puerto.

5. Procedimiento según la reivindicación 3, según el que la diferencia entre cada número de rutas por puerto y el número promedio de rutas por puerto se considera cero si es negativa.

6. Procedimiento según la reivindicación 3, según el que dicha comparación de la diferencia con al menos un umbral comprende una etapa de selección de un primer o un segundo umbral dependiendo de si la diferencia es positiva o negativa.

7. Procedimiento según una cualquiera de las reivindicaciones 1 a 6, que comprende, además, las etapas siguientes:

- identificación (500) de un algoritmo seleccionado para encaminar dicho grupo;

- identificación (505) de un algoritmo utilizado para encaminar dicho grupo; y

- si los identificadores de dicho algoritmo seleccionado y dicho algoritmo utilizado para encaminar dicho grupo son diferentes, notificación (515) de un problema potencial de optimización de encaminamiento de dicho grupo.

8. Procedimiento de una cualquiera de las reivindicaciones 1 a 7, según el que el procedimiento realiza un bucle sobre sí mismo para identificar un problema potencial de encaminamiento de dicho grupo.

9. Procedimiento según una cualquiera de las reivindicaciones 1 a 8, estando implementado el procedimiento en un administrador de grupo de tipo InfiniBand.

10. Programa informático que comprende instrucciones adaptadas a la implementación de cada una de las etapas del procedimiento según una cualquiera de las reivindicaciones anteriores cuando dicho programa es ejecutado en un ordenador.

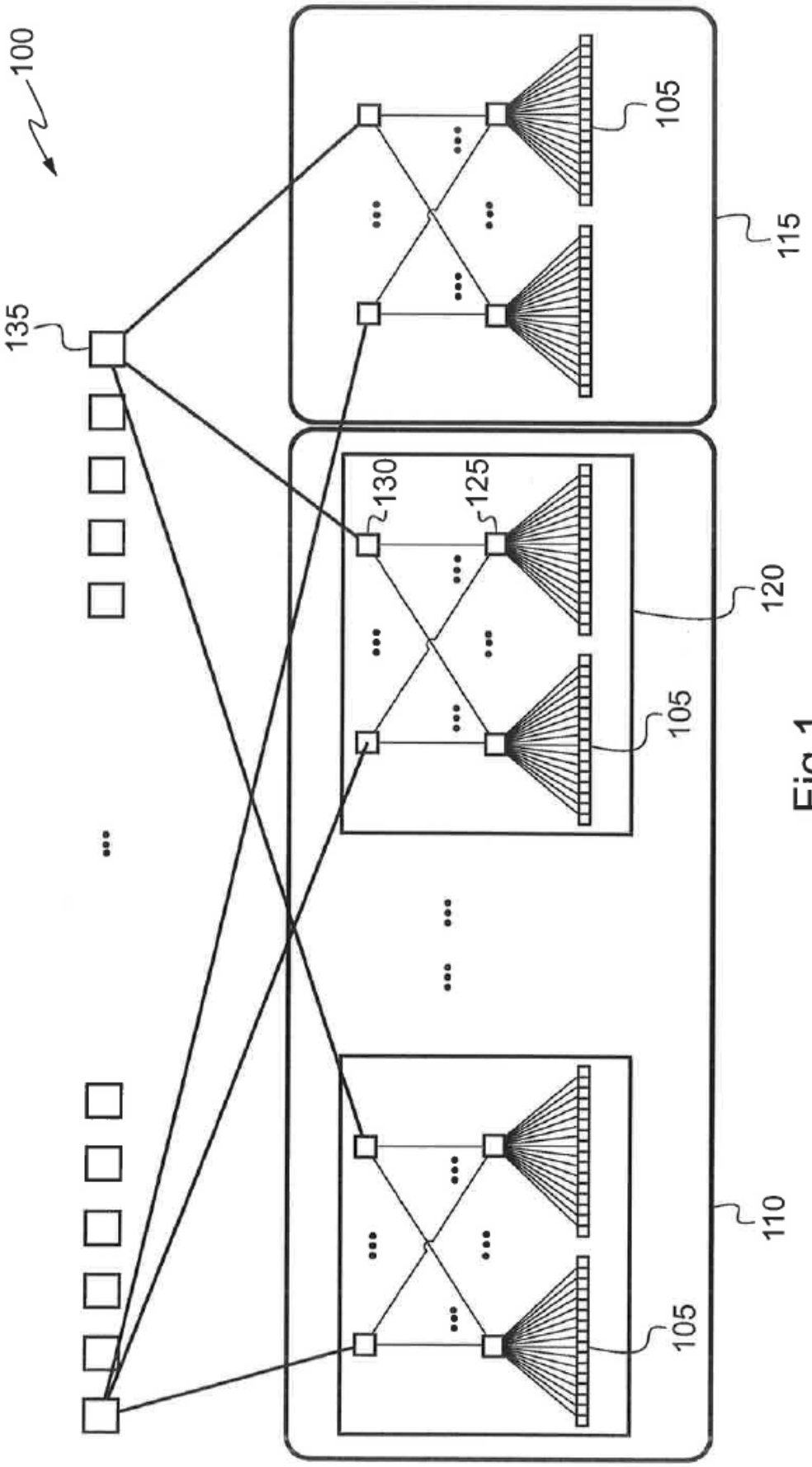

Fig.1

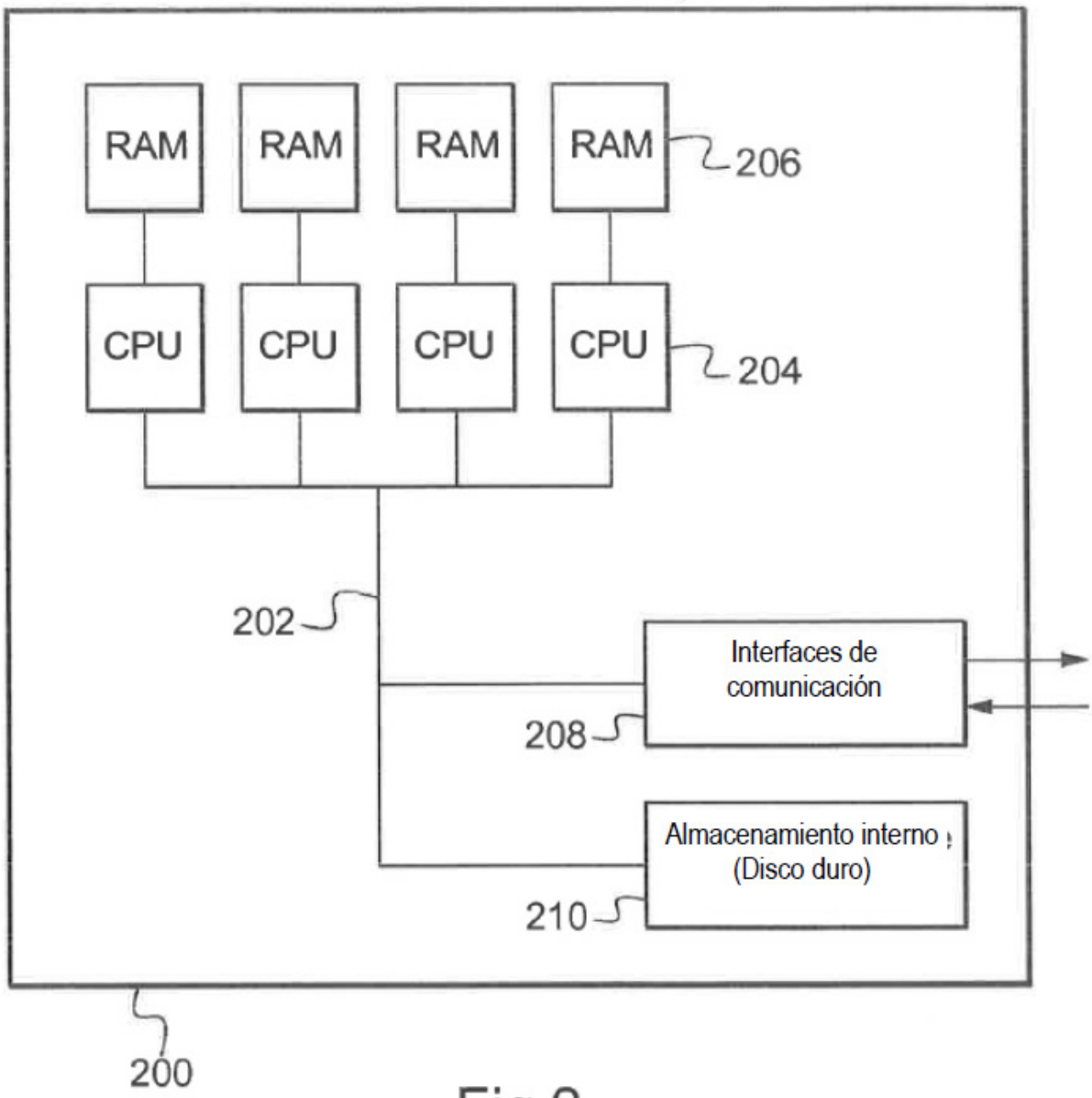
RAM
RAM
RAM
RAM
206
CPU
CPU
CPU
CPU
204
202
Interfaces de comunicación
208
Almacenamiento interno (Disco duro)
210
200

Fig.2

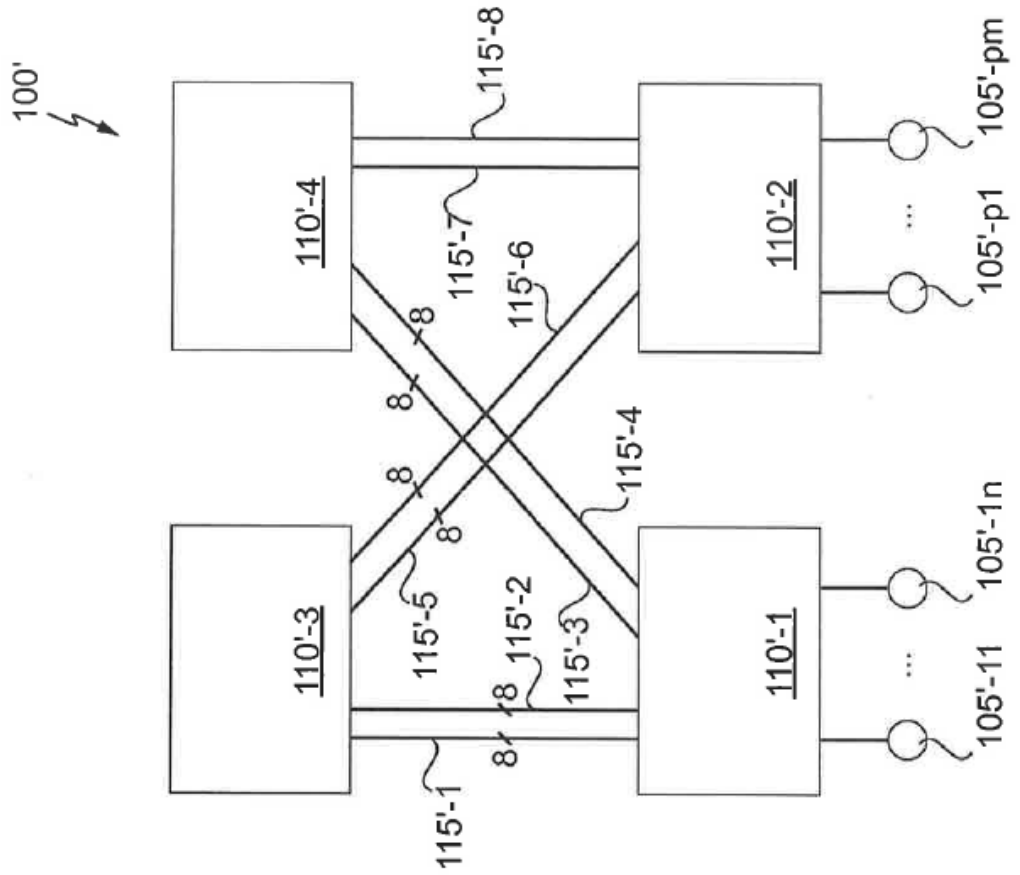
100'
110'-4
110'-2
110'-3
110'-1
115'-8
115'-7
115'-6
115'-4
115'-5
115'-2
115'-3
115'-1
105'-pm
105'-p1
105'-1n
105'-11
8

Fig. 3b

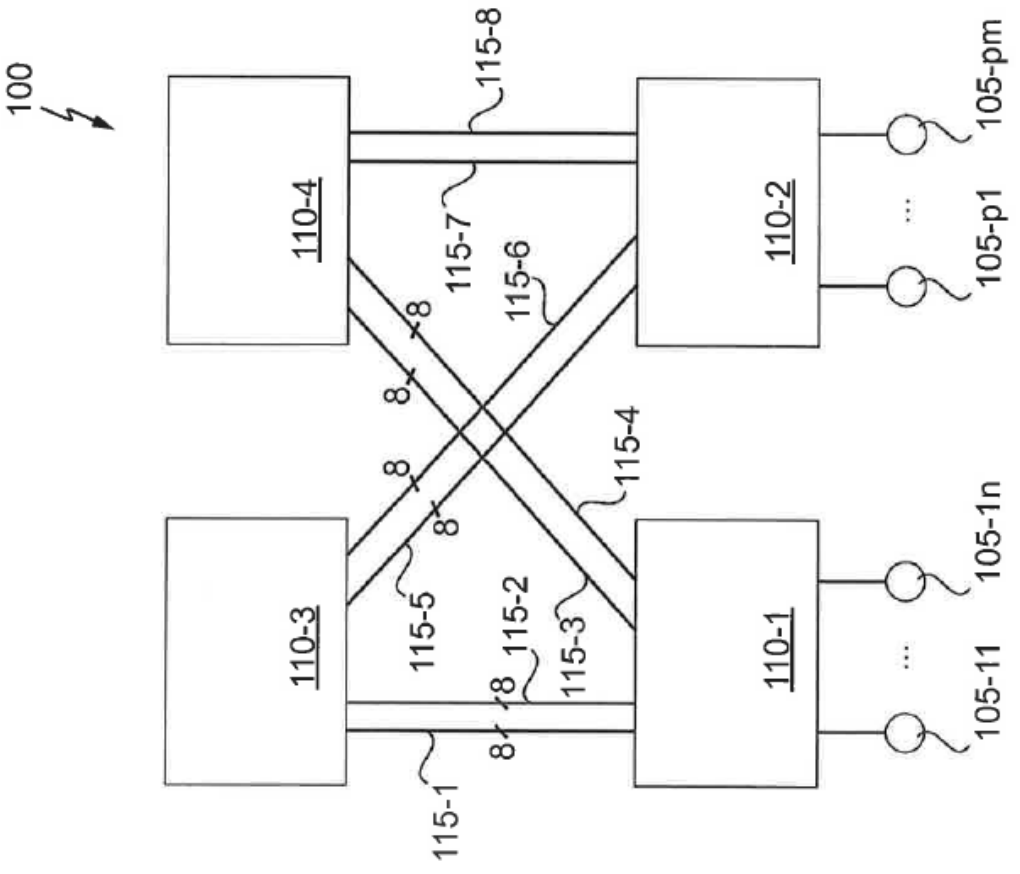
100
110-4
110-2
110-3
110-1
115-8
115-7
115-6
115-4
115-5
115-2
115-3
115-1
105-pm
105-p1
105-1n
105-11
8

Fig. 3a

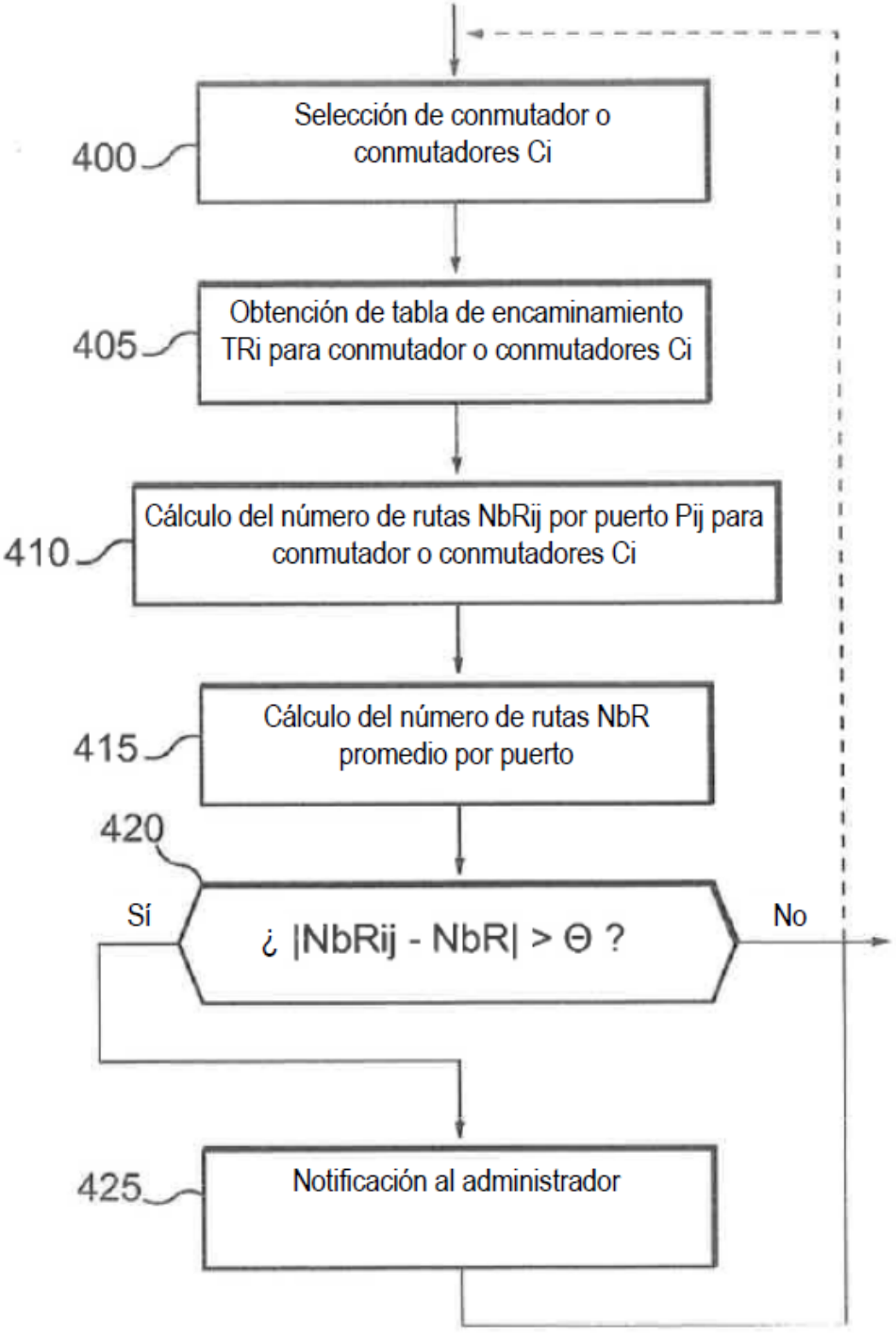
400
Selección de conmutador o conmutadores Ci
405
Obtención de tabla de encaminamiento TRi para conmutador o conmutadores Ci
410
Cálculo del número de rutas NbRij por puerto Pij para conmutador o conmutadores Ci
415
Cálculo del número de rutas NbR promedio por puerto
420
Sí
¿ |NbRij - NbR| > Θ ?
No
425
Notificación al administrador

Fig. 4

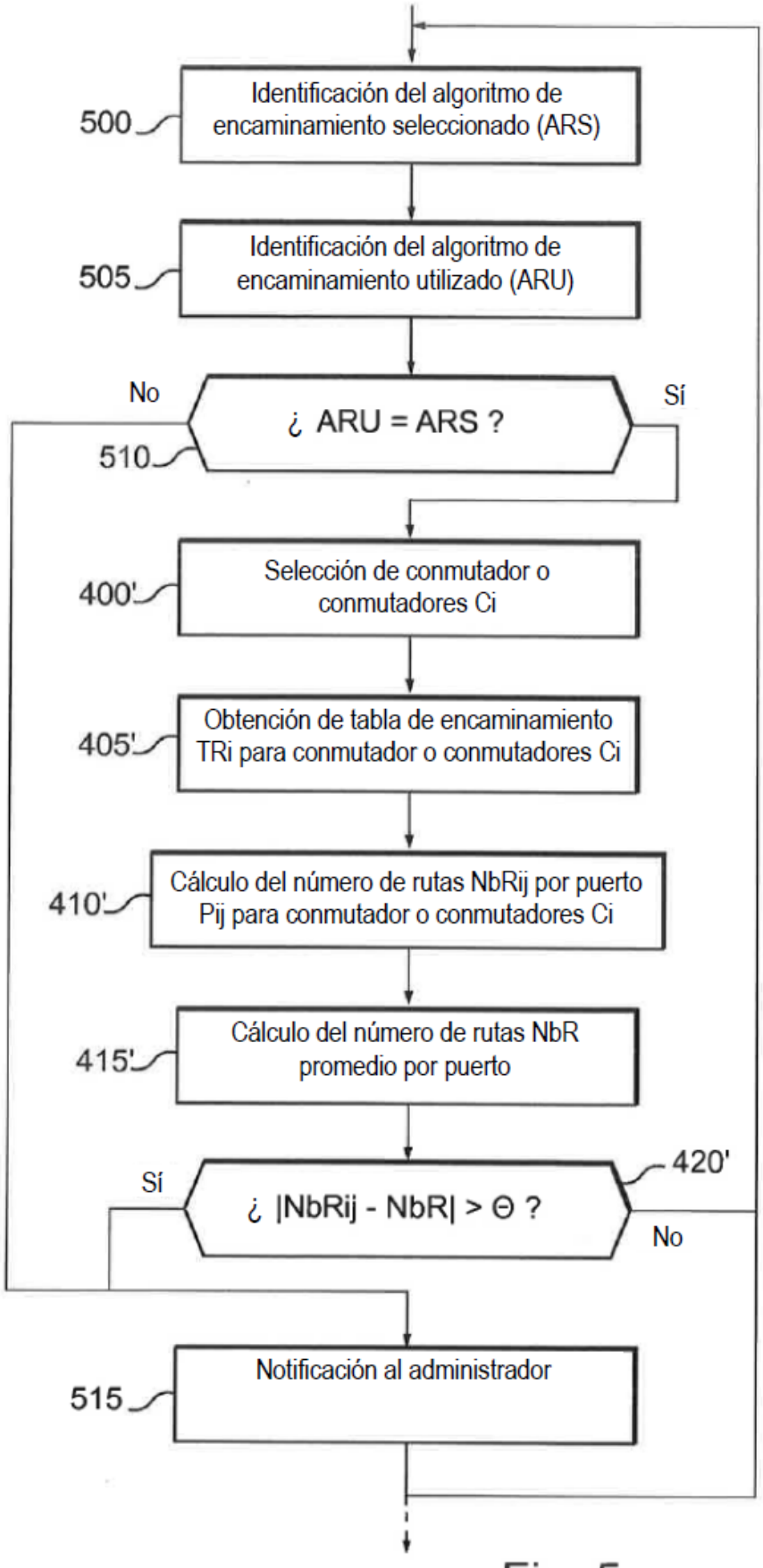
500
Identificación del algoritmo de encaminamiento seleccionado (ARS)
505
Identificación del algoritmo de encaminamiento utilizado (ARU)
No
¿ ARU = ARS ?
Sí
510
400'
Selección de conmutador o conmutadores Ci
405'
Obtención de tabla de encaminamiento TRi para conmutador o conmutadores Ci
410'
Cálculo del número de rutas NbRij por puerto Pij para conmutador o conmutadores Ci
415'
Cálculo del número de rutas NbR promedio por puerto
Sí
¿ |NbRij - NbR| > Θ ?
420'
No
515
Notificación al administrador

Fig. 5