(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 719 754**

(51) Int. CI.:

H04L 12/721 (2013.01)
H04L 12/46 (2006.01)
H04L 12/26 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **03.11.2015 PCT/FR2015/052970**

(87) Fecha y número de publicación internacional: **12.05.2016 WO16071628**

(96) Fecha de presentación y número de la solicitud europea: **03.11.2015 E 15804879 (3)**

(97) Fecha y número de publicación de la concesión europea: **09.01.2019 EP 3216175**

(54) Título: **Procedimiento de supervisión y de control trasladados de un clúster que utiliza una red de comunicación de tipo InfiniBand y programa de ordenador que pone en práctica este procedimiento**

(30) Prioridad:

06.11.2014 FR 1460723

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
12.07.2019

(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

DESCRIPCIÓN

Procedimiento de supervisión y de control trasladados de un clúster que utiliza una red de comunicación de tipo InfiniBand y programa de ordenador que pone en práctica este procedimiento

La presente invención concierne a la supervisión y al control de los clústeres (agrupaciones), en particular de clústeres que utilizan una red de comunicación de tipo InfiniBand y, más en particular, a un procedimiento de supervisión y de control trasladados de un clúster que utiliza una red de comunicación de tipo InfiniBand y a un programa de ordenador que pone en práctica tal procedimiento.

El cálculo de altas prestaciones, también denominado HPC (siglas de *High Performance Computing* en la terminología anglosajona) se desarrolla para la investigación universitaria al igual que para la industria, especialmente en campos técnicos tales como la aeronáutica, la energía, la climatología y las ciencias de la vida. La modelización y la simulación permiten en particular abaratar los costes de desarrollo, acelerar la explotación comercial de productos innovadores, más fiables y menos consumidores de energía. Para los investigadores, el cálculo de altas prestaciones se ha convertido en un medio de investigación imprescindible.

Estos cálculos generalmente se ponen en práctica en sistemas de procesamiento de datos denominados clústeres. Un clúster comprende típicamente un conjunto de nodos interconectados. Ciertos nodos se utilizan para efectuar tareas de cálculo (nodos de cálculo), otros, para almacenar datos (nodos de almacenamiento) y, en general, otro gestiona el clúster (nodo de administración). Cada nodo es, por ejemplo, un servidor que pone en práctica un sistema operativo tal como Linux (Linux es una marca). La conexión entre los nodos se realiza, por ejemplo, con el concurso de enlaces de comunicación Ethernet o InfiniBand (Ethernet e InfiniBand son marcas).

La figura 1 ilustra esquemáticamente un ejemplo de una topología 100 de un clúster, de tipo *fat-tree*. Este último comprende un conjunto de nodos referenciados genéricamente con 105. En este punto, los nodos pertenecientes al conjunto 110 son nodos de cálculo, en tanto que los nodos del conjunto 115 son nodos de servicio (nodos de almacenamiento y nodo de administración). Los nodos de cálculo pueden estar agrupados en subconjuntos 120 denominados islotes de cálculo, denominándose el conjunto 115 islote de servicio.

Los nodos están unidos unos a otros mediante conmutadores (denominados *switch* en la terminología anglosajona), por ejemplo de manera jerárquica. En el ejemplo ilustrado en la figura 1, los nodos están conectados a conmutadores de primer nivel 125 que a su vez están unidos a conmutadores de segundo nivel 130 los cuales, por su parte, están unidos a conmutadores de tercer nivel 135.

Como se ilustra en la figura 2, cada nodo comprende generalmente uno o varios microprocesadores, memorias locales, así como una interfaz de comunicación. Más concretamente, en este punto, el nodo 200 incluye un bus de comunicación 202 al que están unidos:

- unidades centrales de proceso o microprocesadores 204 (o CPU, siglas de *Central Processing Unit* en la terminología anglosajona);
- componentes de memoria de acceso directo 206 (RAM, acrónimo de *Random Access Memory* en la terminología anglosajona) que incluyen registros adaptados para grabar variables y parámetros creados y modificados en el curso de la ejecución de programas (como se ilustra, cada componente de memoria de acceso directo puede estar asociado a un microprocesador); e
- interfaces de comunicación 208 adaptadas para transmitir y recibir datos.

En este punto, el nodo 200 dispone además de medios de almacenamiento interno 212, tales como discos duros, con posibilidad de incluir especialmente el código máquina ejecutable de programas.

El bus de comunicación permite la comunicación y la interoperabilidad entre los diferentes elementos incluidos en el nodo 200 o unidos a él. Los microprocesadores 204 gobiernan y dirigen la ejecución de las instrucciones o porciones de código de soporte lógico del o los programas. En el encendido, el o los programas que están almacenados en una memoria no volátil, por ejemplo un disco duro, se transfieren a la memoria de acceso directo 206.

Se observa en este punto que las prestaciones de un clúster están directamente ligadas a la elección de las rutas que permiten la transferencia de datos entre los nodos, establecidas a través de enlaces de comunicación. Con carácter general, en la configuración de equipo físico de un clúster, se establecen entre los nodos y los conmutadores enlaces de comunicación físicos, determinándose las propias rutas de comunicación en una fase de inicialización a partir de una definición de las conexiones que se tienen que establecer entre los nodos. Según la tecnología de comunicación puesta en práctica, la configuración de las rutas puede ser estática o dinámica.

A título de ilustración, la tecnología InfiniBand permite, dentro de un clúster, una configuración estática de las rutas. Esta configuración utiliza tablas estáticas de encaminamiento (o LFT, siglas de *Linear Forwarding Table* en la terminología anglosajona) en cada conmutador. Cuando se pone en práctica esta tecnología, se puede utilizar un

algoritmo de encaminamiento tal como los algoritmos conocidos con los nombres de FTree, MINHOP, UPDN y LASH.

La elección del algoritmo que se deba utilizar la efectúa típicamente un administrador según, en especial, la topología del clúster. Puede tratarse, por ejemplo, del algoritmo FTree. Sin embargo, si el algoritmo elegido no permite efectuar el encaminamiento, el gestor del clúster (típicamente a cargo del encaminamiento) en general elige automáticamente otro algoritmo, por ejemplo el algoritmo MINHOP (generalmente menos eficiente que el escogido en un principio).

A título de ilustración y de manera simplificada, el algoritmo FTree determina rutas de manera tal que éstas estén repartidas en lo posible a través de los enlaces de comunicación existentes. Con estos fines, en el encaminamiento de una red de comunicación conectada por completo según una arquitectura de tipo *fat-tree*, se considera que cada nodo de la red tiene una importancia idéntica. De este modo, cuando se establece una ruta entre dos nodos de un mismo enlace, el número de rutas que utilizan este enlace, llamado la carga del enlace, aumenta en uno. El algoritmo de encaminamiento, cuando trata de establecer una nueva ruta y se presentan varias posibilidades, compara los niveles de carga asociados a los enlaces en los que se basan estas posibilidades y elige aquella cuyos enlaces tienen el menor nivel de carga.

En el curso de la utilización del clúster, si un enlace o un elemento tal como un nodo o un conmutador acusa un fallo, se efectúa un nuevo encaminamiento.

Al tener la calidad de encaminamiento una influencia directa sobre las prestaciones de un clúster, existe la necesidad de supervisar una configuración de encaminamiento en un clúster que comprende enlaces de comunicación estáticos, y de, en su caso, avisar a un administrador de un potencial problema de encaminamiento.

Como anteriormente se ha observado, la supervisión y el control de un clúster, en particular de los conmutadores, son realizadas típicamente de manera centralizada por un nodo especializado, denominado nodo de administración. Adicionalmente, este nodo de administración ejecuta servicios, por ejemplo servicios de gestión de bases de datos y servicios de gestión de dispositivos específicos (por ejemplo, dispositivos que no son de tipo InfiniBand en un clúster InfiniBand).

La supervisión y el control de un clúster de tipo InfiniBand se realizan con el concurso de paquetes de datos particulares, denominados MAD (acrónimo de *MAnagement Datagrams* en la terminología anglosajona). De este modo, un dispositivo de supervisión y de control direcciona paquetes de tipo MAD a equipos de tipo InfiniBand de un clúster, típicamente conmutadores o adaptadores, los cuales, como respuesta, devuelven paquetes de respuesta de tipo MAD al dispositivo emisor de supervisión y de control.

Un ejemplo del estado de la técnica anterior se encuentra en el documento US 2013/315243.

Sin embargo, se ha observado que puede ser útil, especialmente durante una fase de configuración de un clúster o cuando se producen problemas, ofrecer, además del nodo de administración, medios de supervisión y/o de control de un clúster a distancia o trasladados.

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

Así, la invención tiene por objeto un procedimiento de supervisión y de control trasladados de un clúster que comprende una pluralidad de nodos conectados a una red de comunicación de un primer tipo, comprendiendo un nodo de dicha pluralidad de nodos, denominado nodo retransmisor, una primera interfaz de red conforme a dicho primer tipo de red de comunicación y una segunda interfaz de red conforme a un segundo tipo de red de comunicación, siendo diferenciados dichos primer y segundo tipos, poniéndose en práctica ciertas etapas de dicho procedimiento de supervisión y de control en un ordenador remoto unido a dicho nodo retransmisor mediante una red de comunicación de dicho segundo tipo, comprendiendo el procedimiento las siguientes etapas:

- recepción de al menos un paquete de datos a través de dicha primera interfaz de red;
- encapsulado de dicho al menos un paquete de datos recibido en al menos una trama de datos conforme a un protocolo de dicha red de comunicación de dicho segundo tipo;
- transmisión de dicha al menos una trama de datos a dicho ordenador remoto a través de dicha segunda interfaz de red,

poniéndose en práctica dichas etapas de recepción de al menos un paquete de datos, de encapsulado de dicho al menos un paquete de datos y de transmisión de dicha al menos una trama de datos en dicho nodo retransmisor.

Así, el procedimiento según la invención permite actuar a distancia. Además, ofrece la posibilidad de efectuar un análisis fino de eventos en un clúster. Así, por ejemplo, un simple ordenador portátil unido a un clúster de tipo InfiniBand a través de un enlace Ethernet permite supervisar y gestionar al menos ciertos parámetros de un clúster.

De acuerdo con una forma particular de realización, el procedimiento comprende además las siguientes etapas, puestas en práctica en dicho nodo retransmisor,

- recepción, a través de dicha segunda interfaz de red, de dicho ordenador remoto, de al menos una trama de datos que comprende al menos un paquete de datos;
- desencapsulado de dicha al menos una trama de datos recibida para recuperar dicho al menos un paquete de datos contenido en dicha al menos una trama de datos recibida; y
- transmisión, a través de dicha primera interfaz de red, de dicho al menos un paquete de datos recuperado.

De acuerdo con una forma particular de realización, el procedimiento comprende además las siguientes etapas, puestas en práctica en dicho ordenador remoto,

- recepción, de dicho nodo retransmisor, de al menos una trama de datos que comprende al menos un paquete de datos;
- desencapsulado de dicha al menos una trama de datos recibida de dicho nodo retransmisor para recuperar dicho al menos un paquete de datos contenido en dicha trama de datos recibida de dicho nodo retransmisor; y
- procesamiento de dicho al menos un paquete de datos recuperado contenido en dicha trama de datos recibida de dicho nodo retransmisor, poniéndose en práctica dichas etapas de recepción de dicha al menos una trama de datos, dicha etapa de procesamiento de dicho al menos un paquete de datos recuperado en un módulo de supervisión y de control configurado para su puesta en práctica en un nodo de dicho clúster.

De acuerdo con una forma particular de realización, el procedimiento comprende además las siguientes etapas, puestas en práctica en dicho ordenador remoto,

- recepción, de un módulo de supervisión y de control configurado para su puesta en práctica en un nodo de dicho clúster, de al menos un paquete de datos;
- encapsulado de dicho al menos un paquete de datos recibido de dicho módulo de supervisión y de control configurado para su puesta en práctica en un nodo de dicho clúster, en al menos una trama de datos conforme a un protocolo de dicha red de comunicación de dicho segundo tipo;
- transmisión, a dicho nodo retransmisor, de dicha al menos una trama de datos que comprende dicho al menos un paquete de datos recibido de dicho módulo de supervisión y de control configurado para su puesta en práctica en un nodo de dicho clúster.

De acuerdo con una forma particular de realización, el procedimiento comprende además una etapa previa de memorización, en dicho nodo retransmisor, de una dirección de dicho ordenador remoto, siendo utilizada dicha dirección de dicho ordenador remoto memorizada para encapsular al menos un paquete de datos que ha de transmitirse a dicho ordenador remoto en forma de al menos una trama de datos.

De acuerdo con una forma particular de realización, el procedimiento comprende además una etapa de memorización, en dicho ordenador remoto, de una dirección de dicho nodo retransmisor, siendo utilizada dicha dirección de dicho nodo retransmisor memorizada para encapsular al menos un paquete de datos que ha de transmitirse a dicho nodo retransmisor en forma de al menos una trama de datos.

De acuerdo con una forma particular de realización, al menos un paquete de datos transmitido de dicho nodo retransmisor a dicho ordenador remoto es un paquete de datos de tipo particular que comprende información relativa a dicho clúster.

De acuerdo con una forma particular de realización, dicho primer tipo de red de comunicación es el tipo InfiniBand y según la cual dicho segundo tipo de red de comunicación es el tipo Ethernet.

La invención tiene asimismo por objeto un programa de ordenador que comprende instrucciones adaptadas para la puesta en práctica de cada una de las etapas del procedimiento anteriormente descrito cuando dicho programa es ejecutado en un ordenador, así como un sistema que comprende al menos un nodo retransmisor y al menos un ordenador remoto que comprende medios configurados para poner en práctica cada una de las etapas del procedimiento anteriormente descrito.

Las ventajas que brindan este programa de ordenador y este sistema son similares a las apuntadas anteriormente.

Otras ventajas, finalidades y características de la presente invención se desprenden de la descripción detallada que sigue, llevada a cabo a título de ejemplo no limitativo, con relación a los dibujos que se acompañan, en los cuales:

La figura 1 ilustra un ejemplo de topología de un clúster;
la figura 2 ilustra un ejemplo de arquitectura de un nodo de un clúster;

la figura 3 ilustra un ejemplo de puesta en práctica de la invención según una forma particular de realización;
la figura 4, que comprende las figuras 4a y 4b, ilustra esquemáticamente etapas puestas en práctica en un nodo de un clúster también unido a una red de comunicación de un tipo diferenciado de la propia del clúster, para transmitir datos de supervisión a un dispositivo remoto (figura 4a) y recibir datos de control de este dispositivo remoto (figura 4b); y
la figura 5, que comprende las figuras 5a y 5b, ilustra esquemáticamente etapas puestas en práctica en un ordenador remoto, unido a un nodo retransmisor de un clúster, para recibir datos de supervisión de este nodo retransmisor y procesarlos (figura 5a), así como para generar y transmitir datos de control a este nodo retransmisor (figura 5b).

Con carácter general, la invención está encaminada, de acuerdo con una forma particular de realización, a establecer una pasarela entre la red de comunicación de un clúster, por ejemplo una red de comunicación de tipo InfiniBand, y una red de comunicación de tipo dirigida al público en general, por ejemplo Ethernet.

La figura 3 ilustra un ejemplo de puesta en práctica de la invención según una forma particular de realización.

En este punto, el entorno 300 en el que se pone en práctica comprende un primer conjunto de dispositivos 305 conectados a una misma red de comunicación de un primer tipo, por ejemplo de tipo InfiniBand, y un segundo conjunto de dispositivos 310 conectados a una red de comunicación de un segundo tipo, diferenciado del primero, por ejemplo de tipo Ethernet.

Como se ilustra, el dispositivo 315 está unido a la red de comunicación de primer tipo y a la red de comunicación de segundo tipo (por ejemplo, InfiniBand y Ethernet) a través de sus interfaces de red 320-1 y 320-2 y a través de los conmutadores 325-1 y 325-2, respectivamente.

Este dispositivo permite establecer una pasarela entre las dos redes de comunicación. Con estos fines, comprende un módulo retransmisor 330, típicamente un módulo de lógica.

Así, el dispositivo 315 puede intercambiar datos con dispositivos 335-1 a 335-n utilizando la red de comunicación de primer tipo (por ejemplo, InfiniBand) y con un dispositivo 340 utilizando la red de comunicación de segundo tipo (por ejemplo, Ethernet).

El dispositivo 315, denominado nodo retransmisor más adelante en la descripción, es típicamente un nodo de administración.

Un módulo retransmisor 330, puesto en práctica en este servidor, permite transmitir datos recibidos por el nodo retransmisor 315 por una de las interfaces de red a la otra interfaz de red, y recíprocamente.

Así, el módulo retransmisor 330 permite transmitir datos, en forma de paquetes (por ejemplo, paquetes de tipo MAD), de un dispositivo unido a la red de comunicación de primer tipo (por ejemplo, InfiniBand) al dispositivo 340 del cual con anterioridad se ha obtenido, o se obtiene dinámicamente un identificador, por ejemplo una dirección IP. Los paquetes de datos recibidos se transmiten, de acuerdo con una forma particular de realización, en forma de tramas, a consecuencia de una etapa de encapsulado.

Recíprocamente, el módulo retransmisor 330 permite transmitir datos recibidos por el nodo retransmisor 315, en forma de tramas, del dispositivo 340 hacia un dispositivo unido a la red de comunicación de primer tipo (por ejemplo, InfiniBand). En las tramas recibidas (típicamente en paquetes de datos transmitidos en las tramas) se transmite preferentemente un identificador del o los dispositivos a los cuales se deben transmitir, en forma de paquetes (por ejemplo, paquetes de tipo MAD), datos recibidos.

De acuerdo con una forma particular de realización, las tramas recibidas son desencapsuladas para recuperar paquetes que, encapsulados con anterioridad, han de transmitirse en la red de comunicación de primer tipo, comprendiendo estos paquetes de datos unos identificadores del o los destinatarios.

El dispositivo de supervisión y de control 340 es típicamente un ordenador personal, por ejemplo de tipo PC (siglas de *Personal Computer* en terminología anglosajona) portátil. Comprende una interfaz de red 345, un módulo retransmisor 350 y un módulo de supervisión y de control 355. Estos dos módulos son, típicamente, módulos de lógica.

En este punto, el módulo de supervisión y de control 355 es un módulo estándar de supervisión y de control, utilizado generalmente en un nodo de administración para supervisar y controlar la debida ejecución de ciertas operaciones efectuadas en un clúster.

Así, típicamente está diseñado para procesar directamente paquetes de datos recibidos de la red de comunicación del clúster, es decir, en este punto, de la red de comunicación de tipo InfiniBand, en particular, paquetes de tipo

MAD.

El módulo retransmisor 350 puesto en práctica en el dispositivo de supervisión y de control 340 permite transmitir datos recibidos del nodo retransmisor 315, por ejemplo en forma de tramas, al módulo de supervisión y de control 355. Con estos fines, las tramas recibidas, de acuerdo con una forma particular de realización, son desencapsuladas para direccionar los datos contenidos en estas tramas en forma de paquetes (por ejemplo, paquetes de tipo MAD) al módulo de supervisión y de control 355.

Recíprocamente, el módulo retransmisor 350 permite transmitir datos recibidos del módulo de supervisión y de control 355, por ejemplo en forma de paquetes (por ejemplo, paquetes de tipo MAD), al nodo retransmisor 315, por ejemplo en forma de tramas.

De acuerdo con una forma particular de realización, se obtiene con anterioridad o se obtiene dinámicamente un identificador del nodo retransmisor 315 al que se deben transmitir los datos recibidos, por ejemplo una dirección IP. Siempre de acuerdo con una forma particular de realización, los paquetes recibidos son encapsulados para que sean transmitidos en forma de tramas.

Así, el módulo retransmisor 350 permite “engañar” al módulo de supervisión y de control 355 actuando como si fuera puesto en práctica en un dispositivo directamente unido a la red de comunicación del clúster, por ejemplo una red de comunicación de tipo InfiniBand. Este módulo se puede poner en práctica en forma de una biblioteca particular o con el concurso de una función de sobrecarga (denominada *overloading* en la terminología anglosajona) utilizando la variable de entorno conocida con el nombre de LD_PRELOAD en el entorno Unix (Unix es una marca).

Siempre de acuerdo con una forma particular de realización, el módulo retransmisor 330 es un demonio, es decir, un proceso que se ejecuta en segundo plano, ejecutado por un nodo que tiene al menos dos interfaces de red, para recibir datos en una interfaz de red y transmitirlos en otra interfaz de red.

Así, permite recibir paquetes de datos de control de clúster (por ejemplo, paquetes de tipo MAD) de una red de comunicación de tipo InfiniBand, encapsularlos en tramas y transmitirlos en una red de comunicación de tipo Ethernet. Recíprocamente, permite recibir tramas de datos de una red de comunicación de tipo Ethernet, desencapsularlas y transmitirlas en forma de paquetes (por ejemplo, paquetes de tipo MAD) en una red de comunicación de tipo InfiniBand.

Los paquetes de datos recibidos de la red de comunicación de tipo InfiniBand pueden ser recibidos de un dispositivo del clúster como respuesta a una petición previa, por ejemplo una petición procedente del dispositivo de supervisión y de control, o de manera autónoma.

Igualmente, los paquetes de datos emitidos por el módulo de supervisión y de control con destino a uno o varios dispositivos del clúster pueden ser emitidos como respuesta a datos recibidos con anterioridad de uno o de varios dispositivos del clúster o de manera autónoma.

La figura 4, que comprende las figuras 4a y 4b, ilustra esquemáticamente etapas puestas en práctica en un nodo de un clúster también unido a una red de comunicación de un tipo diferenciado de la propia del clúster, para transmitir datos de supervisión a un dispositivo remoto (etapas 400 a 410 de la figura 4a) y recibir datos de control de este dispositivo remoto (etapas 415 a 425 de la figura 4b). El nodo que pone en práctica las etapas 400 a 425 es, por ejemplo, el nodo retransmisor 315 descrito con referencia a la figura 3.

Las etapas 400 a 410 y las etapas 415 a 425 son ejecutadas de manera diferenciada, típicamente en paralelo.

Como se ilustra, para transmitir datos de control de un clúster entre un dispositivo de este clúster y un dispositivo remoto (es decir, no unido directamente al clúster), tiene por objeto una primera etapa la recepción de los datos (etapa 400).

De acuerdo con una forma particular de realización, los datos intercambiados entre dispositivos del clúster se transmiten en forma de paquetes que comprenden un identificador de destinatario. Este identificador puede estar ligado a un solo dispositivo (transmisión llamada *unicast*) o a varios dispositivos (transmisión llamada *multicast*). Un identificador de dispositivo puede ser, por ejemplo, una dirección local, por ejemplo una dirección conocida con el nombre de LID.

De acuerdo con esta forma de realización, los datos recibidos en el curso de la etapa 400 son paquetes que comprenden, como destinatario, el identificador del nodo retransmisor que pone en práctica las etapas 400 a 410. Estos paquetes son recibidos por una primera interfaz de red del nodo retransmisor, por ejemplo una interfaz de tipo InfiniBand.

En una etapa siguiente (etapa 405), los datos recibidos son encapsulados en una o varias tramas de datos cuyo

formato está definido por el protocolo puesto en práctica en la red de comunicación que une el nodo retransmisor que pone en práctica las etapas 400 a 410 y el dispositivo al que se deben transmitir los datos (es decir, el dispositivo utilizado para supervisar y controlar el clúster).

Estas tramas comprenden un identificador del dispositivo al que se deben transmitir los datos, por ejemplo la dirección IP (siglas de *Internet Protocol* en la terminología anglosajona) de la interfaz de red utilizada de este dispositivo. De acuerdo con una forma particular de realización, este identificador está memorizado con anterioridad en el nodo retransmisor.

Después de haber sido encapsulados, los datos que han de transmitirse se transmiten (etapa 410) a través de una segunda interfaz de red del nodo retransmisor que pone en práctica las etapas 400 a 410, que ofrece un acceso a la red de comunicación que une este nodo al dispositivo al que se deben transmitir los datos, por ejemplo una red de comunicación de tipo Ethernet.

Paralelamente, para transmitir datos de control de un clúster entre un dispositivo remoto (es decir, no unido directamente al clúster) y un dispositivo de este clúster, tiene por objeto una primera etapa la recepción de los datos de control (etapa 415).

De acuerdo con una forma particular de realización, los datos intercambiados entre este dispositivo remoto y el nodo del clúster que pone en práctica las etapas 415 a 425 (es decir, el nodo retransmisor) se transmiten en forma de tramas que encapsulan los datos que han de intercambiarse, pudiendo a su vez estas últimas estar organizadas en forma de paquetes de datos, comprendiendo estas tramas un identificador del nodo retransmisor. Este identificador puede ser, por ejemplo, la dirección IP de la segunda interfaz de red del nodo retransmisor.

En una etapa siguiente (etapa 420), los datos recibidos en forma de tramas son desencapsulados para recuperar los datos en un formato compatible con la red de comunicación del clúster, típicamente paquetes de datos que comprenden un identificador de uno o varios dispositivos en el clúster.

Después de haber sido desencapsulados, los datos que han de transmitirse se transmiten (etapa 425) a través de la primera interfaz de red del nodo retransmisor, que ofrece un acceso a la red de comunicación del clúster.

La figura 5, que comprende las figuras 5a y 5b, ilustra esquemáticamente etapas puestas en práctica en un ordenador remoto, unido a un nodo retransmisor de un clúster, para recibir datos de supervisión de este nodo retransmisor y procesarlos (etapas 500 a 515 de la figura 5a) así como para generar y transmitir datos de control a este nodo retransmisor (etapas 520 a 535 de la figura b).

El dispositivo que pone en práctica las etapas 500 a 535 es, por ejemplo, el dispositivo 340 descrito con referencia a la figura 3.

Al igual que las etapas 400 a 410 y las etapas 415 a 425, las etapas 500 a 515 y las etapas 520 a 535 son ejecutadas de manera diferenciada, típicamente en paralelo.

Como se ilustra, para procesar, típicamente analizar, datos que permiten supervisar y controlar un clúster, a partir de un ordenador remoto (es decir, no unido directamente al clúster), tiene por objeto una primera etapa la recepción de los datos transmitidos por el nodo retransmisor (etapa 500).

Estos datos son recibidos por una interfaz de red del ordenador remoto, uniendo esta interfaz de red este último a un nodo retransmisor a través de una red de comunicación (por ejemplo, red de comunicación de tipo Ethernet).

De acuerdo con una forma particular de realización, los datos intercambiados entre el nodo retransmisor y el ordenador remoto se transmiten en forma de tramas, por ejemplo de tramas Ethernet, que comprenden uno o varios paquetes de datos. Cada trama comprende un identificador del ordenador remoto, por ejemplo la dirección IP de su interfaz de red.

En una etapa siguiente (etapa 505), los datos recibidos son desencapsulados a efectos de recuperar los datos de origen (es decir, previo al encapsulado) conformes al protocolo de comunicación utilizado por el clúster, por ejemplo paquetes de datos conformes al estándar InfiniBand.

Los datos recuperados se transmiten entonces a un módulo de supervisión y de control del ordenador remoto (etapa 510), el cual los procesa (etapa 515).

Como anteriormente se ha descrito, el módulo de supervisión y de control es, de acuerdo con una forma particular de realización, un módulo de supervisión y de control configurado para su puesta en práctica en un nodo del clúster. Dicho de otro modo, el módulo de supervisión y de control está configurado para procesar datos conformes a un protocolo utilizado por la red de comunicación del clúster.

Paralelamente, para transmitir datos de control de un clúster entre el dispositivo remoto y un dispositivo de este clúster, tiene por objeto una primera etapa generar los datos de control (etapa 520).

Estos datos son generados típicamente en el módulo de supervisión y de control que con anterioridad ha procesado datos recibidos del clúster (habiéndose recibido estos últimos típicamente como respuesta a una petición anterior).

Después de haber sido recibidos (etapa 525), por ejemplo del módulo de supervisión y de control, los datos de control son encapsulados en una o varias tramas de datos cuyo formato está definido por el protocolo puesto en práctica en la red de comunicación que une el ordenador remoto al nodo retransmisor (etapa 530).

Se trata, por ejemplo, de tramas Ethernet. Estas comprenden un identificador del nodo retransmisor al que se deben transmitir los datos, por ejemplo su dirección IP obtenida con anterioridad por el ordenador remoto.

Los datos encapsulados son transmitidos entonces por el ordenador remoto, a través de una interfaz de red, al nodo retransmisor (etapa 535). Este último los transfiere entonces a la red de comunicación del clúster según lo descrito con referencia a la figura 4b.

Interesa señalar que los algoritmos descritos con referencia a las figuras 4 y 5 se pueden por en práctica, por ejemplo, en un dispositivo similar al descrito con referencia a la figura 2, en forma de un programa de ordenador.

Obviamente, para satisfacer necesidades específicas, una persona perita en la materia de la invención podrá introducir modificaciones en la anterior descripción.

REIVINDICACIONES

1. Procedimiento de supervisión y de control trasladados de un clúster que comprende una pluralidad de nodos (315, 335) conectados a una red de comunicación de tipo InfiniBand (305), comprendiendo un nodo (315) de dicha pluralidad de nodos, denominado nodo retransmisor, una primera interfaz de red (210-1) conforme a dicha red de comunicación de tipo InfiniBand y una segunda interfaz de red (320-2) conforme a una red de comunicación de tipo Ethernet, estando unido dicho nodo retransmisor a un ordenador remoto (340) mediante una red de comunicación de tipo Ethernet (310), comprendiendo el procedimiento las siguientes etapas:

- recepción (400), por parte del nodo retransmisor, de al menos un paquete de datos de supervisión a través de dicha primera interfaz de red;
- encapsulado (405), por parte del nodo retransmisor, de dicho al menos un paquete de datos de supervisión recibido en al menos una trama de datos de supervisión conforme a un protocolo de dicha red de comunicación de tipo Ethernet;
- transmisión (410), por parte del nodo retransmisor, de dicha al menos una trama de datos de supervisión a dicho ordenador remoto a través de dicha segunda interfaz de red,

en el que el ordenador remoto pone en práctica un módulo de supervisión y de control (355) y un módulo retransmisor (350) configurado para transmitir datos recibidos del nodo retransmisor (315) al módulo de supervisión y de control (355), de modo que el módulo retransmisor (350) está configurado para “engañar” al módulo de supervisión y de control (355) actuando como si fuera puesto en práctica en un dispositivo directamente unido a la red de comunicación de tipo InfiniBand.

2. Procedimiento según la reivindicación 1 que comprende además las siguientes etapas,

- recepción (415), por parte del nodo retransmisor, a través de dicha segunda interfaz de red, de dicho ordenador remoto, de al menos una trama de datos de control que comprende al menos un paquete de datos de control;
- desencapsulado (420), por parte del nodo retransmisor, de dicha al menos una trama de datos de control recibida para recuperar dicho al menos un paquete de datos de control contenido en dicha al menos una trama de datos de control recibida; y
- transmisión (425), por parte del nodo retransmisor, a través de dicha primera interfaz de red, de dicho al menos un paquete de datos de control recuperado.

3. Procedimiento según la reivindicación 1 o la reivindicación 2 que comprende además las siguientes etapas,

- recepción (500), por parte del ordenador remoto, de dicho nodo retransmisor, de al menos una trama de datos que comprende al menos un paquete de datos;
- desencapsulado (505), por parte del ordenador remoto, de dicha al menos una trama de datos recibida de dicho nodo retransmisor para recuperar dicho al menos un paquete de datos contenido en dicha trama de datos recibida de dicho nodo retransmisor; y
- procesamiento (515), por parte del ordenador remoto, de dicho al menos un paquete de datos recuperado contenido en dicha trama de datos recibida de dicho nodo retransmisor, poniéndose en práctica dicha etapa de procesamiento de dicho al menos un paquete de datos recuperado en el módulo de supervisión y de control, estando dicho módulo de supervisión y de control configurado para procesar datos conformes a un protocolo utilizado por la red de comunicación de tipo InfiniBand.

4. Procedimiento según una cualquiera de las reivindicaciones 1 a 3 que comprende además las siguientes etapas, puestas en práctica en dicho ordenador remoto,

- recepción (525), de un módulo de supervisión y de control configurado para su puesta en práctica en un nodo de dicho clúster, de al menos un paquete de datos;
- encapsulado (530) de dicho al menos un paquete de datos recibido de dicho módulo de supervisión y de control configurado para su puesta en práctica en un nodo de dicho clúster, en al menos una trama de datos conforme a un protocolo de dicha red de comunicación de dicho segundo tipo;
- transmisión (535), a dicho nodo retransmisor, de dicha al menos una trama de datos que comprende dicho al menos un paquete de datos recibido de dicho módulo de supervisión y de control configurado para procesar datos conformes a un protocolo utilizado por la red de comunicación de tipo InfiniBand.

5. Procedimiento según una cualquiera de las reivindicaciones 1 a 4, que comprende además una etapa previa de memorización, en dicho nodo retransmisor, de una dirección de dicho ordenador remoto, siendo utilizada dicha dirección de dicho ordenador remoto memorizada para encapsular al menos un paquete de datos que ha de transmitirse a dicho ordenador remoto en forma de al menos una trama de datos.

6. Procedimiento según una cualquiera de las reivindicaciones 1 a 5, que comprende además una etapa de

memorización, en dicho ordenador remoto, de una dirección de dicho nodo retransmisor, siendo utilizada dicha dirección de dicho nodo retransmisor memorizada para encapsular al menos un paquete de datos que ha de transmitirse a dicho nodo retransmisor en forma de al menos una trama de datos.

7. Procedimiento según una cualquiera de las reivindicaciones 1 a 6, según el cual al menos un paquete de datos transmitido de dicho nodo retransmisor a dicho ordenador remoto es un paquete de datos de tipo MAD que comprende información relativa a dicho clúster.

8. Programa de ordenador que comprende instrucciones adaptadas para la puesta en práctica de cada una de las etapas del procedimiento según una cualquiera de las reivindicaciones anteriores cuando dicho programa es ejecutado en un ordenador.

9. Sistema que comprende al menos un nodo retransmisor y al menos un ordenador remoto que comprende medios configurados para poner en práctica cada una de las etapas del procedimiento según una cualquiera de las reivindicaciones 1 a 7.

Fig. 1

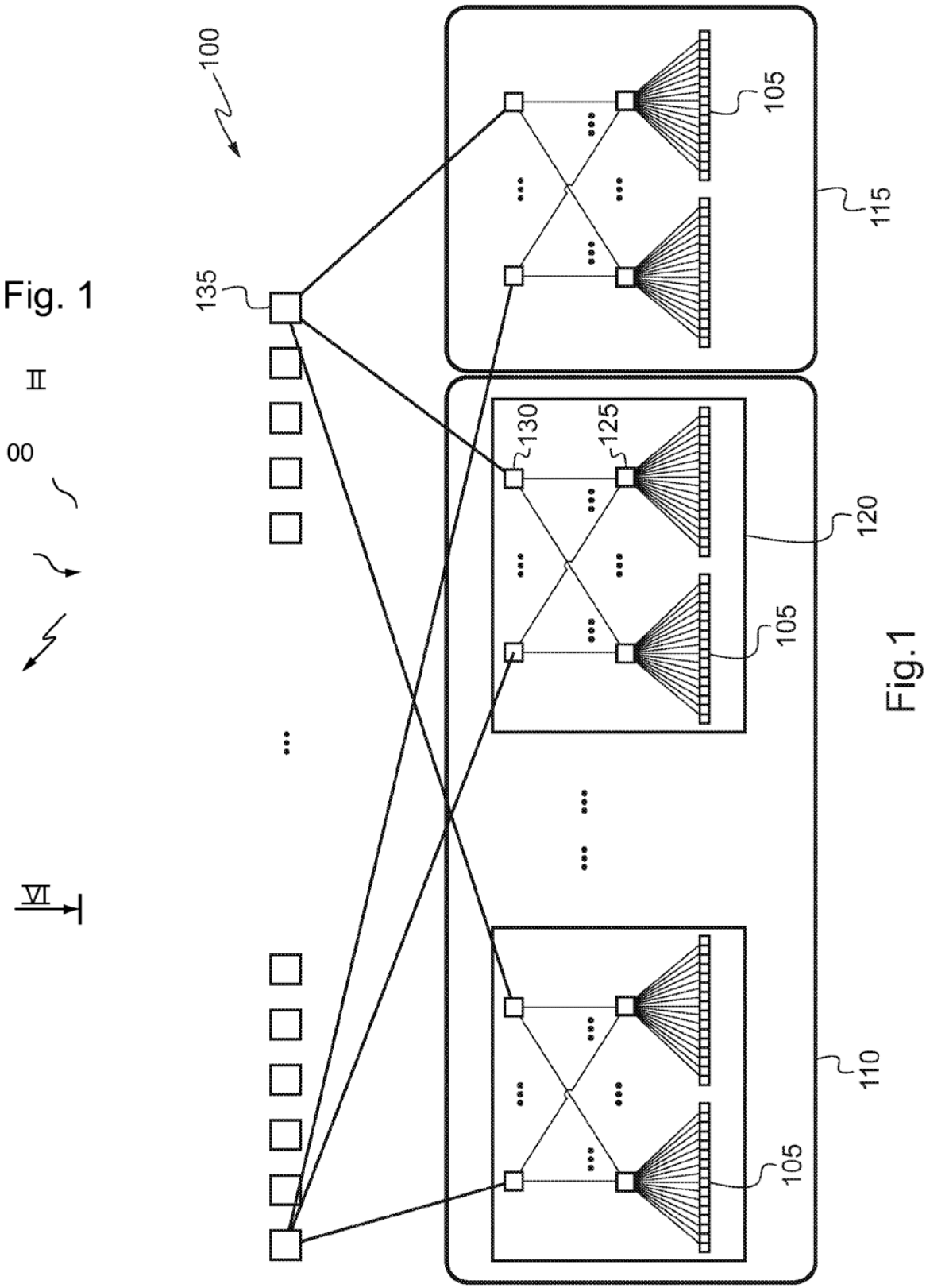


Fig.1

Fig. 1


II
00
VI

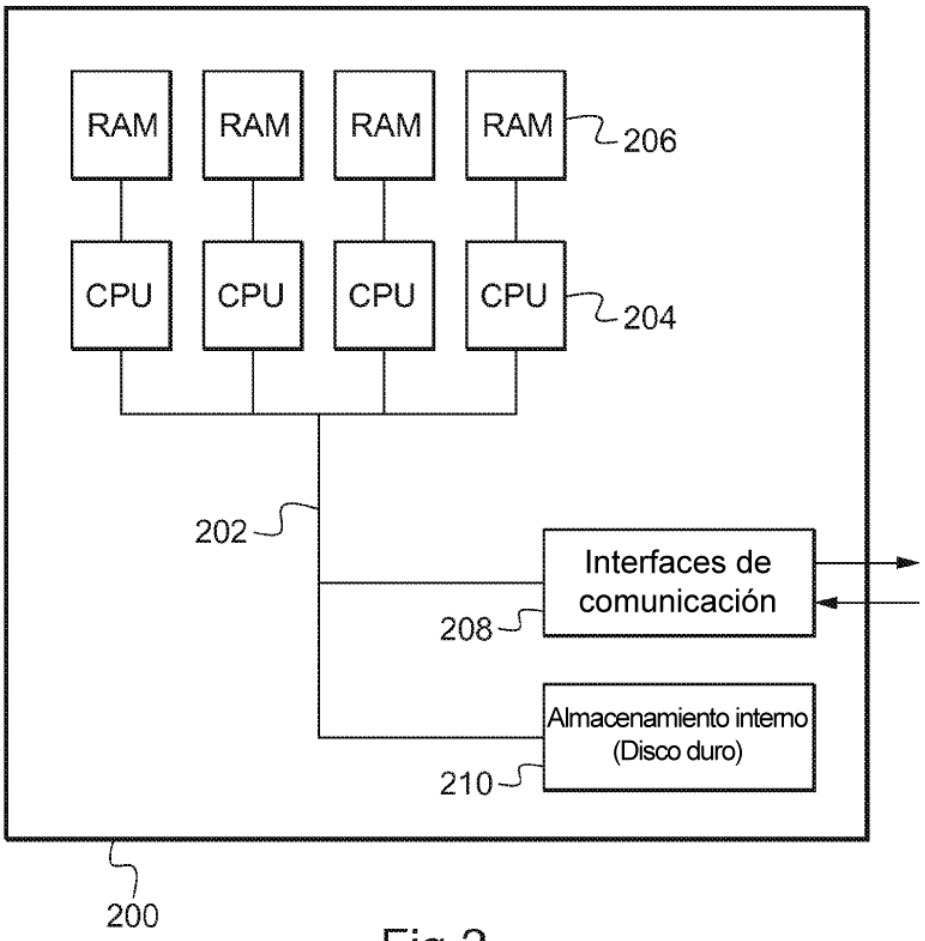

RAM
RAM
RAM
RAM
206
CPU
CPU
CPU
CPU
204
202
Interfaces de comunicación
208
Almacenamiento interno (Disco duro)
210
200

Fig.2

Fig. 1

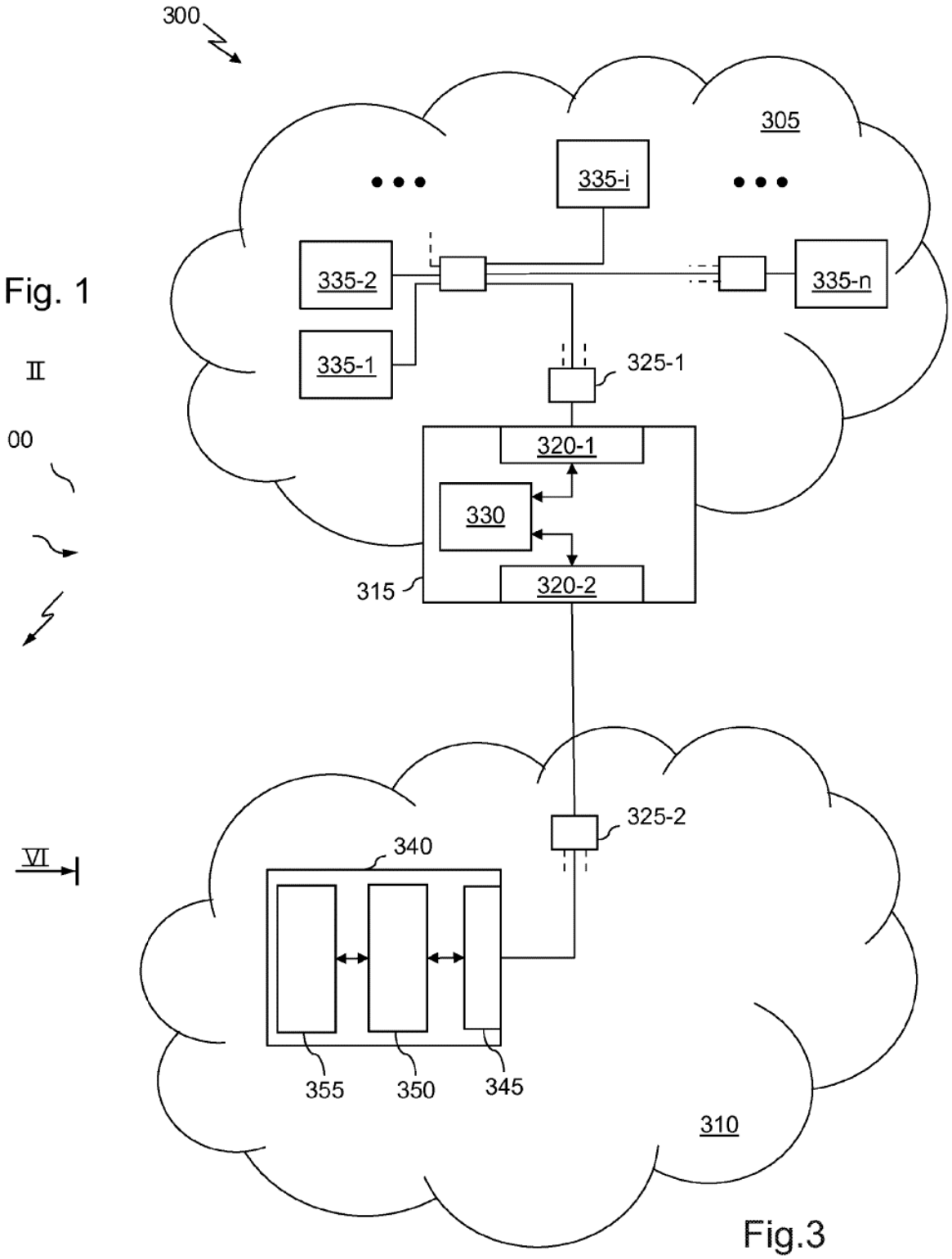
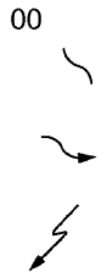

300
305
335-i
335-2
335-n
335-1
325-1
320-1
330
320-2
315
325-2
340
355
350
345
310

Fig.3

Fig. 1



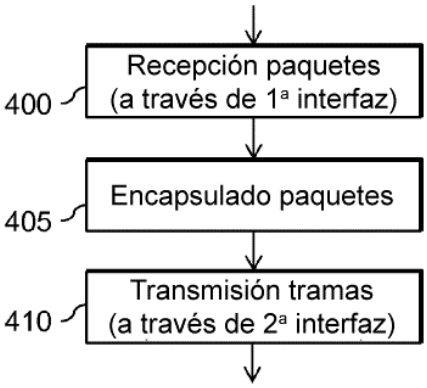


Figura 4a

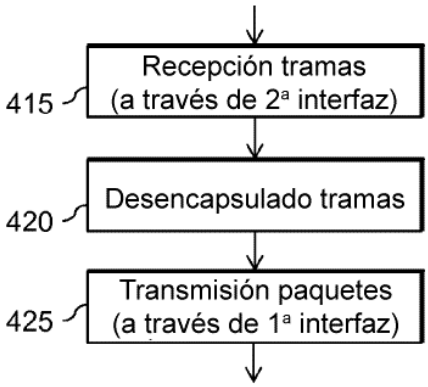


Figura 4b

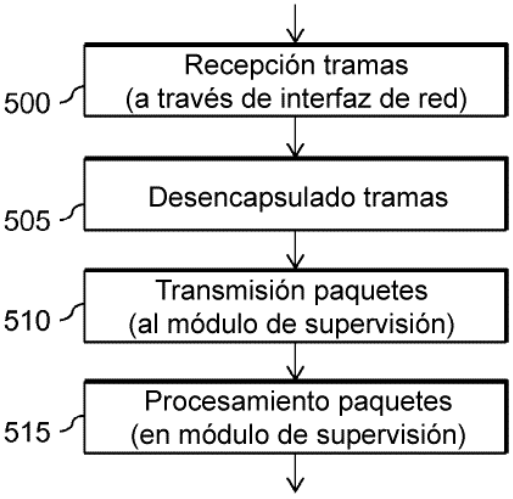


Figura 5a

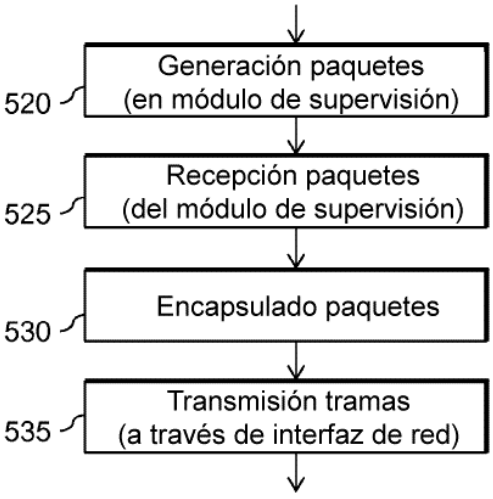


Figura 5b