(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 614 614**

(51) Int. CI.:

H04L 12/715 (2013.01)
H04L 29/08 (2006.01)
H04L 29/12 (2006.01)
H04L 12/46 (2006.01)
H04L 12/707 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **28.05.2010 PCT/US2010/036757**

(87) Fecha y número de publicación internacional: **02.12.2010 WO10138936**

(96) Fecha de presentación y número de la solicitud europea: **28.05.2010 E 10781357 (8)**

(97) Fecha y número de publicación de la concesión europea: **09.11.2016 EP 2436156**

(54) Título: **Igualación de carga en dominios de capa-2**

(30) Prioridad:

28.05.2009 US 182057 P
26.10.2009 US 605388

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
01.06.2017

(73) Titular/es:

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

(72) Inventor/es:

PATEL, PARVEEN;
MALTZ, DAVID;
GREENBERG, ALBERT;
YUAN, LIHUA y
KERN, RANDY

(74) Agente/Representante:

DE ELZABURU MÁRQUEZ, Alberto

DESCRIPCIÓN

Igualación de carga en dominios de capa-2

Antecedentes

Los igualadores o compensadores de carga pueden ser una parte crítica de una infraestructura de red que pueden distribuir un conjunto de peticiones sobre un conjunto de servidores capaces de procesar las peticiones. Los igualadores de carga convencionales pueden incluir pares de dispositivos, cada uno de los cuales es un hardware dedicado o especializado. Debido al uso de este hardware dedicado, los igualadores de carga convencionales tienden a costar mucho dinero. Otro inconveniente es que usan una estrategia de crecimiento escalado: una pareja única de igualadores de carga pueden gestionar un número de peticiones concurrentes limitadas por la capacidad del hardware. Se adquieren igualadores de carga más potentes que contienen hardware de mayor capacidad para gestionar peticiones adicionales. Una optimización de Retorno Directo de Servidor (DSR) puede ser útil con respecto a la reducción de cuellos de botella de tráfico en una red. Sin embargo, un inconveniente de los igualadores de carga convencionales es que esta técnica está típicamente limitada a una única red de área local virtual (VLAN) de la red.

El documento EP 1 494 422 describe un método de igualación de carga en el cual una dirección de fuente y una dirección de destino de VIP de paquetes entrantes son reemplazadas por una dirección de origen y una dirección de objetivo de una infraestructura de igualación de carga. En particular, el método incluye obtener una pareja de fuente/destino a partir de un paquete; acceder a una tabla de mapeo de encapsulación utilizando la pareja fuente/destino para localizar una entrada de mapeo de encapsulación; extraer un identificador de flujo a partir de la entrada de mapeo de encapsulación; y reemplazar parte del paquete con el identificador de flujo para producir un paquete encapsulado.

El documento titulado “Towards a Next Generation Data Center Architecture: Scalability and Commoditization”, por Albert Greenberg et al, proporciona un debate general de una arquitectura de red que utiliza conmutadores y servidores de capa-2 en los cuales la función de igualación de carga es descompuesta en un grupo de servidores regulares, con el resultado de que se puede distribuir un hardware de servidor de igualación de carga entre bastidores de un centro de datos, facilitando una mejor escalabilidad.

Compendio

La invención proporciona un método y un sistema de igualación de carga tal como se reivindica a continuación en la presente memoria.

Breve descripción de los dibujos

Los dibujos adjuntos representan implementaciones de los conceptos contenidos en la presente aplicación. Las características de las implementaciones representadas pueden ser comprendidas con mayor facilidad en referencia a la siguiente descripción tomada en conjunto con los dibujos adjuntos. Los mismos números de referencia son utilizados en diferentes dibujos siempre que sea posible para indicar elementos similares. Además, el numeral más a la izquierda de cada número de referencia dirige la figura y una explicación asociada en la cual se ha presentado primero el número de referencia.

Las Figuras 1 a 5 representan entornos de red que pueden utilizar algunos de los conceptos presentes de acuerdo con algunas implementaciones.

La Figura 6 representa una arquitectura de igualación de carga escalable que puede utilizar alguno de los conceptos presentes de acuerdo con algunas implementaciones.

Las Figuras 7 a 8 representan algunos componentes introducidos en las Figuras 1 a 6 de acuerdo con algunas implementaciones de los conceptos presentes.

La Figura 9 representa una técnica de mapeado de aleatorización o generación de claves (hashing) que es coherente con algunos de los conceptos presentes de acuerdo con algunas implementaciones.

Las Figuras 10 y 11 representan diagramas de flujo que pueden implementar algunos de los conceptos de igualación de carga escalable de acuerdo con algunas implementaciones.

Descripción detallada

Introducción/visión general

Un igualador de carga de red puede ayudar a mejorar la utilización de recursos en una red, mediante la clasificación de paquetes entrantes en sesiones y, distribuyendo el tráfico de paquetes para sesiones individuales hacia un(os) recurso(s) escogido(s) (por ejemplo, servidor(es)). Para ayudar en la reducción de la aparición de cuellos de botella de tráfico de paquetes en el igualador de carga, se puede utilizar una técnica de optimización tal como Retorno

Directo de Servidor (DSR). DSR permite el tráfico de paquetes salientes de la red para eludir al igualador de carga en lugar de tener que pasar a través de él como lo hace el tráfico de paquetes entrante. Sin embargo, esta técnica está típicamente limitada a una única red de área local virtual (VLAN) de la red. En contraste, la Figura 1 representa una vista de alto nivel de algunos de los conceptos presentes.

Ejemplos de red

La Figura 1 representa un entorno de red 100 donde cliente(s) externo(s) cliente externo 102 puede(n) comunicarse con un sistema 104 de igualación de carga escalable a través de la Internet 106. La igualación o difusión de la carga puede ser considerada como uno de cualesquiera medios mediante los cuales un dispositivo de red puede difundir tráfico a través de un conjunto de siguientes saltos válidos.

Un sistema 104 de igualación de carga escalable puede incluir una capa 108 funcional de igualación de carga que es escalable en cuanto a que puede soportar una cantidad esencialmente ilimitada de dispositivos objetivo indicados como 110. En este caso, la expresión “esencialmente ilimitado” puede significar de forma general tantos dispositivos objetivo como sean requeridos por un sistema 104 de igualación de carga escalable controlador de entidad. Por ejemplo, el número de dispositivos objetivo puede ser de decenas o cientos de miles, o más. La capa 108 de funcionalidad de igualación de carga está configurada de modo tal que las comunicaciones procedentes del cliente externo 102 pueden pasar a través de, y ser distribuidas por, la funcionalidad de igualación de carga, hacia dispositivos objetivo individuales, tal como se representa mediante la flecha 112. Sin embargo, las comunicaciones retornantes representadas por la flecha 114 no necesitan pasar a través de la capa 108 de funcionalidad de igualación de carga en el camino de regreso hacia el cliente externo 102.

Brevemente, algunas implementaciones pueden obtener una capa 108 de funcionalidad de igualación de carga utilizando técnicas de emisión de paquetes de dominio inter capa-2. En algunos casos, estas técnicas de emisión de paquetes de dominio inter capa-2 pueden permitir técnicas de optimización de igualación de carga, tales como DSR, para ser utilizadas a través de múltiples sub-redes de IP y permitiendo así el uso de un número esencialmente ilimitado de unos dispositivos objetivo 110. Por razones de escalabilidad y de otras, las redes que utilizan el Protocolo de Internet pueden dividir huéspedes que comparten un prefijo de bit común en sus Direcciones de IP en una subred de IP. Típicamente, el alcance de una subred simple está confinado al alcance de una VLAN simple. La concesión del uso de dispositivos objetivo 110 con Direcciones de Protocolo de Internet (IP) por parte de diferentes subredes puede eliminar limitaciones significativas de los diseños anteriores de igualadores de carga. Subredes de IP individuales pueden quedar asociadas con uno de los diferentes dominios de capa-2 del sistema 104 de igualación de carga escalable. En una o más realizaciones, los paquetes entrantes individuales de un flujo de paquetes pueden ser encapsulados utilizando, por ejemplo, encapsulación IP-en-IP. Esto puede ser realizado, por ejemplo, mediante un multiplexor (MUX o Mux) de la funcionalidad 108 de igualación de carga.

Los paquetes entrantes encapsulados pueden ser encaminados hacia recursos o dispositivos objetivo 110 sobre el sistema 104 de igualación de carga escalable mediante el paso a través de la funcionalidad 108 de igualación de carga antes de alcanzar a los dispositivos objetivo individuales. En al menos algunas realizaciones, la funcionalidad de igualación de carga puede utilizar una técnica de optimización, tal como una DSR, para reducir/minimizar el tráfico de flujo de paquetes en la funcionalidad de igualación de carga. Los dispositivos objetivo (por ejemplo, servidores) pueden estar asociados con, y por tanto abarcar, las múltiples sub-redes de IP o VLANs. Los componentes (por ejemplo, componentes de software) asociados con dispositivos objetivo individuales pueden desencapsular paquetes entrantes recibidos para obtener información de IP. Los resultados (paquetes salientes) pueden ser a continuación encaminados fuera del sistema 104 de igualación de carga escalable hacia el cliente externo 102 (por ejemplo, el cliente que recibe uno o más de los paquetes entrantes) sin pasar a través de (es decir, atravesando) la capa 108 de funcionalidad de igualación de carga. Brevemente, el sistema 104 de igualación de carga escalable puede permitir nuevas funcionalidades, incluyendo la funcionalidad asociada al Protocolo de Resolución de Direcciones Gratuitas (G-ARP) y de difusión de carga. Estos conceptos son ampliados a continuación.

La Figura 2 representa otro entorno 200 ejemplar de red según una o más realizaciones. El entorno 200 de red proporciona componentes o estructuras ejemplares que pueden realizar los conceptos presentados anteriormente en relación a la Figura 1. El entorno 200 de red puede incluir cliente(s) externo(s) 202 que se comuniquen con un sistema 204 de igualación de carga escalable a través de la Internet 106 u otra red. El sistema 204 de igualación de carga escalable puede incluir un conjunto de encaminadores 206, un conjunto de igualadores (208) de carga dinámicos (DLBs) y un conjunto de dispositivos objetivo 210. En este caso, el conjunto de encaminadores 206 están indicados como encaminadores 206(1) y 206(n). El conjunto de DLBs 208 están indicados como DLBs 208(1) y 208(n) que incluyen los multiplexadores (o MUXes) 212(1) y 212(n), respectivamente. El conjunto de dispositivos objetivo 210 está indicado como servidores de aplicación 214(1) y 214(n) e igualadores de carga locales 216(1) y 216(n).

Las flechas de puntos, indicadas de forma general en 218, muestran potenciales caminos de comunicación entre componentes del sistema 204 de igualación de carga escalable. Las flechas continuas en negrita 220(1) y 220(2) muestran dos caminos potenciales de flujo de paquetes a través del entorno 200 de red desde el cliente externo 202 hacia el servidor de aplicación 214(1). La flecha continua en negrita 222 representa un camino de flujo de paquetes de retorno desde el servidor de aplicación 214(1) hacia el cliente externo 202. Por ejemplo, las flechas en negrita

220(1) y 220(2) pueden representar una consulta de búsqueda procedente del cliente externo 202 que es gestionada por el servidor de aplicación 214(1). Como tal, el servidor de aplicación 214(1) puede ser denominado como el 'dispositivo objetivo'. Aunque en este caso el dispositivo objetivo es un servidor de aplicación en nivel de aplicación, deberá notarse y entenderse que el dispositivo objetivo ejemplar podría ser adicional o alternativamente otro tipo de dispositivo objetivo, tal como un igualador de carga local – por ejemplo, un igualador de carga en nivel de aplicación. Nótese que mientras que el flujo de paquetes entrante (es decir, las flechas en negrita 220(1) y 220(2)) pasa a través de un miembro del conjunto de DLBs 208, el flujo de paquetes de retorno saliente (es decir, la flecha en negrita 222) no pasa necesariamente a través del igualador de carga y en su lugar elude a los DLBs. Como resultado, se puede reducir o minimizar la aparición de cuello de botella sobre el tráfico de flujo de paquetes en uno o más de los DLBs. En al menos algunas realizaciones, esto se logra mediante la utilización de una técnica de optimización DSR. A continuación, se describen técnicas de optimización DSR ejemplares.

En al menos algunas realizaciones, los MUX(es) 212(1) y/o 212(n) en uno o más de los DLBs 208(1) y 208(n) pueden usar encapsulación de IP-en-IP para transmitir el flujo de paquetes hacia el dispositivo objetivo 210. Aunque se proporcionan ejemplos de encapsulación específicos, la encapsulación puede estar realizada por cualquier medio al direccionamiento de un paquete para su transmisión a lo largo de un camino o una parte de un camino. Además, un componente de desencapsulación 222(1) a 222(n) en el dispositivo objetivo puede desencapsular un paquete o unos paquetes del flujo de paquetes entrante y enviar los resultados (es decir, el flujo de paquetes saliente) de vuelta hacia el cliente externo 202. En un caso, los componentes de desencapsulación 222(1) a 222(n) pueden estar indicados en el dispositivo objetivo 210 como componentes de software que son ejecutables por un procesador del dispositivo objetivo.

En esta implementación, los encaminadores 206 pueden utilizar Multicaminos de Coste Equivalente (ECMP) para difundir cargas de paquetes a través de los MUXes 212(1) y 212(n) de los DLBs 208. Además, los MUXes pueden ofrecer una aleatorización coherente a paquetes que son enviados hacia los dispositivos objetivo 210. En algunas de las presentes implementaciones, los DLBs 208 y los dispositivo objetivo 210 pueden ser implementados en un único dispositivo, tal como un(os) servidor(es). Por ejemplo, un dispositivo único de cálculo, tal como un servidor, puede incluir los DLBs 208(1) con el MUX 212(1) y el servidor de aplicación 214(1). En otras implementaciones, los DLBs pueden ser dispositivos independientes respecto a los dispositivos objetivo.

Durante el funcionamiento, cada uno de los DLBs 208 en el ejemplo puede ser configurado para proporcionar una interfaz de programa de aplicación (API) para gestionar IP virtual (VIP) para mapeos de IP directo (DIP) (por ejemplo, VIP → {$Ranura_1$, $Ranura_2$, $Ranura_3$, ..., $Ranura_N$}) de un mapa VIP-DIP. Unas ranuras individuales son asignadas a un DIP. Un DIP único puede aparecer en múltiples ocasiones en este mapa de VIP a DIP. Este mapa de VIP a DIP puede ser referido como un MapaVip.

Aunque anteriormente se describe como un mapeo entre una dirección de VIP única y una lista de direcciones de DIP, se ha de entender que cada dirección puede estar asociada además con un número de puerto (por ejemplo, un puerto de protocolo de control de transmisión (TCP) tal como el puerto 80). En esta generalización, una dirección de VIP o una dirección de VIP y un número de puerto pueden ser mapeados para una lista consistente en entradas que son o bien sólo una dirección de DIP o bien una dirección de DIP y un número de puerto. Una dirección de DIP única puede aparecer múltiples veces, ya sea sola, con un número de puerto diferente, o con el mismo número de puerto, según cualquier combinación. También pueden haber múltiples VIP o combinaciones de VIP con número de puerto que se mapeen para listas idénticas de DIPs y combinaciones de DIP con número de puerto. Los MUXes 212(1) a 212(n) individuales de los DLBs 208 pueden ser cada uno de ellos configurados para aleatorizar los campos de cabecera a partir de los paquetes individuales del flujo de paquetes entrante y enviar los paquetes individuales hacia una dirección de IP apropiada asociada con el(los) dispositivo(s) objetivo 210. Por ejemplo, considérese un paquete entrante ejemplar. Uno o ambos de los DLBs puede aleatorizar el paquete entrante ejemplar y escoger una ranura (por ejemplo, {$Ranura_1$, $Ranura_2$, $Ranura_3$, ..., $Ranura_N$} mediante el cálculo de:

$$Ranura_i = \text{Aleatorizar (campos de cabecera de paquete)} \text{ módulo } N$$

donde N es el número de ranuras en el mapa VIP-DIP. Los MUX(es) del(de los) DLB(s) pueden enviar a continuación el paquete entrante ejemplar a una dirección indicada en la $Ranura_i$. Una ventaja potencial de este diseño es que los paquetes que son parte del mismo flujo (por ejemplo, un flujo TCP donde todos los paquetes comparten el mismo quinteto de dirección de fuente de IP, dirección de IP de destino, puerto de fuente de TCP, puerto de destino de TCP y número de protocolo de IP) pueden ser reenviados hacia el mismo dispositivo objetivo 210 independientemente de cual DLB 208 procesa el paquete.

La Figura 3 representa otro entorno 300 de red ejemplar que ofrece una alternativa al entorno 200 de red descrito anteriormente. Brevemente, el entorno 300 de red es similar al entorno 200 de red. Sin embargo, en el entorno 300 de red, los igualadores de carga locales (LLBs) pueden estar pensados como una capa de intervención entre el DLB y los dispositivos objetivo. Específicamente, el entorno 300 de red incluye un cliente externo 302 que se comunica con el sistema 304 de igualación de carga escalable a través de la Internet u otra red 306. El sistema 304 de igualación de carga incluye una capa de encaminador 308, una capa de DLB 310, una capa de LLB 312 y una capa 314 de dispositivo objetivo. En este caso, la capa 314 de dispositivo objetivo incluye servidores de aplicación 314(1) a 314(n). La capa de LLB 312 incluye LLBs 312(1) a 312(n).

Los componentes de desencapsulación 316(1) a 316(n) son residentes en los LLBs 312(1) a 312(n), respectivamente. En esta configuración, la comunicación del cliente externo puede ser encapsulada en la capa de DLB 310 y desencapsulada después de su recepción en la capa de LLB 312. La comunicación puede ser entonces redirigida hacia el servidor de aplicación 314(1) a 314(n) apropiado. Cualquier comunicación retornante hacia el cliente externo 302 puede eludir las capas de DLB y de LLB 310 y 312, respectivamente. Eludiendo las capas de DLB y LLB se pueden evitar cuellos de botella potenciales y/o conservar los recursos del sistema para comunicaciones entrantes.

La Figura 4 representa otro ejemplo de alto nivel de componentes de un entorno 400 de red de un sistema de igualación de carga escalable. En este caso, estos componentes incluyen generadores de consultas 402(1) a 402(n), encaminadores de acceso (AR) 404(1) a 404(n), conmutadores de agregación de capa-2 406(1) a 406(n) y conmutadores cabecera de bastidor (ToR) 408(1) a 408(n). Los ToRs pueden comunicarse con varios componentes de bastidor de servidores tales como MUXes (M), monitores de salud (H), servidores (S), igualadores de carga (B).

Para un VIP1 cumpliendo un servicio particular, los ARs 404(1) a 404(n) pueden estar configurados con N rutas, apuntando cada una de estas rutas a su siguiente-salto a una dirección de IP intermedia (IIP) (IIP1 a IIPN) que tiene el mismo coste. Sobre los AR, las rutas pueden ser todas siguientes-saltos para el VIP. Por lo tanto, el AR puede distribuir por igual el tráfico entre las N direcciones de IIP. Estas rutas podrían ser configuradas como rutas estáticas sobre el AR con métricas equivalentes (es decir, rutas estáticas de coste idéntico (descritas a continuación en relación a la Figura 5)). Alternativamente, estas rutas podrían ser establecidas dinámicamente a través de un agente o locutor de protocolo de encaminamiento (por ejemplo, el Protocolo de Pasarela de Límite (BGP) o Primer Camino Abierto Más Corto (OSPF)) que tenga la sesión apropiada con el AR. Además, el AR puede ser configurado para anunciar el VIP. Los IIPs pueden ser distribuidos a través de los MUX's (M). Un MUX, además de en su propia dirección de IP (MIP), puede ser configurado en una o más direcciones de IIP de manera tal que pueda responder la petición de ARP para las IIPs configuradas.

Por lo tanto, un MUX individual puede recibir una parte de un tráfico dirigido. Después de recibir un paquete, el MUX individual puede ejecutar un algoritmo de aleatorización coherente para seleccionar un DLB activo para dirigir el tráfico.

Los MUX's pueden utilizar el mismo algoritmo de aleatorización coherente basado en el mismo conjunto de DLBs activos. Por lo tanto, un paquete puede ser dirigido hacia el mismo DLB sin importar qué MUX lo recibe desde el AR 404(1) a 404(n). Nótese que cuando se añade o se quita un nuevo DLB al o del conjunto, esto puede desencadenar algún cambio de la configuración local; sin embargo, se conservan las conexiones existentes.

La Figura 5 representa un entorno 500 de red y una técnica asociada para configurar N rutas estáticas de coste idéntico. En este caso, el entorno 500 de red incluye un encaminador de acceso 404(1) (presentado en la Figura 4), los IIP(1) a IIP(n), los MUXes 212(1) a 212(n) y los DLBs 208(1) a 208(n) (presentados en la Figura 2). El entorno 500 de red puede configurar N rutas estáticas de coste idéntico para cada VIP. El SALTO_SIGUIENTE de estas rutas estáticas de coste idéntico apunta a una dirección IIP(1) a IIP(n) de IP Intermedio (IIP). Estas direcciones IIP pueden ser obtenidas a partir de un grupo de direcciones separado, independiente del grupo de VIP y DIP. Esta implementación puede además activar la difusión de carga de manera que el tráfico sea distribuido de forma equitativa hacia las N direcciones de IIP.

En otra realización se podría utilizar un protocolo de encaminamiento, tal como una conexión BGP con los encaminadores, para informarles de los MUXes que están activos y tomando paquetes para cada VIP.

Varias implementaciones pueden abordar el asunto de cómo conservar conexiones de larga duración a medida que los módulos MUX vienen y van. Una solución utilizada en algunas implementaciones puede ser retener el estado de los flujos individuales manejados en cada MUX, y entregar una copia de este estado a los MUXes individuales a medida que son incorporados al sistema de igualación de carga escalable. Con el fin de gestionar la incorporación o la eliminación de MUXes sin interrumpir conexiones existentes, una alternativa es crear información de estado cada vez que se gestiona por primera vez una nueva conexión por parte de cualquier MUX. Este estado puede ser compartido entre MUXes o bien directamente mediante un mecanismo igual a igual o indirectamente enviándolo a una memoria centralizada lógicamente a partir de la cual cualquier MUX que necesite gestionar paquetes para una conexión pueda determinar el DIP al que otros MUXes han enviado los paquetes de esa conexión.

Una implementación alternativa que requiere mucha menos compartición del estado, y por tanto que puede ser mucho más escalable, es que los MUX o bien estén enviando paquetes, utilizando un mapeo actual entre las VIPs y las DIPs (es decir, un MapaVip), o durante un periodo de transición en el que cambian de enviar paquetes utilizando un MapaVip (V) a enviar paquetes utilizando otro MapaVip (V'). En esta realización, se puede hacer que los MUXes estén de acuerdo en cuanto a V, V' y a su estado de transición actual (es decir, en cuanto a si están en el estado de transición entre V y V' o si han comenzado todos a utilizar sólo V' para todos los paquetes). Cuando no están en régimen transitorio, todos los MUXes envían todos los paquetes utilizando el MapaVip actual. Cuando están en régimen transitorio, los MUXes crean una recomendación de estado local siempre que encuentran un paquete para una nueva conexión (por ejemplo, un paquete TCP SYN). Cuando se envían paquetes distintos de los que indican una nueva conexión, el MUX comprueba si dispone de estado para esa conexión. Si tiene estado, el MUX envía el

paquete utilizando el MapaVip V' nuevo; en caso contrario envía el paquete utilizando el MapaVip V antiguo.

Brevemente, en al menos algunas configuraciones, el MUX 212(1) a 212(n) puede disponer de los siguientes componentes principales: (1) el módulo de IIP que reclama la propiedad de un IIP al encaminador y recibe el tráfico para ese IIP, (2) el módulo de aleatorización coherente que determina cúal DLB 208(1) a 208(n) envía el tráfico, (3) el transformador de paquetes que modifica el paquete, (4) el monitor DLB local. Cualquiera o todos estos componentes pueden ser implementados sobre servidores ampliamente disponibles (es decir, servidores básicos) y/o sobre los encaminadores en diferentes implementaciones. Los componentes de MUX son descritos con mayor detalle a continuación en relación a las Figuras 6 a 8.

El módulo (IIP(1) a IIP(n)) de IIP puede ser responsable de registrar el MUX 212(1) a 212(n) en el encaminador mediante el protocolo ARP. Básicamente, el módulo de IIP puede establecer en el encaminador un mapeo de IP-MAC de la dirección de IIP y la dirección de MUX MAC.

Considérese la función ejemplar 'booleano AddIP (Dirección IIP de IP)': En esta función ejemplar, la dirección IIP puede ser incorporada como una dirección de IP secundaria sobre la interfaz de MUX. Nótese que es posible que un MUX tenga múltiples direcciones de IP secundarias. La 'AddIP()' puede hacer que la pila de la red MUX emita 3 peticiones ARP (G-ARP) gratuitas, que pueden actualizar la tabla ARP del encaminador (o desencadenar una detección de conflicto de dirección de IP en sí mismo).

Con fines explicativos, considérese la función ejemplar "EliminaIP (Dirección IIP de IP)": Esta función ejemplar puede eliminar la dirección IIP de la interfaz de MUX. Considérese además la función ejemplar "EnviaARP()." Esta función ejemplar puede forzar la emisión de una petición G-ARP. Esta petición G-ARP puede ser emitida a modo de medida preventiva para la corrección del mapeo de IIP-MAC.

G-ARP y Detección de Conflicto de Dirección

Cuando se incorpora una dirección de IP a la interfaz, el sistema operativo (OS) puede difundir una G-ARP (dentro del mismo dominio L2). Esta petición G-ARP puede solicitar la dirección de IP que está reclamando. Si ninguna otra máquina responde con esta dirección de IP, la dirección de IP puede ser incorporada con éxito. En caso contrario, se puede detectar un conflicto de dirección de IP y la pila del MUX puede evitar que la máquina reclame esta dirección de IP. Esto puede ocurrir si otro MUX ha reclamado el IIP (por ejemplo, conmutación por error) y ha fallado en su eliminación. Este escenario puede ser gestionado a través de medidas exteriores (por ejemplo, mediante el apagado de la máquina que se defiende).

Cuando un nuevo MUX, con finalidad de ejemplo, MUX “B”, necesita reemplazar al MUX “A” (por ejemplo, debido a un tiempo de inactividad planificado del MUX A y/o fallo del sistema en MUX A) el nuevo MUX B puede incorporar el(s) IIP(s) del MUX A a su propia interfaz.

En al menos una realización, un módulo tal como el descrito anteriormente puede dirigir flujos de paquetes hacia uno o más módulos con seguimiento o inspección de estado en un grupo de servidores, donde el módulo con seguimiento de estado puede mantenerse según estado del flujo. En este caso, los paquetes entrantes pueden fluir a través de rutas desde el cliente al módulo hacia el módulo con seguimiento de estado hacia el servidor objetivo que maneja la petición asociada. El flujo saliente puede encaminarse desde el servidor objetivo hacia el módulo con seguimiento de estado hacia el cliente. Según estado del flujo en el módulo con seguimiento de estado puede permitir que módulos con seguimiento de estado individuales apliquen regulación en el nivel de flujo para soportar características de igualación de carga adicionales. En particular, el módulo con seguimiento de estado puede, por ejemplo, inspeccionar cookies o URLs para personalizar la igualación de carga hacia el servidor objetivo para depender de la aplicación, en la petición del cliente, y/o del rol y/o la carga y/o condiciones de los servidores y elementos de red. Esta realización puede ser ventajosa debido a que puede difundir cargas de trabajo de estado intensivas y de CPU hacia tantos servidores como sea necesario.

En al menos una realización, el módulo puede adaptar su encaminamiento hacia el módulo con seguimiento de estado para depender de información más importante que la información de cabecera transportada en las cabeceras de aplicación y de TCP/IP. En particular, para soportar una función de acceso directo, tal como sucede en Windows 7 ®, el módulo puede aprender o participar en protocolos criptográficos, permitiendo la desencriptación de partes de un paquete. El elección de servidor objetivo del módulo con seguimiento de estado puede depender entonces de estas partes desencriptadas. El mecanismo puede ser construido de manera que el servidor objetivo retorna el flujo saliente hacia el módulo con seguimiento de estado que está disponible (y potencialmente el más apropiado) para manejarlo. Esto puede beneficiarse del uso de CPUs programables para implementar el módulo.

En al menos una realización, el módulo puede incluir la dirección destino original en alguna parte de la cabecera de paquete, tal como una opción de Protocolo de Internet (IP), y envíar el paquete al dispositivo objetivo. El dispositivo objetivo puede extraer esta información a partir del cabecera de paquete y utilizarla para enviar paquetes salientes directamente hacia la fuente (por ejemplo, un cliente externo), donde algunos de los paquetes no pasan a través del módulo.

La Figura 6 representa una arquitectura ejemplar 600 de sistema de igualación de carga escalable que puede

realizar los conceptos descritos anteriormente y a continuación. En este caso, la arquitectura 600 de sistema de igualación de carga escalable puede incluir un gestor 602 de igualación de carga escalable, estando un rol de MUX representado en 604 y estando un rol de DIP representado en 606. La arquitectura 600 de sistema de igualación de carga puede incluir además un monitor de salud 608, la sonda de salud 610 y un gestor de ruta 612. El rol de MUX 604 puede comprender a un controlador de MUX 614 funcionando en un modo de usuario 616 y a un activador 618 de MUX funcionando en un modo kernel 620. EL rol de DIP 606 puede comprender un controlador de DIP 622 funcionando en un modo de usuario 624 y a un activador 626 de desencapsulación funcionando en un modo kernel 628.

El gestor 602 de igualación de carga escalable puede ser concebido como el punto de entrada para interacciones con la arquitectura 600 de sistema de igualación de carga escalable. El gestor 602 de igualación de carga escalable puede proporcionar una API que se puede usar para gestionar una instancia de los conceptos de igualación de carga escalable. Una instancia de igualación de carga escalable puede ser especificada utilizando una configuración XML o una API.

EL gestor 602 de igualación de carga escalable puede ser responsable de la configuración del mapeo VIP:DIP en máquinas de MUX y de asegurar que las máquinas MUX permanecen en sincronía. Adicionalmente, el gestor 602 de igualación de carga escalable puede facilitar también la conservación de conexiones de larga duración cuando los DIPs son incorporados o adecuadamente eliminados de un grupo. Se describe esta característica con mayor detalle a continuación en relación a la Figura 9.

Para aumentar la disponibilidad, el gestor 602 de igualación de carga escalable puede ser replicado y se puede utilizar un algoritmo de selección principal para asegurar una coherencia de estado.

Un rol de MUX 604 puede ser configurado con una o más direcciones de IP intermedias (IIPs). Tal como se ha citado anteriormente, en relación a la Figura 4, un encaminador, tal como el encaminador 404(1) puede estar configurado para enviar hacia un conjunto de IIPs los paquetes destinados al VIP. El MUX configurado con un IIP dado realizará el procesamiento de MUX para paquetes redirigidos hacia ese IIP.

El controlador de MUX 614 puede controlar el activador 618 de MUX. El controlador de MUX puede exportar una API de servicios web que es utilizada por el Gestor 602 de Igualación de Carga Escalable para controlar el MUX. En algunas implementaciones, el controlador de MUX puede realizar la siguiente funcionalidad:

1. descargar el mapa VIP:DIP en el activador;

2. informar al activador acerca de conexiones de larga duración;

3. recoger estadísticas procedentes del activador;

4. configurar IIP sobre la interfaz de red;

5. transmitir paquetes G-ARP sobre la red para el IIP especificado para atraer al MUX cualquier paquete dirigido hacia el IIP por los encaminadores u otros huéspedes en la red.

El activador 618 de MUX puede implementar la funcionalidad de modificación del paquete base. El activador de MUX puede aleatorizar los campos de cabecera de un paquete entrante, recoger un DIP basado en el valor de aleatorización y en el mapa de VIP actual y encapsular el paquete para su transmisión. Además del mapa, el activador 618 de MUX puede además mantener una memoria caché de mapeo aleatorización:DIP de todas las conexiones de larga duración para cada VIP.

El controlador de DIP 622 puede controlar el activador 626 de desencapsulación sobre la máquina de DIP. Similar al Controlador de MUX 614, el controlador de DIP 622 puede exportar una API de servicios web que es utilizada por el gestor 602 de igualación de carga escalable para controlar y consultar a la máquina de DIP. En algunas implementaciones, el controlador de DIP 622 puede realizar las siguientes funciones:

1. configurar los VIPs en una interfaz de bucle de retorno;

2. configurar la desencapsulación para VIPs específicos;

3. consultar a la máquina de DIP acerca de conexiones actualmente activas;

4. consultar acerca de la salud de la máquina de DIP (esto es opcional, dependiendo de la implementación del monitor de salud).

El activador 626 de desencapsulación puede desencapsular paquetes IP-en-IP que están destinados al VIP específico. Esta característica ayuda a evitar la interrupción de las comunicaciones existentes con aplicaciones específicas. Por ejemplo, si hay una aplicación que está utilizando conectores en bruto para enviar un IP-en-IP (por ejemplo, una app VPN de red privada virtual), entonces el activador 626 de desencapsulación no los desencapsula.

El gestor de ruta 612 puede ser responsable de configurar los encaminadores cuando las máquinas de MUX son incorporadas o eliminadas del grupo. El gestor de ruta puede utilizar un protocolo de encaminamiento, tal como OSPF o BGP, o una interfaz para configurar rutas estáticas en los encaminadores.

El monitor de salud 608 puede ser responsable de mantener el estado de salud de las máquinas de MUX y DIP y posiblemente para rutas implicadas en el procesado de peticiones. Para este fin, el monitor de salud puede supervisar uno o más parámetros de red que pueden ser de valor en la determinación de la salud de la red y/o de los componentes de red. El gestor 602 de igualación de carga escalable puede usar el monitor de salud 608 a modo de fuente autoritaria de información de salud acerca de los MUXes y los DIPs. Si el monitor de salud 608 notifica al gestor 602 de igualación de carga escalable acerca de un evento de cambio en la salud, el gestor de igualación de carga escalable puede tomar la acción apropiada de añadir o eliminar ese nodo del grupo correspondiente.

Visto en perspectiva, el monitor de salud 608 puede ser utilizado para supervisar la salud de los MUXes, los DLBs y/o las rutas a esas máquinas.

En al menos algunas implementaciones, el monitor de salud 608 puede consistir en tres módulos, un marcador de VPN, un monitor de MUX y un monitor de DLB. DLB puede proporcionar una interfaz de HTTP. El monitor de salud 608 puede emplear varios tipos de sondas de salud 610 para establecer la salud de los componentes objetivo. Por ejemplo, el monitor de salud puede emitir un “obtener http” para requerir un pequeño archivo de texto/xml de DLB. Si el archivo contiene la 'palabra mágica' que el monitor de salud y el DLB han acordado, entonces el monitor de salud puede considerar que el DLB está activo y en funcionamiento, y determinar si un DLB o un MUX están funcionando como se espera. Adicionalmente, en al menos algunas realizaciones, los componentes del monitor de salud pueden operar en dispositivos separados de los dispositivos de MUX.

La sonda de salud 610 puede ser utilizada por el monitor de salud 608. Por ejemplo, el monitor de salud puede utilizar varias sondas de salud para realizar su trabajo. Las sondas de salud 610 pueden supervisar activamente un aspecto de la salud de la máquina objetivo, por ejemplo, una sonda de ping (verificación mediante paquetes de Internet) supervisa la conectividad y la vitalidad de la máquina. Otras sondas de salud pueden simplemente consultar la máquina/rol acerca de su salud – la máquina/rol puede ser responsable de mantener un registro de su salud, consultándolo la sonda simplemente de forma periódica.

Si una sonda de HTTP tiene éxito, esto puede indicar que todo está activo y en funcionamiento. Pero, debido a que funciona sobre TCP, es posible que el DLB pueda estar funcionando sin conectores u otros recursos por un tiempo. También es posible que durante un ataque de denegación de servicio (DoS) un DLB pueda estar funcionando sin recursos (por ejemplo, conectores) durante un periodo de tiempo prolongado. Una solución a esto puede ser mantener una conexión de HTTP persistente. Sin embargo, la mayoría de las implementaciones de servidor/navegador darán por terminadas las conexiones de TCP persistentes. Por ejemplo, algunos navegadores pueden dar por terminada una conexión persistente después de 60 segundos. Por lo tanto, el monitor de salud está preparado para renovar una conexión persistente si está cerrada, y no debería necesariamente ver el cierre de una conexión persistente como una indicación de un fallo de DIP.

Si otro MUX puede reemplazar al MUX fallido, ya que todos los MUXes funcionan sobre la misma función de aleatorización coherente, los paquetes serán dirigidos hacia el mismo DLB. Por tanto, el flujo (por ejemplo, una conexión de TCP) no debe verse alterada.

Un grupo aparte de MUXes puede estar disponible a modo de pausa caliente de los MUXes activos. El monitor de salud 608, después de detectar un fallo de MUX, puede iniciar uno o más MUXes para quedarse con los IIPs del MUX fallido. En el mismo momento, el monitor de salud puede desconectar el MUX fallido. Para gestionar tiempo muerto planificado para MUXes, se pueden utilizar técnicas similares a las usadas para una pausa en caliente. Debido a que los MUXes funcionan es un modo sin estado, algunas implementaciones pueden desconectar con seguridad un MUX después extraer de él todos los paquetes.

En al menos una realización, el tiempo muerto planificado del DLB puede ser gestionado a través de una transición de mapa de MUX con seguimiento de estado.

1. MUX está utilizando un MapaVip (V) que utiliza DLB (D).

2. MUX recibe la notificación de que un DLB (D) va a ser desactivado en tiempo T;

3. MUX calcula un nuevo MapaVip (V') que no utiliza DLB (D).

4. MUX pone el activador en (V->V' modo de transición);

5. En la transición, se conserva una tabla de estado y cada TCP SYN causará una nueva entrada en la tabla;

a. Si un paquete concuerda con una entrada en la tabla de estado, entonces se trata de un flujo nuevo y por tanto usa V';

b. En caso contrario, se utiliza la V antigua;

Nótese: durante este periodo de transición, cualquier nuevo flujo cambiará al nuevo MapaVip (V'), evitando DLB (D).

6. DLB (D) sigue contando el número de conexiones de TCP activas (con VIP). Cuando el contador llega a cero, notifica al MUX que la transición se ha completado.

7. Alternativamente, el MUX puede identificar conexiones de larga duración como conexiones que no concuerdan con ninguna entrada de la tabla de estado.

8. Cuando se alcanza el tiempo T, se fuerza la transición V->V'. El MUX enviará todo el tráfico basado en V'.

En una realización, el tiempo muerto planificado de MUX es gestionado a través de los pasos siguientes:

1. Establecer el MapaVip en el nuevo MUX (M');

2. Establecer el antiguo MUX (M) para dirigir todo el tráfico VIP hacia M', M' dirigirá el tráfico a los DLBs según lo habitual;

3. Eliminar el IIP del antiguo MUX (M);

4. Incorporar el IIP al nuevo MUX (M'); y,

5. El encaminador debe iniciar el envío hacia el nuevo MUX.

En al menos una realización, el monitor de salud 608 puede enviar sondas periódicas a MUXes y DLBs para supervisar fallos inesperados. Cuando se detecta un fallo de DLB, el monitor de salud puede ordenar al MUX que actualice su MapaVip para evitar el uso del DLB fallido. Cuando se detecta un fallo de MUX, el monitor de salud puede ordenar a otro MUX en la misma VLAN que instale el IIP (y use G-ARP para anunciar al encaminador). En al menos una realización, el monitor de salud puede emitir sondas de MantenerActivo cada dos segundos, y anunciar la muerte del MUX/DLB después de 3 fallos consecutivos.

Para lograr una conmutación por error de MUX rápida (<< 1 segundo) para VIPs en misión crítica, se puede utilizar un grupo virtual de MUXes para cada IIP. El coste de esta conmutación por error rápida puede ser más uso de red durante funcionamiento normal. Se pueden utilizar los siguientes pasos para gestionar los MUXes y el IIP para el VIP:

A. Cada IIP puede ser una dirección de multidifusión. Cada VIP tiene un grupo de MUXes asignados al mismo.

B. El MUX principal del grupo es el portador real del IIP.

C. El MUX maestro emite un anuncio multidifusión de que él es el MUX activo para este VIP a todos los miembros en este grupo. Este anuncio es emitido a alta velocidad (<< 1 segundo). Este anuncio evita asimismo que otros MUXs inicien el proceso de elección del nuevo maestro.

D. Debido a que el IIP puede ser una dirección multidifusión, el encaminador de aguas arriba replica cada paquete que ha recibido hacia los MUX miembros (el maestro y todos los secundarios) del grupo de VIP.

E. El MUX de apoyo designado almacena los paquetes durante un tiempo T especificado.

F. El MUX maestro realiza la función de igualación de carga sobre los paquetes y los dirige a los DLBs.

G. Si no se recibe ningún anuncio de que el MUX-Maestro-Está-Operativo durante un tiempo T dado, el MUX de apoyo designado comenzará a igualar la carga y a dirigir todos los paquetes en su memoria provisional.

H. Los apoyos en este grupo iniciarán un proceso de elección del nuevo maestro. En algunas configuraciones, el MUX de apoyo designado puede convertirse en el nuevo maestro.

I. El paso G puede hacer que el DLB reciba algunos paquetes dos veces, pero TCP tolera paquetes duplicados y pérdidas transitorias de paquetes suficientemente bien. Nótese que una pérdida de paquetes puede no ocurrir siempre que el encaminador de aguas arriba esté operativo y funcionando correctamente.

La Figura 7 representa una configuración ejemplar de un MUX 212(1) (presentado en la Figura 2) según una o más realizaciones. Tomadas en conjunto, las Figuras 7 y 8 representan cómo los paquetes pueden ser encapsulados y desencapsulados a lo largo de un camino.

La Figura 7 comprende el modo de usuario 702 y el modo kernel 704, pero se concentra en la funcionalidad proporcionada por el activador 618 de MUX de los MUX's en el modo kernel. En este caso, el activador de MUX está implementado a modo de extensión de la capa de IP de la pila de redes.

En este ejemplo, un paquete 706 es recibido por el activador 618 de MUX, tal como procedente de un servidor de aplicación. El paquete comprende una dirección de cliente fuente en 708 y una dirección VIP de destino en 710. El

paquete migra hacia la capa 712 de Tarjeta de Interfaz de Red (NIC) física, y la capa 714 de Especificación de Interfaz de Activador de Red (NDIS). El paquete es gestionado por el enviador 716 del activador del MUX en la capa de IP 718. El enviador encapsula el paquete 706 para generar el paquete 720. Este paquete comprende la dirección de cliente fuente en 708 y la dirección de VIP de destino en 710 encapsuladas mediante una dirección 722 de MUX fuente y una dirección 724 de DIP de destino. De ese modo, el paquete 706 original es encapsulado en el paquete 720 de modo que da la impresión que procede del MUX 212(1) en lugar del cliente 708.

El MUX 212(1) puede implementar una igualación de carga de capa-4, también conocido como mapeo VIP:DIP. El tráfico procedente de los clientes puede ser enviado hacia uno de los nodos de MUX por parte del nivel 1 (típicamente a través de encaminamiento Multi Camino de Coste Equivalente (ECMP)). Cuando el MUX 212(1) recibe un paquete 706, puede aleatorizar los campos de cabecera de paquete (es flexible en términos de qué campos son aleatorizados) y puede coger un DIP basado en esta aleatorización. (Un ejemplo de este proceso se describe a continuación en relación a la Figura 9). El MUX puede a continuación encapsular el paquete 706 original en una nueva cabecera de IP que indica al DIP elegido como el destino (es decir, dirección 724 de DIP de destino) y al MUX como la dirección de IP de fuente. (De modo alternativo, el MUX puede utilizar el emisor original a modo de fuente de IP).

Los nodos MUX en una agrupación de igualación de carga pueden utilizar la misma función de aleatorización. Adicionalmente, los nodos de MUX pueden mantener el estado durante la incorporación y la adecuada eliminación de los DIPs. Esto puede permitir que los paquetes de un flujo determinado sean dirigidos hacia el mismo servidor en el siguiente nivel independientemente de qué MUX recibe el paquete.

La Figura 8 representa un ejemplo del rol de DIP 606 presentado anteriormente en relación con la Figura 6. Brevemente, en este caso, el activador 626 de desencapsulación de DIP puede realizar desencapsulación del paquete 720 encapsulado presentado en la Figura 7. En esta configuración, el activador de desencapsulación de DIP está implementado como una extensión de la capa de IP de la pila de redes. Tal como se ha citado anteriormente, la Figura 7 ofrece un ejemplo para realizar encapsulación en el extremo frontal de un camino de transmisión, la Figura 8 ofrece un ejemplo en el extremo trasero de desencapsulación del paquete 706 original presentado anteriormente.

En este ejemplo, el activador 626 de desencapsulación puede recibir el paquete 720 encapsulado. El activador de desencapsulación puede eliminar la encapsulación (es decir, la dirección 722 de MUX fuente y la dirección 724 de DIP de destino) para producir el paquete 706 una vez que el paquete encapsulado viaja a lo largo del camino y está listo para su transmisión hacia la dirección 710 de VIP de destino.

El MUX 212(1) y el rol de DIP 606 descritos anteriormente pueden ser utilizados con los conceptos presentes para facilitar la encapsulación de un paquete, tal como el paquete 706 que está asociado con una dirección de aplicación (es decir, la dirección 710 de VIP de destino), con una dirección de ubicación (es decir, la dirección 724 de DIP de destino) de manera que el paquete 706 puede ser transportado sobre una infraestructura de capa-3 y ser entregado en última instancia a la dirección 710 de VIP de destino de capa-2. Además, el paquete encapsulado puede viajar a través del camino seleccionado definido mediante encapsulación y el camino seleccionado puede ser fácilmente reseleccionado para paquetes subsiguientes para evitar congestión.

Además, esta configuración puede facilitar crecimiento y contracción libres de interrupciones (o con interrupciones reducidas) de un grupo de nodos de red (es decir, componentes de sistemas de igualación de carga escalables 104, 204 y/o 304). Brevemente, el estado de un sistema de igualación de carga escalable tiende a no ser estático. Por ejemplo, más servidores de aplicación pueden conectarse a la red y/o ciertos servidores de aplicación pueden desconectarse de la red, ciertos conmutadores pueden entrar o salir, las comunicaciones son iniciadas y finalizadas, etc. Los conceptos presentes pueden permitir una transición suave desde un mapeo de sistema de igualación de carga escalable a un nuevo mapeo de sistema de igualación de carga escalable. Por ejemplo, los conceptos presentes pueden rastrear las comunicaciones existentes o activas de un mapeo existente. Algunas implementaciones pueden intentar mantener la continuidad para aquellas comunicaciones activas que utilizan el mapeo existente mientras que utilizan un nuevo mapeo que refleja los cambios del sistema de igualación de carga escalable para nuevas comunicaciones. Estas implementaciones pueden de ese modo realizar una transición “suave” desde el mapeo antiguo hacia el nuevo mapeo en un modo que es relativamente imperceptible.

La Figura 9 representa un método ejemplar método ejemplar 900 de mapeo de un espacio de aleatorización a un grupo de DIP. Por ejemplo, el mapeo puede permitir la eliminación de un DIP de un grupo de VIP sin interrupción del tráfico que no va hacia el DIP afectado. Por ejemplo, un primer mapeo entre un espacio de aleatorización (es decir, valores de aleatorización potenciales) y un grupo de DIPs disponibles se representa en 902. Un segundo mapeo entre el espacio de aleatorización y un grupo diferente de DIPs disponibles se representa en 904. En este caso, el segundo mapeo 904 ocurrió como resultado de la desactivación de una DIP 1 (es decir, quedando no disponible) tal como se ha indicado en 906. Observando inicialmente al primer mapeo primer mapeo 902, los valores de aleatorización son mapeados a DIP 1 en 908(1), 908(2) y 908(3), a DIP 2 en 910(1), 910(2) y 910(3), a DIP 3 en 912(1), 912(2) y 912(3), y a DIP 4 en 914(1), 914(2) y 914(3). De ese modo, los valores de aleatorización son distribuidos entre los DIPs disponibles en un modo que pueden reducir o evitar los cuellos de botella.

Con la pérdida de DIP 1 en 906, esta implementación redistribuye la carga de DIP 1 entre los restantes DIP disponible de un modo que evita la sobrecarga repentina de cualquier DIP individual disponible. Por ejemplo, en el segundo mapeo 904, la primera parte de la aleatorización que fue mapeada a DIP 1 en 908(1) en el primer mapeo 902 es reasignada a DIP 2 tal como se indica en 916. La segunda parte de DIP 1, la 908(2), es reasignada a DIP 3 tal como se indica en 918. La tercera parte de DIP 1, 908(3), es reasignada a DIP 4 tal como se indica en 920. De ese modo, esta implementación redistribuye de forma imperceptible el flujo de paquetes desde una distribución de 4 vías como se observa en el primer mapeo 902 a una distribución de 3 vías como se observa en el segundo mapeo 904 según un modo equitativo que puede evitar la sobrecarga de cualquiera de los DIPs restantes, y por tanto evita la potencial creación de un cuello de botella asociado a cualquier DIP sobrecargado.

Con fines de una explicación más detallada, considérese un MUX (tal como el MUX 212(1) que tiene un mapa M de VIP-DIP que determina el mapeo de un VIP respecto uno o más servidores de aplicación (DLBs). Ahora considérese un escenario donde M debe ser cambiado a M'. Utilizando las técnicas descritas, M puede ser cambiada a M' suavemente. Debido a que podrían existir conexiones de han estado largamente en marcha, se puede definir de modo opcional un tiempo límite T. El MUX puede a continuación cambiar de M a M' una vez que se ha alcanzado T o cuando el cambio suave ha sido completado.

A continuación se describe un ejemplo de una forma de cambiar suavemente de M a M':

Para un paquete P, el MUX puede calcular tanto H(P) como H'(P), donde H(P) puede ser calculado utilizando el mapa M y H'(P) puede ser calculado usando el mapa M'.

- si H(P)=H'(P), dirigir hacia H(P) es equivalente a dirigir hacia H'(P);

- si H(P)!=H'(P) y P es SYN (paquete TCP SYN, que puede iniciar una conexión de TCP), P puede ser utilizado para establecer una conexión nueva, que debería ir hacia H'(P), además el inserto aleatorizar(P) → H'(P) puede ser insertado dentro de una tabla de estado S de manera que ese flujo puede ser reconocido como que ha sido desplazado a M';

- si H(P)!=H'(P) y P no es SYN, y aleatorizar(P) no está en S, esto puede ser parte de una conexión activa hacia H(P), así que debe continuar hacia H(P);

- si H(P)!=H'(P) y P no es SYN, y aleatorizar(P) está en S, esto puede ser parte de una conexión activa que ha sido ya desplazada a M', así que debe continuar hacia H'(P);

- cuando se alcanza T o todos los DLBs informan que la transmisión se ha realizado, el mapeo puede ser cambiado de M a M', y la tabla de estado S puede ser limpiada.

Correspondientemente, el DLB puede ser informado acerca de la misma transición M -> M', y a continuación puede calcular si él (es decir, el DLB) está afectado por esta transmisión.

Si un DLB decide que está siendo echado durante la transición, puede agotar suavemente la conexión que tiene.

Para una conexión de HTTP persistente, el servidor de DLB HTTP puede desactivar el 'MantenerActivo HTTP'. Como tal, el servidor de DLB HTTP puede terminar la conexión de TCP subyacente con un FIN (paquete de TCP FIN, que completa una conexión de TCP). El FIN puede ser concebido como un indicador en la cabecera de TCP que indica que el emisor de este paquete desea terminar la conexión. Un cliente externo puede reiniciar una conexión. Sin embargo, esto probablemente iniciará un nuevo procedimiento de enlace, por lo cual el MUX puede encaminar la nueva conexión de TCP hacia el nuevo DLB.

De forma alternativa, las conexiones de HTTP persistentes pueden ser gestionadas como las conexiones de TCP establecidas descritas a continuación.

- Una conexión de TCP establecida puede permanecer estar inactiva u ocupada durante el periodo de transición, y se podría esperar que HTTP la cerrara. Algunas acciones potenciales son:

1. Permitir que la conexión de TCP expire en el lado del cliente. Básicamente esta técnica simplemente ignora estas conexiones de TCP.

2. Forzar la transmisión de un TCP RST al cliente cuando se ha alcanzado el tiempo T de manera que el cliente es informado. Enviar un RST no requiere que se disponga del número de secuencia correcto. Como tal, esta técnica puede simplemente enumerar a través de las conexiones “establecidas” y matar todas las conexiones establecidas.

3. El MUX puede mantener el estado para conexiones persistentes hasta que los DLBs determinan que las conexiones afectadas por la transición han terminado.

- Cuando el número de conectores de TCP abiertos es cero, el MUX puede ser informado de que el nodo puede ser eliminado de forma segura del grupo.

En resumen, las presentes implementaciones pueden utilizar encapsulación de IP-en-IP de manera que se puede utilizar DSR a través de potencialmente todos los dispositivos objetivo en lugar de sólo de una sub-red. Además, los igualadores de carga pueden ser implementados a modo de niveles lógicos escalables como se desee. Los conceptos pueden conservar además las conexiones durante las transiciones del sistema. Por ejemplo, los DIPs pueden ser incorporados o eliminados, las cargas pueden ser re-quilibradas y/o la capacidad del sistema puede ser ajustada mientras se realizan suavemente conexiones en transición. Se puede alcanzar una aleatorización coherente en la capa de MUX para permitir la escalabilidad y para permitir la eliminación de DIPs fallidos sin mantener el estado. Además, la supervisión del sistema, el control, y/o las funciones de gestión, pueden ser recolocadas junto a funciones de igualación de carga. Esto puede permitir a un maestro asegurar la continuidad de las direcciones entre los MUXes, entre otras ventajas potenciales.

Primer método ejemplar

La Figura 10 representa un diagrama de flujo de un método 1000 que describe los pasos o acciones de un ejemplo asociado con conservación de conexiones de larga duración con respecto a la expansión de un grupo de DIP para un VIP, según una o más realizaciones.

El método puede ser implementado en conexión con cualquier hardware, software apropiado (por ejemplo, incluyendo firmware), o cualquier combinación de los mismos. En algunos casos, el método puede estar almacenado en un medio de almacenamiento legible por un ordenador que puede ser ejecutado por un procesador de un dispositivo de cálculo para realizar el método. Adicionalmente, uno o más pasos del método pueden ser repetidos cualquier número de veces. Adicional o alternativamente, uno o más pasos pueden ser omitidos en al menos algunas realizaciones.

En el paso 1002, se identifican nuevas conexiones para una red o un sistema de igualación de carga escalable. En al menos algunas realizaciones, esto se puede lograr mediante la búsqueda de TCP SYN.

En el paso 1004, el estado es conservado para las nuevas conexiones.

En el paso 1006, se utiliza una aleatorización antigua o existente para conexiones antiguas o existentes, y se puede utilizar una aleatorización nueva para nuevas conexiones.

En el paso 1008, se consultan a los DIPs. En al menos algunas realizaciones, esto puede incluir consultar DIPS para conexiones de larga duración que han de ser conservadas. De forma alternativa, el sistema de igualación de carga puede determinar conexiones activas mediante la interpretación de las cabeceras de paquete.

En el paso 1010, el estado para las nuevas conexiones ha expirado.

En el paso 1012, el estado para las conexiones conservadas ha expirado. En al menos algunas realizaciones, esto puede incluir expirar el estado de conexiones conservadas a medida que terminan en los DIPs.

Se ofrece el método 1000 por razones explicativas y no debería ser considerado de forma limitante. Por ejemplo, un método alternativo que se puede utilizar durante una transición puede utilizar el siguiente algoritmo:

1. Identificar un paquete de inicio de nueva conexión mediante la interpretación de las cabeceras de paquete (por ejemplo, buscar el TCP SYN);

2. Si se trata de un paquete de inicio de nueva conexión, transmitirlo sólo de acuerdo con el mapa nuevo;

3. Enviar también el paquete según ambos, mapa antiguo y mapa nuevo;

4. Identificar las conexiones antiguas mediante la consulta de los DIPs o rastreando el estado en los igualadores de carga durante un cierto periodo de tiempo;

5. Enviar las conexiones antiguas en función del mapa antiguo y las nuevas conexiones en función del mapa nuevo; y,

6. Estado de expiración de las conexiones antiguas después de la expiración de un tiempo límite o cuando terminan en las DIPs.

Segundo método ejemplar

La Figura 11 representa un diagrama de flujo que describe los pasos o las acciones de un método ejemplar 1100. El método puede ser implementado en conexión con cualquier hardware, software disponible (por ejemplo, incluyendo firmware), o cualquier combinación de los mismos. En algunos casos, el método puede estar almacenado en un medio de almacenamiento legible por un ordenador que puede ser ejecutado por un procesador de un dispositivo de cálculo para realizar el método. Adicionalmente, uno o más pasos del método pueden ser repetidos cualquier número de veces. Adicional o alternativamente, uno o más pasos pueden ser omitidos en al menos algunas realizaciones.

En el paso 1102, los paquetes de red pueden ser difundidos entre una serie de módulos. En al menos una realización, los módulos son módulos de MUX configurados para ser implementados sobre servidores y/o encaminadores. La difusión puede ser independiente de las características individuales de los paquetes, excepto que paquetes que van a destino pueden ser entregados a un módulo de MUX que contenga el estado necesario para gestionar paquetes para ese destino. En al menos algunas realizaciones, los paquetes de red individuales son difundidos entre módulos que utilizan un encaminador ECMP.

En el paso 1104, los paquetes de red pueden ser encapsulados en módulos individuales. En al menos algunas realizaciones la encapsulación del paquete comprende una encapsulación de IP-en-IP y/o conserva una o más direcciones de VIP hacia las cuales el paquete fue transmitido. En cuanto a eso, deberá notarse que una característica potencialmente valiosa de las técnicas descritas en la presente memoria está asociada con la encapsulación de paquetes de red en base a características de los paquetes (por ejemplo, el quinteto de dirección de fuente de IP, dirección de IP de destino, número de Protocolo IP, puerto de fuente de TCP y/o puerto de destino de TCP) de manera que los paquetes que forman parte de la misma petición pueden, en algunas realizaciones, ser todos ellos gestionados por el mismo dispositivo objetivo, independientemente de qué módulo MUX encapsula al paquete.

En el paso 1106, se puede escoger un dispositivo objetivo para el cual se encapsulan los paquetes de red que utilizan un estado compartido entre los módulos. En al menos algunas realizaciones, el estado compartido entre los módulos es un espacio clave de una función de aleatorización coherente. Adicional o alternativamente, en al menos algunas realizaciones, el estado compartido entre los módulos puede ser cambiado en respuesta al fallo del dispositivo objetivo.

En el paso 1108, los paquetes de red pueden ser enviados desde los módulos.

En el paso 1110, se puede supervisar la salud de los dispositivos objetivo, los módulos de MUX, los encaminadores y las rutas entre varios componentes.

Conclusión

Aunque las técnicas, métodos, dispositivos y sistemas, etc., pertenecientes a escenarios de igualación de carga son descritos en un lenguaje específico respecto a características estructurales y/o acciones metodológicas, debe entenderse que el objeto en cuestión definido en las reivindicaciones adjuntas no está necesariamente limitado a las características o acciones específicas descritas. Más bien, las características o acciones específicas son descritas como formas ejemplares de implementar los métodos, dispositivos, sistemas, etc. reivindicados.

REIVINDICACIONES

1. Un método (1100) de igualación de carga que comprende:

difundir (1102) paquetes de red entre una serie de módulos (212);

encapsular (1104) un paquete de red individual de un flujo de paquetes en un módulo individual (212) de una capa (108) de igualación de carga de manera que conserva una dirección de fuente de un dispositivo (102, 202) de cliente externo y una dirección destino del paquete de red individual tal como es recibido por el módulo individual;

escoger (1106) un dispositivo objetivo (110, 210) para el cual se ha encapsulado el paquete de red individual para proporcionar una paquete encapsulado que utiliza un estado compartido entre los módulos (212), en el que el paquete encapsulado comprende el paquete de red individual con la dirección de fuente conservada y la dirección de destino conservada y en el que una dirección de fuente del paquete encapsulado se refiere al módulo individual y una dirección de destino del paquete encapsulado se refiere al dispositivo objetivo;

enviar (1108) el paquete encapsulado desde el módulo individual (212) hacia el dispositivo objetivo;

desencapsular el paquete encapsulado mediante un componente (222) de desencapsulación del dispositivo objetivo (110, 210); y,

encaminar al menos algunos paquetes salientes del flujo de paquetes hacia el dispositivo (102, 202) de cliente externo sin pasar a través de la capa (208) de igualación de carga.

2. Método (1100) según la reivindicación 1, en el que el estado compartido entre los módulos (212) de la serie es un espacio clave de una función de aleatorización coherente.

3. Método (1100) según la reivindicación 1, en el que los paquetes de red individuales son difundidos entre los módulos (212) de la serie utilizando encaminamiento Multi-Camino de Coste Equivalente (ECMP).

4. Método (1100) según la reivindicación 1, que comprende además supervisar (1110) la salud del dispositivo objetivo (110, 210).

5. Método (1100) según la reivindicación 1, en el que el estado compartido entre los módulos (212) de la serie es alterado en respuesta a un fallo del dispositivo objetivo (110, 210).

6. Método (1100) según la reivindicación 1, que comprende además cambiar los módulos (212) activos de la serie por módulos de reserva en base a uno o más de unos parámetros de carga o a uno o más de otros parámetros para no provocar un tiempo muerto o inactivo de un determinado servicio.

7. Método (1100) según la reivindicación 1, en el que el dispositivo objetivo (110, 210) es un miembro de un conjunto de dispositivos objetivo (110, 210), y en el que en un caso en el que se recibe una indicación de que uno o más dispositivos objetivo (110, 210) existentes del conjunto quedarán no disponibles o uno o más nuevos dispositivos objetivo (110, 210) estarán disponibles, pasando a una configuración que difunde paquetes de red asociados con futuras comunicaciones hacia un nuevo conjunto de dispositivos objetivo (110, 210) mientras continua enviado paquetes de red asociados con comunicaciones activas dirigidas hacia los dispositivos objetivo (110, 210) del conjunto.

8. Método (1100) según la reivindicación 1, en el que la encapsulación de los paquetes de red comprende una encapsulación de IP-en-IP.

9. Medio de almacenamiento legible por un ordenador, que comprende instrucciones almacenadas en el mismo que, cuando son ejecutadas por un dispositivo de procesamiento, hacen que el dispositivo de procesamiento realice el método (1100) de cualquier de las reivindicaciones precedentes.

10. Sistema (104, 204) que comprende:

una capa (108) de igualación de carga configurada para encapsular paquetes entrantes individuales de un flujo de paquetes procedentes de un dispositivo (102, 202) de cliente externo para proporcionar paquetes encapsulados, en el que cada paquete encapsulado comprende el paquete entrante individual con la dirección de fuente conservada y la dirección de destino conservada, y el que una dirección de fuente del paquete encapsulado se refiere a la capa de igualación de carga y una dirección de destino del paquete encapsulado se refiere al dispositivo objetivo, estando la capa (108) de igualación de carga configurada adicionalmente para encaminar los paquetes encapsulados hacia los dispositivos objetivo (110, 210) del sistema (104, 204), en el que los dispositivos objetivo (110, 210) se extienden sobre una o más sub-redes de protocolo de Internet (IP), y en el que los paquetes encapsulados pasan a través de uno o más igualadores de carga (208, 216) de la capa (108) de igualación de carga antes de alcanzar a los dispositivos objetivo (110, 210) individuales; y,

en el que los dispositivos objetivo (110, 210) individuales comprenden un componente (222) de desencapsulación

configurado para desencapsular paquetes procedentes de la capa (108) de igualación de carga, y en el que los dispositivos objetivo (110, 210) individuales están configurados para encaminar (222) al menos algunos de los paquetes salientes del flujo de paquetes hacia el dispositivo (102, 202) de cliente externo sin pasar a través de cualquiera de los uno o más igualadores de carga (208).

11. Sistema (104, 204) según la reivindicación 10, en el que la capa (108) de igualación de carga está configurada para encapsular los paquetes entrantes individuales utilizando una o ambas de las opciones de IP de encapsulación de IP-en-IP o modificación de paquetes.

12. Sistema (104, 204) según la reivindicación 11, en el que la capa (108) de igualación de carga comprende al menos un igualador (208) de carga dinámico y al menos un multiplexor (212) y en el que al menos un multiplexor (212) está configurado para encapsular los paquetes entrantes individuales.

13. Sistema (104, 204) según la reivindicación 11, en el que la capa (108) de igualación de carga comprende al menos un multiplexor (212) y en el que al menos un multiplexor (212) está configurado para encapsular los paquetes entrantes individuales utilizando encapsulación de IP-en-IP.

14. Sistema (104, 204) según la reivindicación 11, en el que los dispositivos objetivo (110) individuales abarcan múltiples redes virtuales de área local.

15. Sistema (104, 204) según la reivindicación 11, en el que uno o más de los igualadores de carga (208, 216) comprende igualadores (208) de carga dinámicos que están configurados para proporcionar una interfaz de programa de aplicación para gestionar IP virtual (VIP) para mapeos de IP directo (DIP) de la capa (108) de igualación de carga.

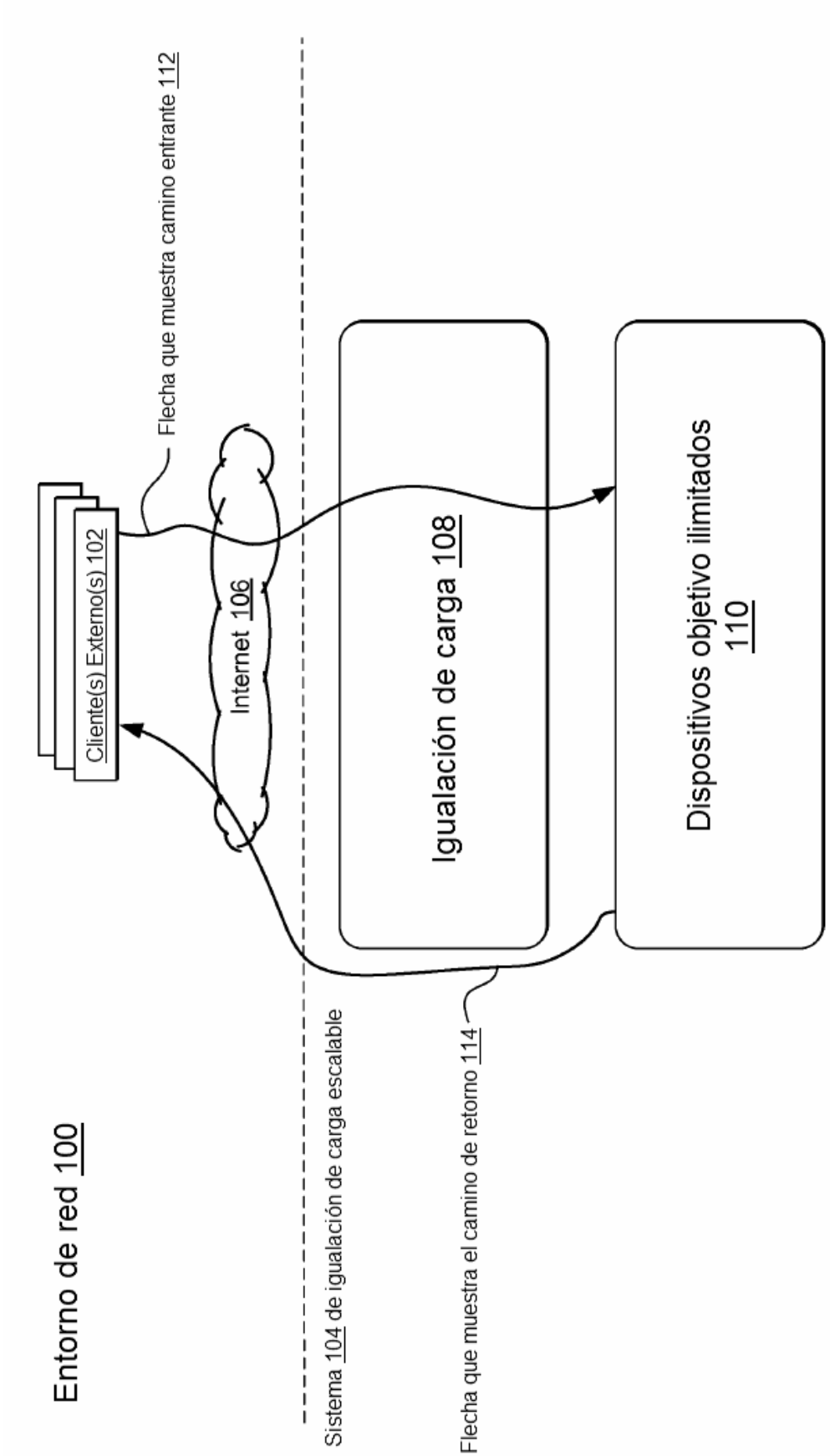
Entorno de red 100
Cliente(s) Externo(s) 102
Flecha que muestra camino entrante 112
Internet 106
Igualación de carga 108
Dispositivos objetivo ilimitados 110
Sistema 104 de igualación de carga escalable
Flecha que muestra el camino de retorno 114

FIG. 1

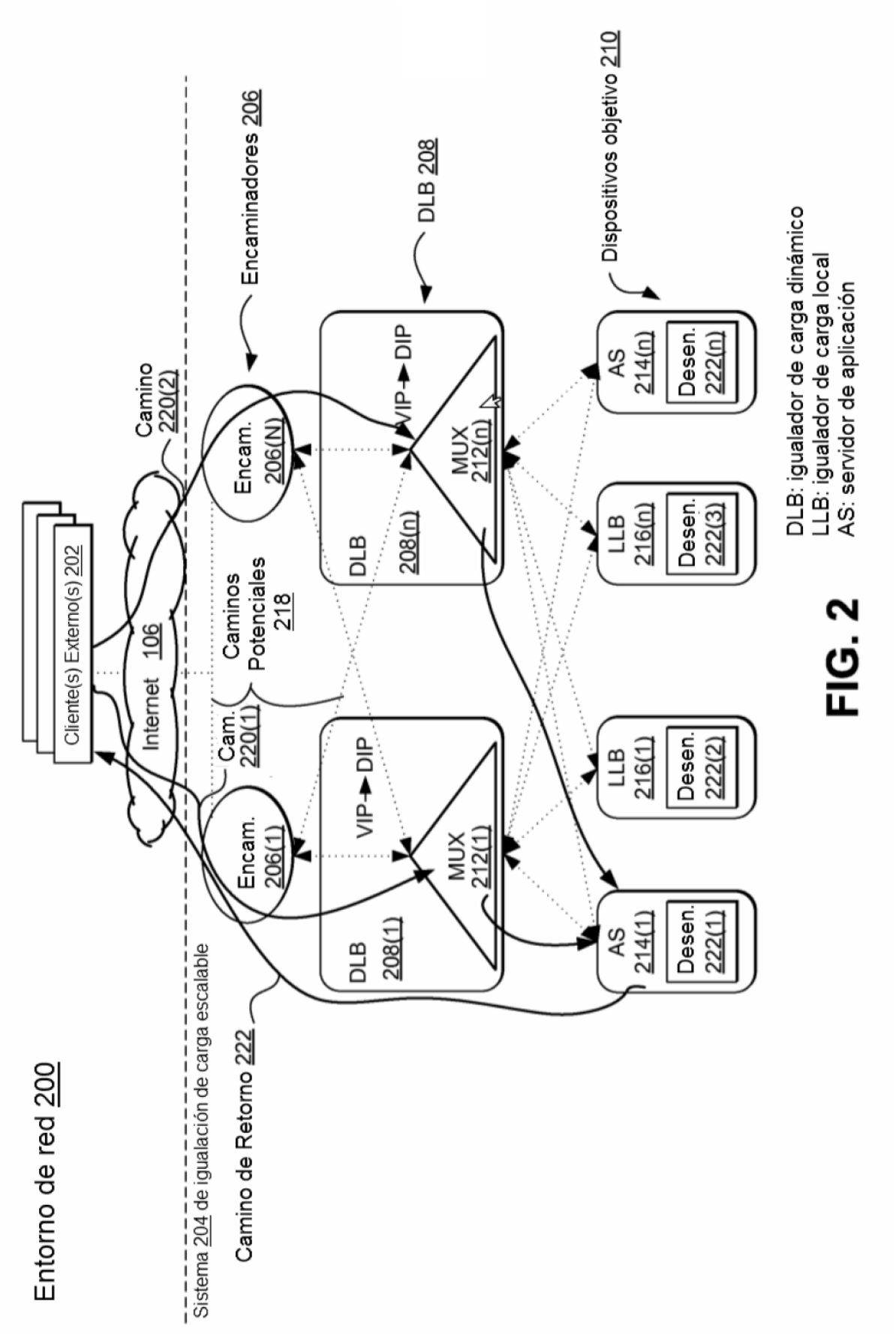
Entorno de red 200
Sistema 204 de igualación de carga escalable
Cliente(s) Externo(s) 202
Internet 106
Camino 220(2)
Encaminadores 206
Encam. 206(N)
Encam. 206(1)
Caminos Potenciales 218
Cam. 220(1)
Camino de Retorno 222
DLB 208
DLB 208(n)
DLB 208(1)
VIP→DIP
MUX 212(n)
MUX 212(1)
Dispositivos objetivo 210
AS 214(n)
AS 214(1)
LLB 216(n)
LLB 216(1)
Desen. 222(n)
Desen. 222(3)
Desen. 222(2)
Desen. 222(1)
DLB: igualador de carga dinámico
LLB: igualador de carga local
AS: servidor de aplicación

FIG. 2

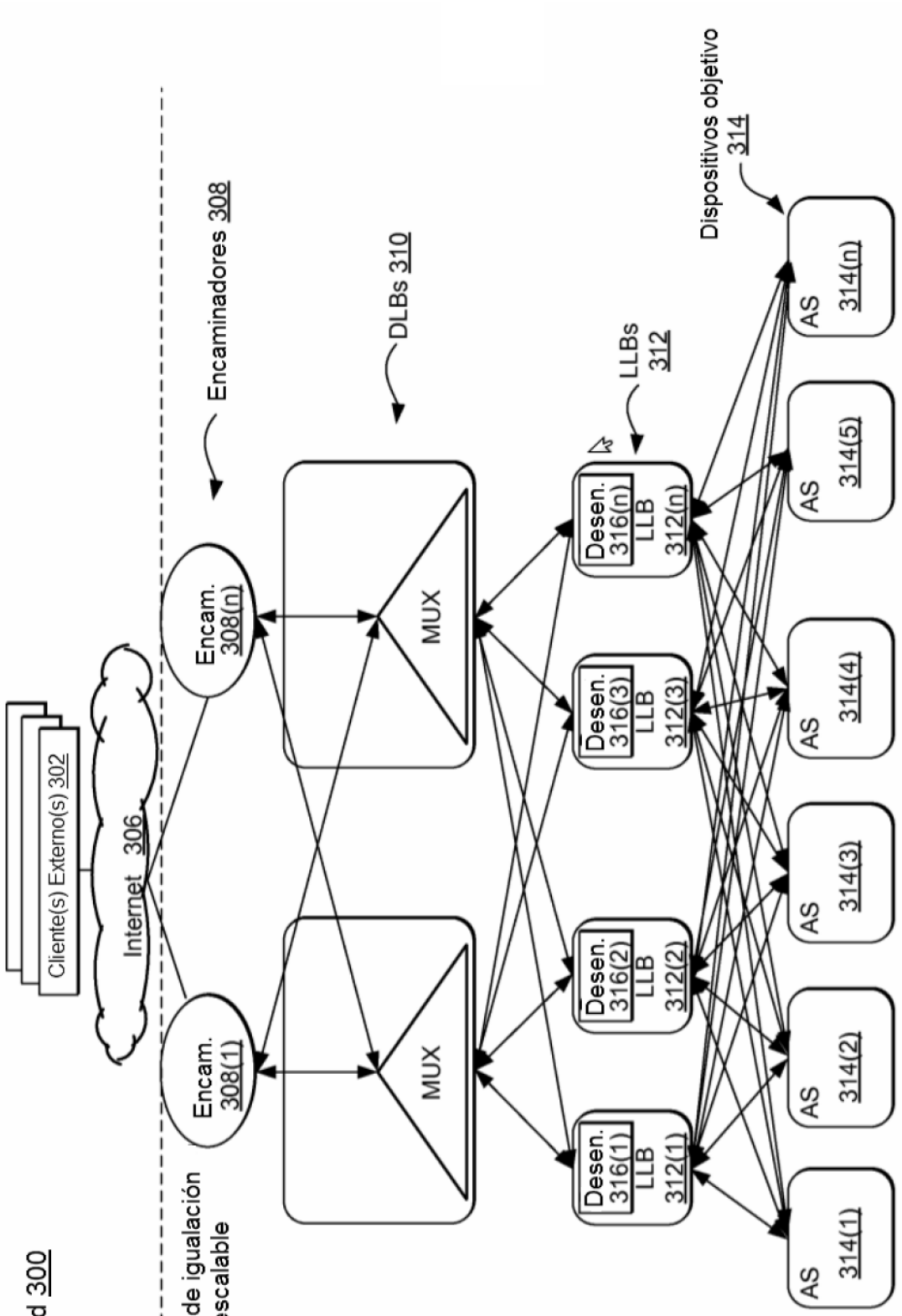
Entorno de red 300
Sistema 304 de igualación de carga escalable
Cliente(s) Externo(s) 302
Internet 306
Encaminadores 308
Encam. 308(1)
Encam. 308(n)
DLBs 310
MUX
MUX
LLBs 312
Desen. 316(1) LLB 312(1)
Desen. 316(2) LLB 312(2)
Desen. 316(3) LLB 312(3)
Desen. 316(n) LLB 312(n)
Dispositivos objetivo 314
AS 314(1)
AS 314(2)
AS 314(3)
AS 314(4)
AS 314(5)
AS 314(n)

FIG. 3

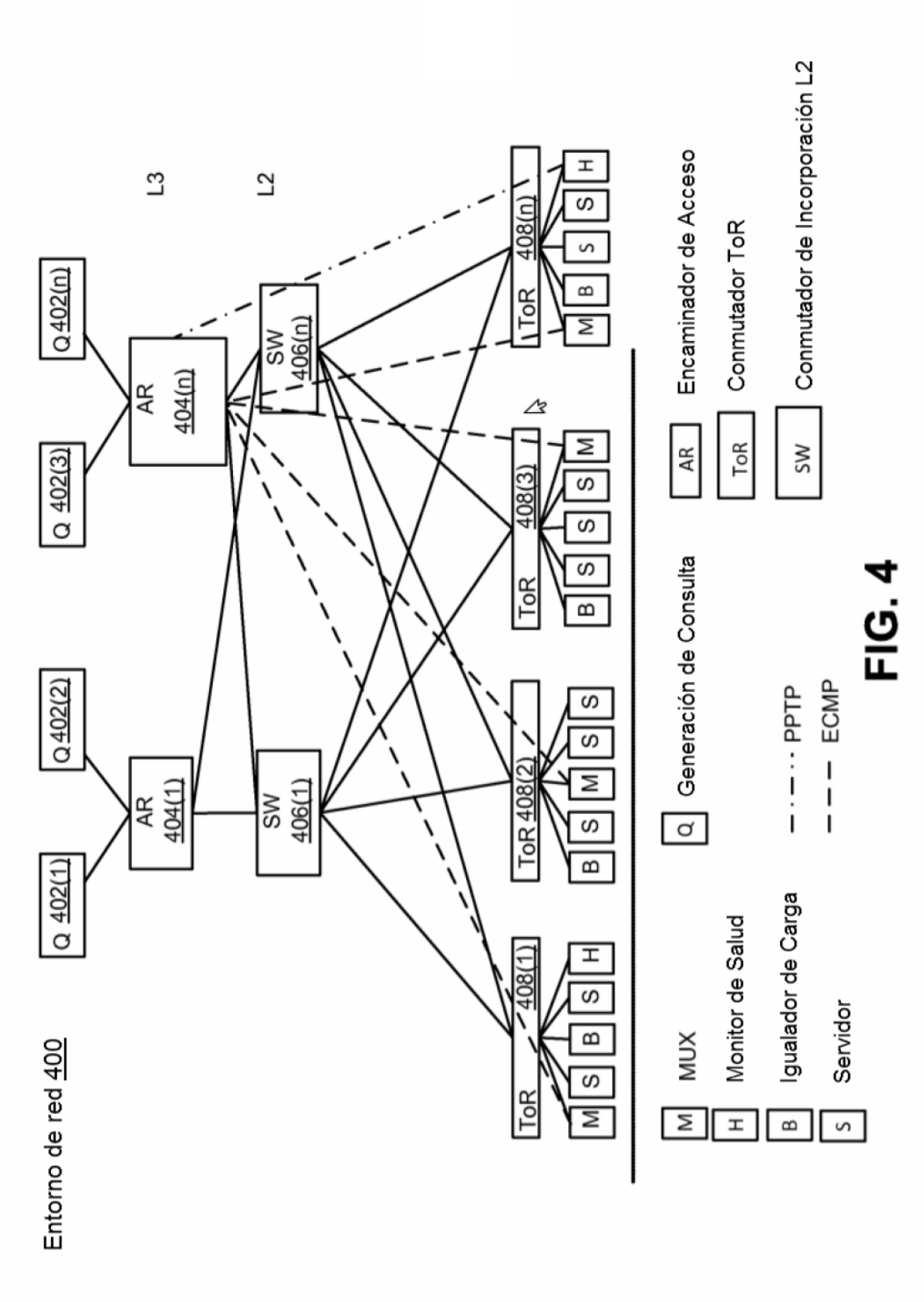
Entorno de red 400
L3
L2
Q 402(1)
Q 402(2)
Q 402(3)
Q 402(n)
AR 404(1)
AR 404(n)
SW 406(1)
SW 406(n)
ToR 408(1)
ToR 408(2)
ToR 408(3)
ToR 408(n)
M MUX
H Monitor de Salud
B Igualador de Carga
S Servidor
Q Generación de Consulta
AR Encaminador de Acceso
ToR Conmutador ToR
SW Conmutador de Incorporación L2
PPTP
ECMP

FIG. 4

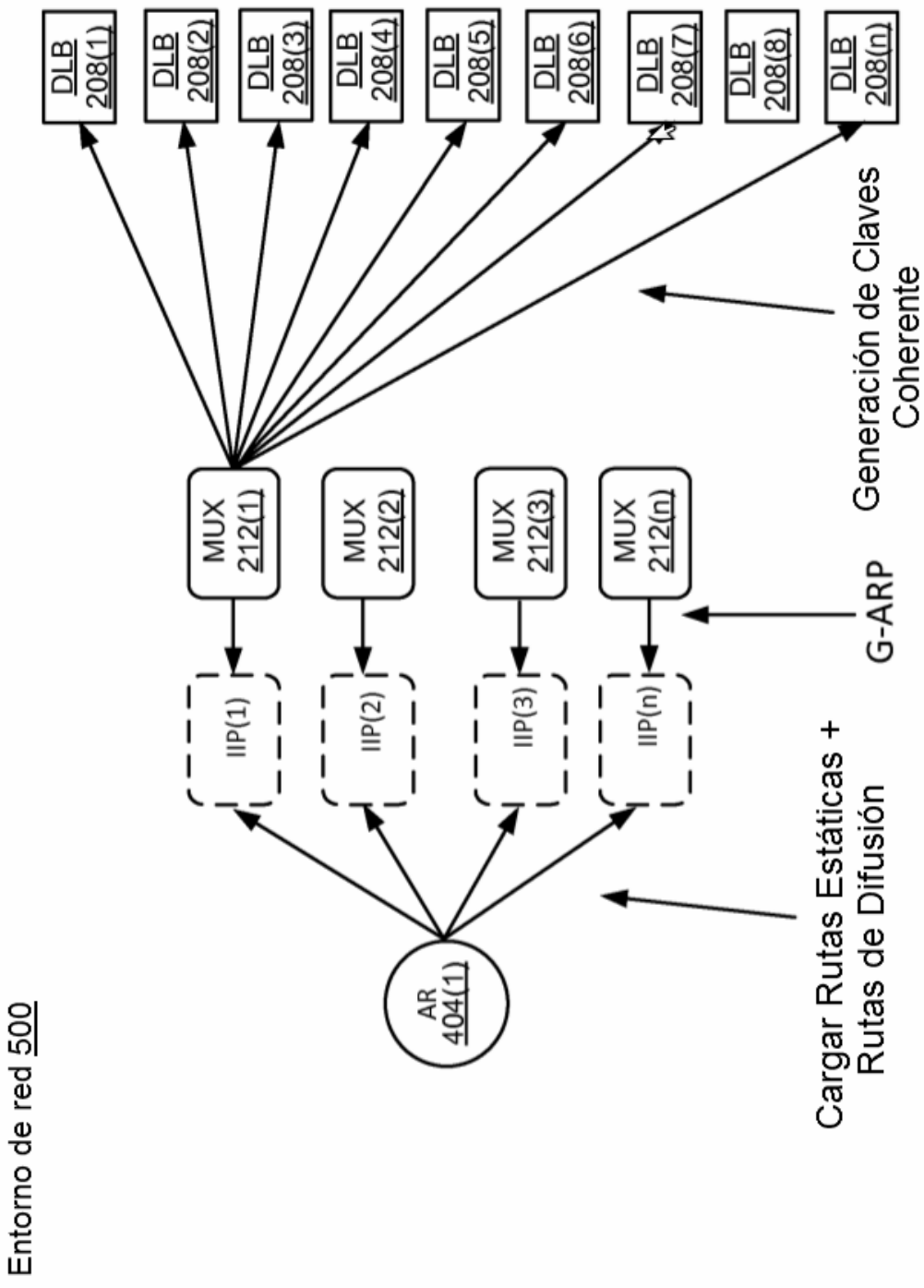
Entorno de red 500
AR 404(1)
IIP(1)
IIP(2)
IIP(3)
IIP(n)
MUX 212(1)
MUX 212(2)
MUX 212(3)
MUX 212(n)
DLB 208(1)
DLB 208(2)
DLB 208(3)
DLB 208(4)
DLB 208(5)
DLB 208(6)
DLB 208(7)
DLB 208(8)
DLB 208(n)
Cargar Rutas Estáticas + Rutas de Difusión
G-ARP
Generación de Claves Coherente

FIG. 5

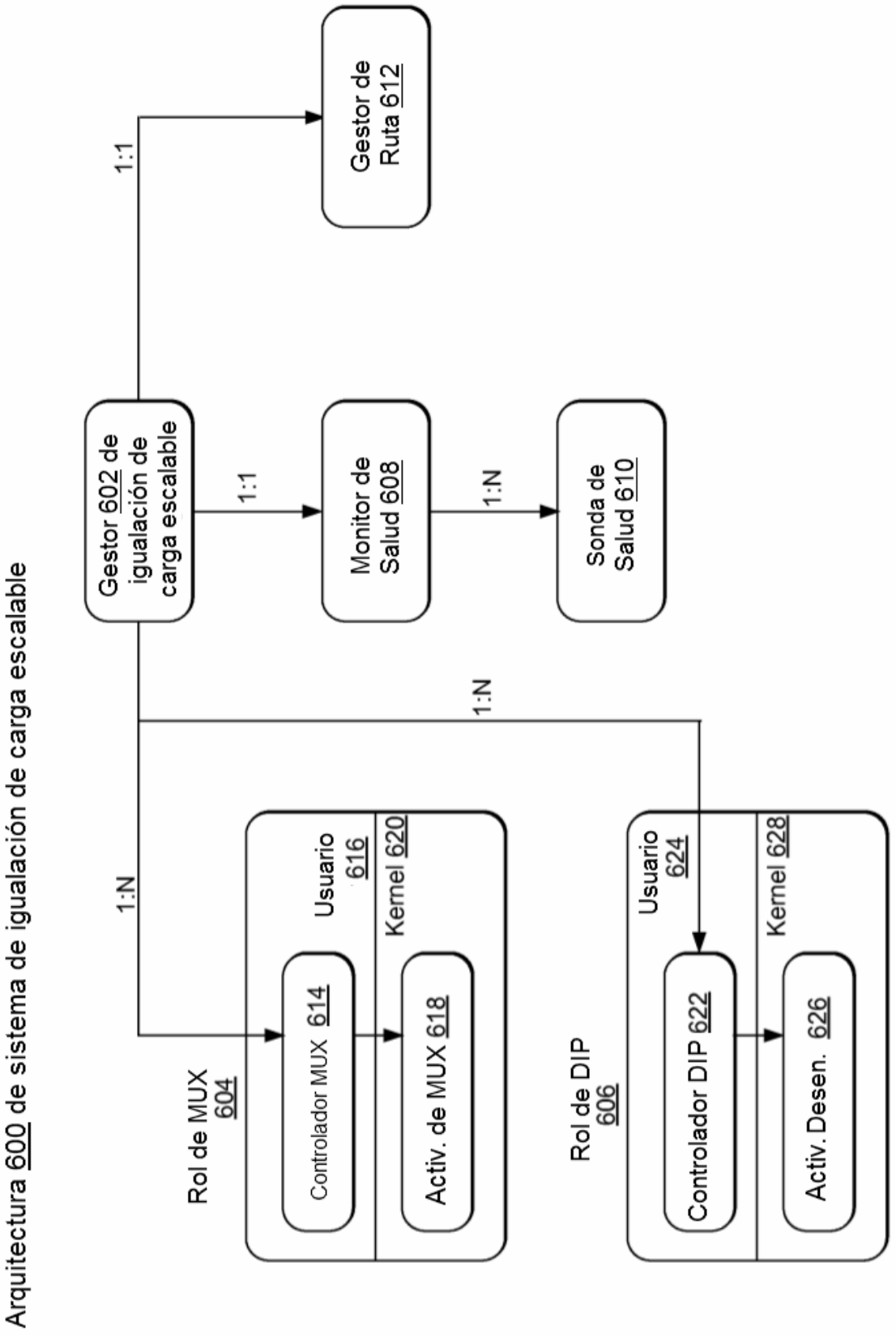
Arquitectura 600 de sistema de igualación de carga escalable
Gestor 602 de igualación de carga escalable
1:1
Gestor de Ruta 612
1:1
Monitor de Salud 608
1:N
Sonda de Salud 610
1:N
1:N
Rol de MUX 604
Usuario 616
Kernel 620
Controlador MUX 614
Activ. de MUX 618
Rol de DIP 606
Usuario 624
Kernel 628
Controlador DIP 622
Activ. Desen. 626

FIG. 6

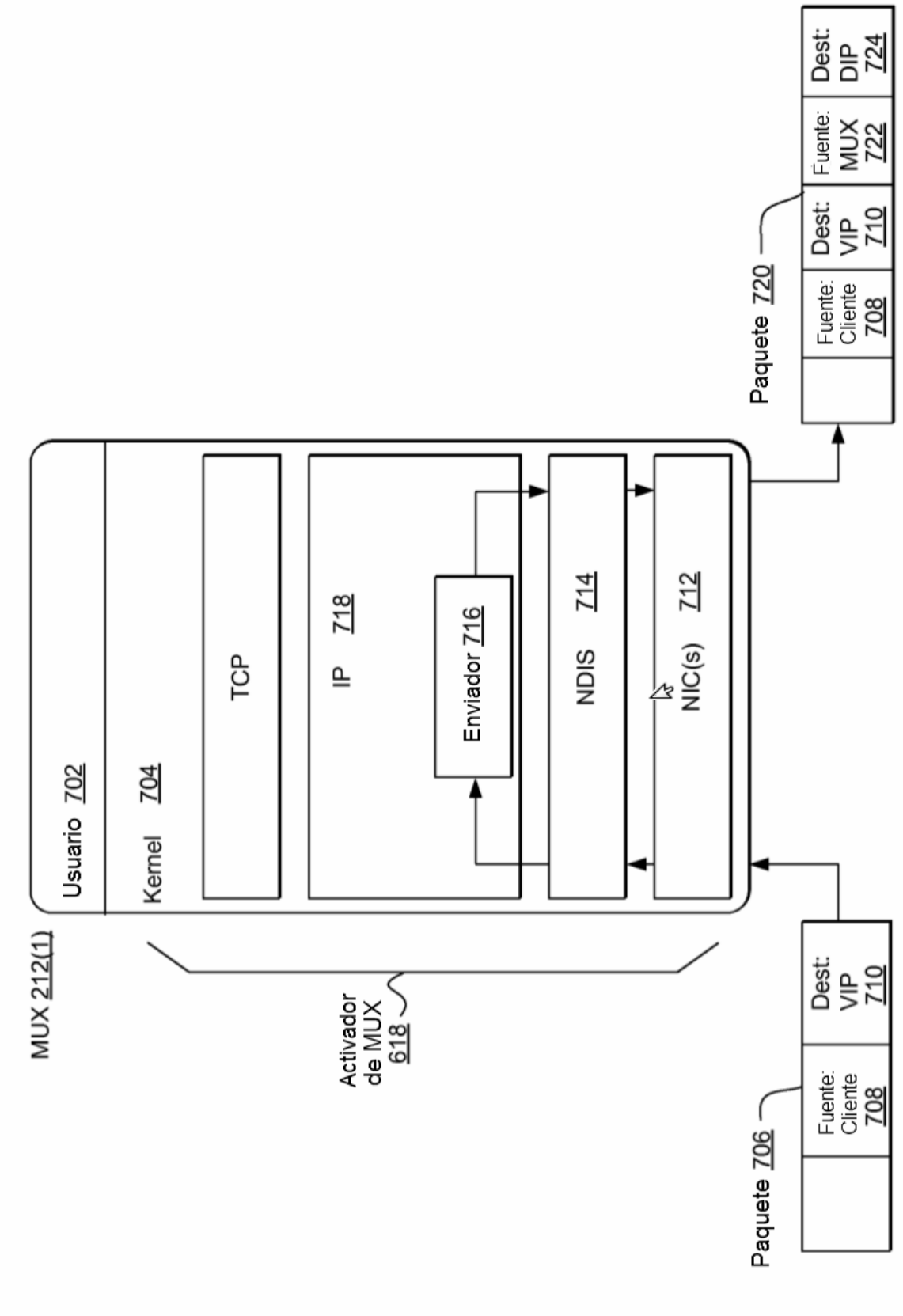
MUX 212(1)
Usuario 702
Kernel 704
TCP
IP 718
Enviador 716
NDIS 714
NIC(s) 712
Activador de MUX 618
Paquete 706
Fuente: Cliente 708
Dest: VIP 710
Paquete 720
Fuente: Cliente 708
Dest: VIP 710
Fuente: MUX 722
Dest: DIP 724

FIG. 7

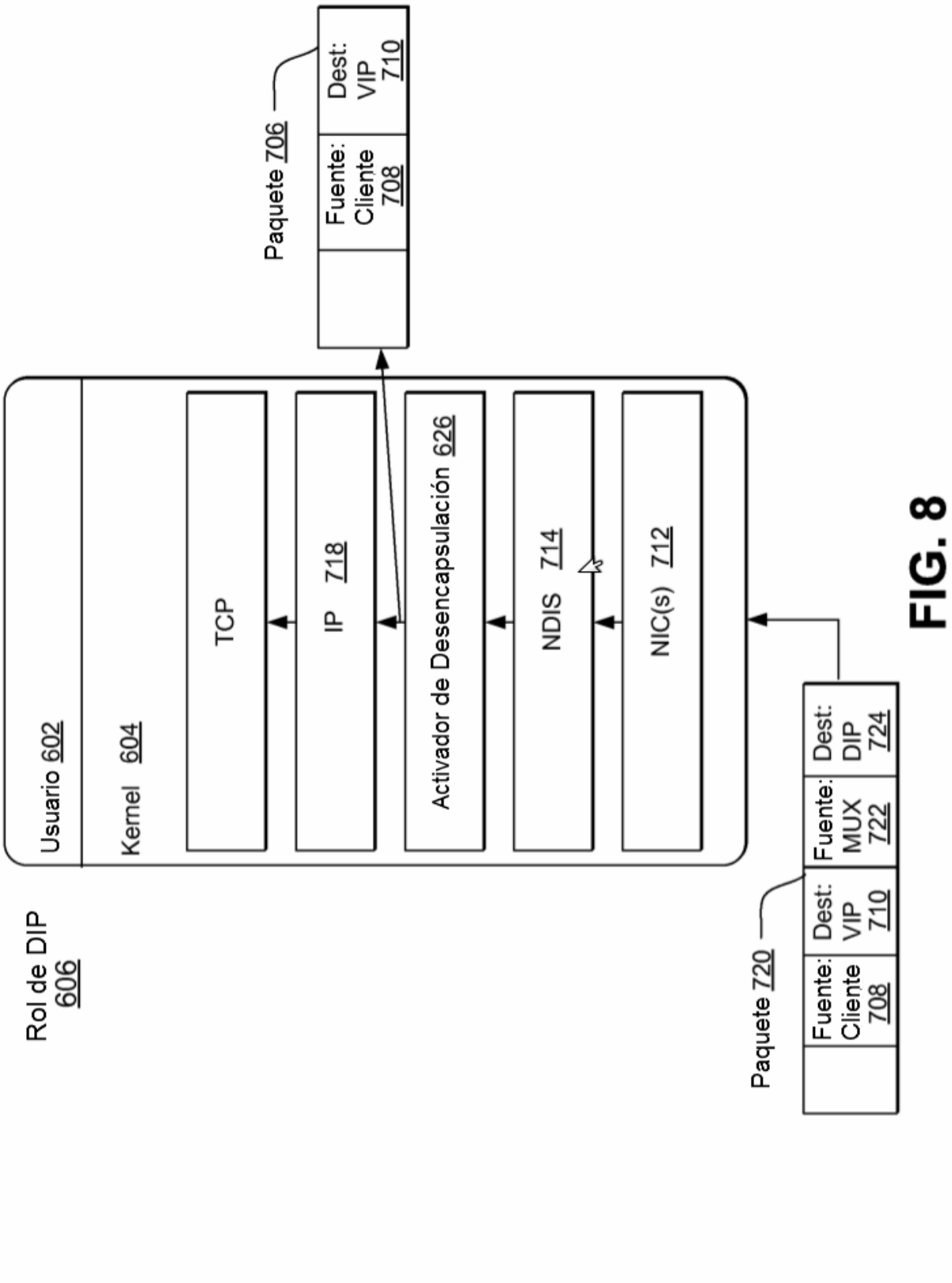
Paquete 706
Fuente: Cliente 708
Dest: VIP 710
Usuario 602
Kernel 604
TCP
IP 718
Activador de Desencapsulación 626
NDIS 714
NIC(s) 712
Rol de DIP 606
Paquete 720
Fuente: Cliente 708
Dest: VIP 710
Fuente: MUX 722
Dest: DIP 724

FIG. 8

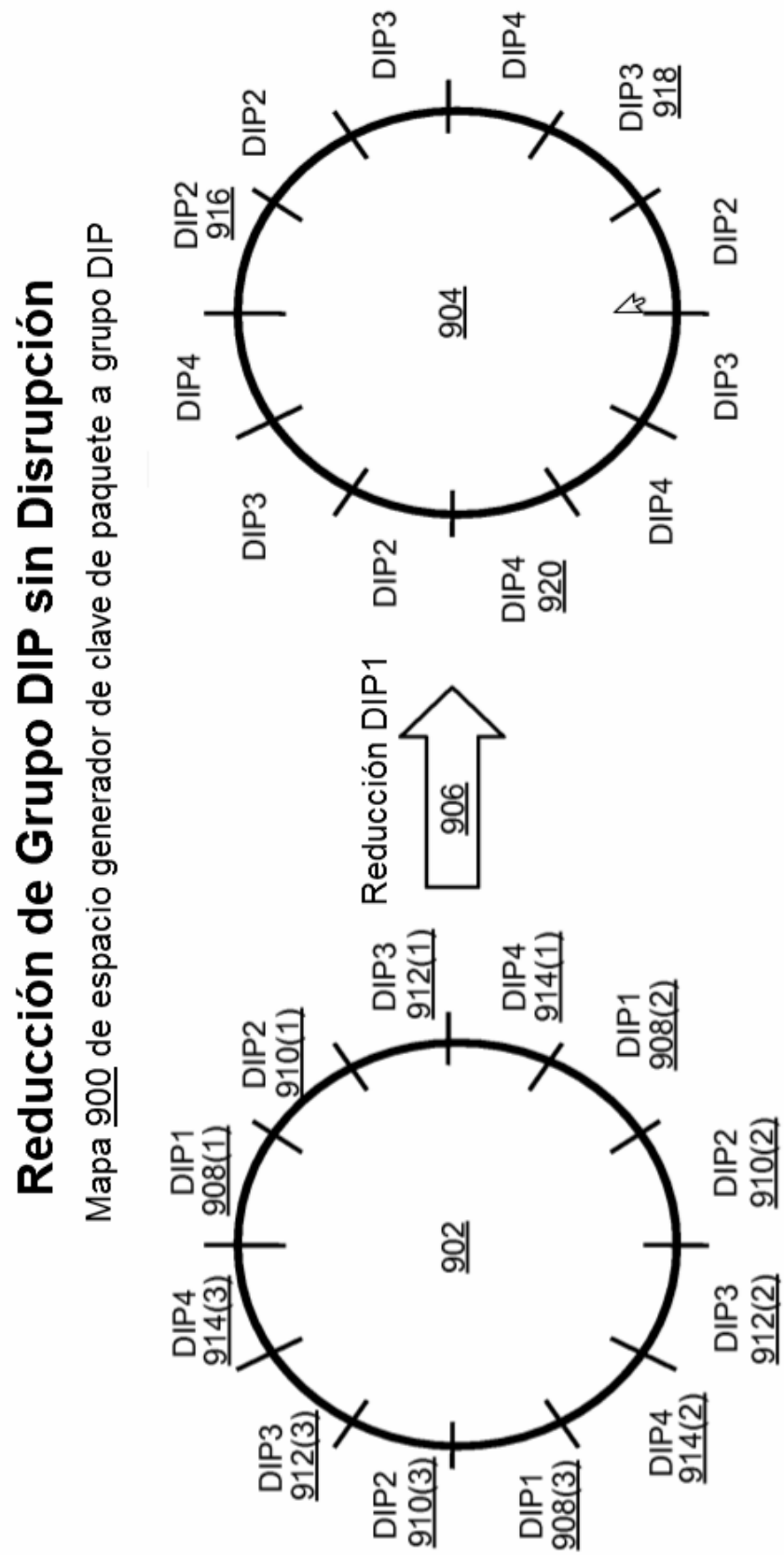
Reducción de Grupo DIP sin Disrupción
Mapa 900 de espacio generador de clave de paquete a grupo DIP
902
DIP1 908(1)
DIP2 910(1)
DIP3 912(1)
DIP4 914(1)
DIP1 908(2)
DIP2 910(2)
DIP3 912(2)
DIP4 914(2)
DIP1 908(3)
DIP2 910(3)
DIP3 912(3)
DIP4 914(3)
Reducción DIP1
906
904
DIP2 916
DIP3 918
DIP4 920

FIG. 9

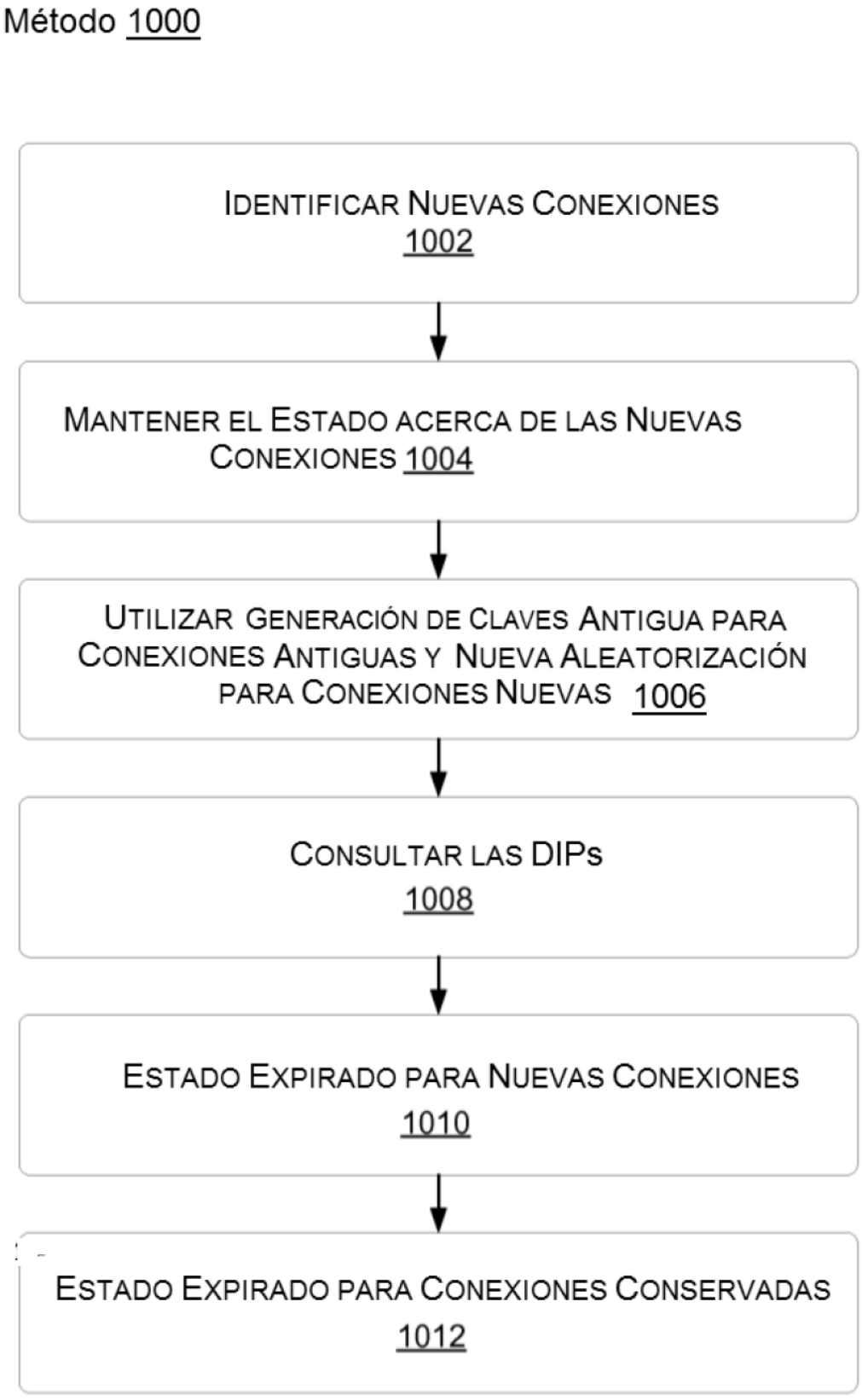
Método 1000
IDENTIFICAR NUEVAS CONEXIONES 1002
MANTENER EL ESTADO ACERCA DE LAS NUEVAS CONEXIONES 1004
UTILIZAR GENERACIÓN DE CLAVES ANTIGUA PARA CONEXIONES ANTIGUAS Y NUEVA ALEATORIZACIÓN PARA CONEXIONES NUEVAS 1006
CONSULTAR LAS DIPs 1008
ESTADO EXPIRADO PARA NUEVAS CONEXIONES 1010
ESTADO EXPIRADO PARA CONEXIONES CONSERVADAS 1012

FIG. 10

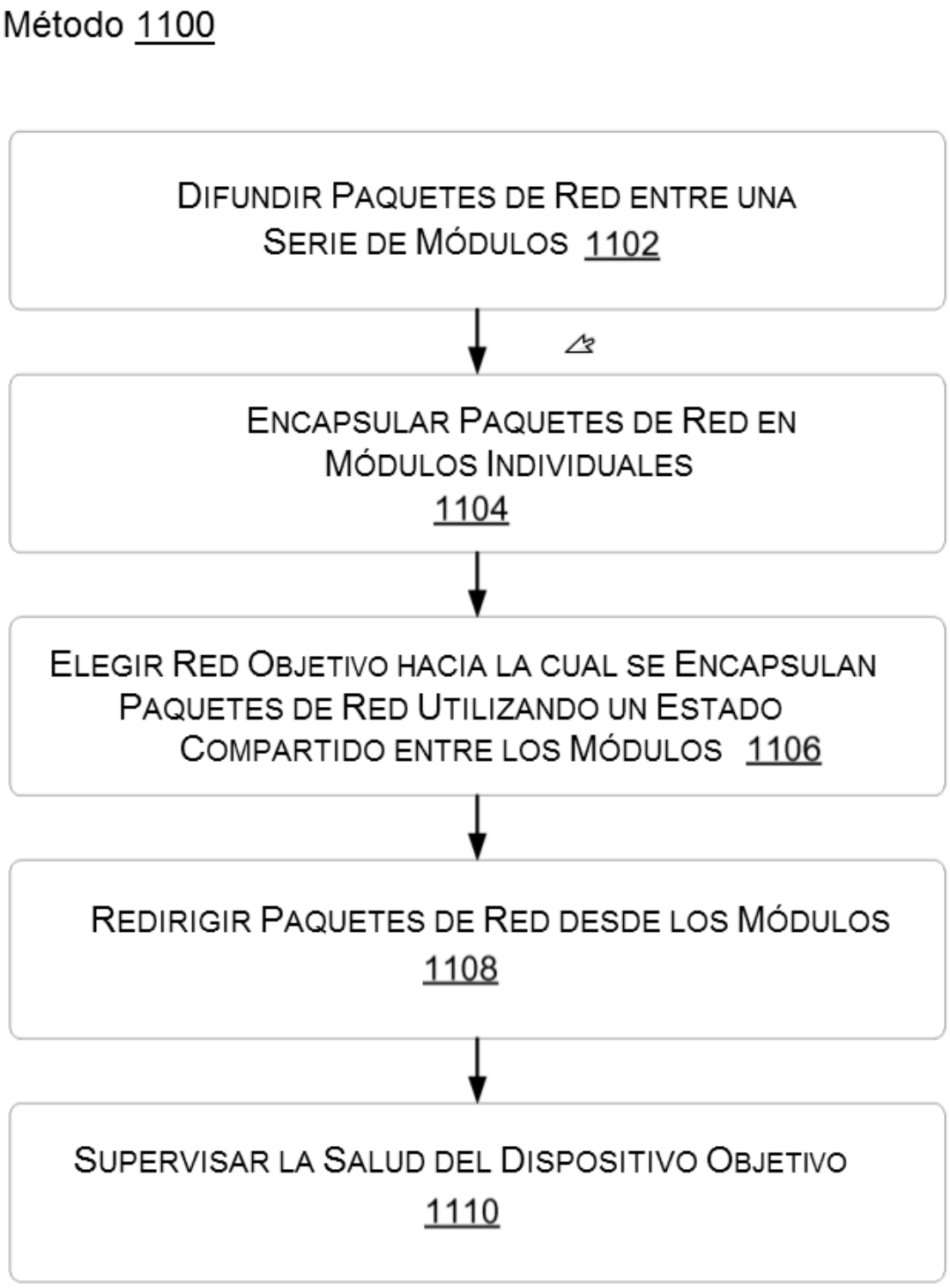
Método 1100
DIFUNDIR PAQUETES DE RED ENTRE UNA SERIE DE MÓDULOS 1102
ENCAPSULAR PAQUETES DE RED EN MÓDULOS INDIVIDUALES 1104
ELEGIR RED OBJETIVO HACIA LA CUAL SE ENCAPSULAN PAQUETES DE RED UTILIZANDO UN ESTADO COMPARTIDO ENTRE LOS MÓDULOS 1106
REDIRIGIR PAQUETES DE RED DESDE LOS MÓDULOS 1108
SUPERVISAR LA SALUD DEL DISPOSITIVO OBJETIVO 1110

FIG. 11