(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 682 358**

(51) Int. CI.:

G06F 15/16 (2006.01)
H04L 29/08 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **12.03.2014 PCT/US2014/024386**

(87) Fecha y número de publicación internacional: **02.10.2014 WO14159605**

(96) Fecha de presentación y número de la solicitud europea: **12.03.2014 E 14776562 (2)**

(97) Fecha y número de publicación de la concesión europea: **06.06.2018 EP 2972957**

(54) Título: **Método y aparato para implementar el almacenamiento en memoria caché de contenido distribuido en una red de distribución de contenido**

(30) Prioridad:

14.03.2013 US 201361783666 P
22.11.2013 US 201314087595

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
20.09.2018

(73) Titular/es:

MASSACHUSETTS INSTITUTE OF TECHNOLOGY (100.0%)
77 Massachusetts Avenue
Cambridge, MA 02139, US

(72) Inventor/es:

MEDARD, MURIEL;
CALMON, FLAVIO, DU PIN y
ZENG, WEIFEI

(74) Agente/Representante:

CURELL AGUILÁ, Mireia

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

DESCRIPCIÓN

Método y aparato para implementar el almacenamiento en memoria caché de contenido distribuido en una red de distribución de contenido.

Campo técnico

La materia objeto dada a conocer en la presente se refiere, en general, a la distribución de datos y, más particularmente, a la distribución de contenido digital desde uno o más proveedores de contenido a una pluralidad de usuarios finales.

Antecedentes

Las redes de distribución de contenido (CDN) son sistemas que se usan para distribuir contenido digital a usuarios finales en una red. Normalmente, las CDN son usadas por proveedores de contenido para distribuir su contenido al público. Los proveedores de contenido pueden mantener sus propias CDN o pueden pagar a un operador independiente de CDN para que distribuya su contenido. Habitualmente, una CDN acordará la distribución de contenido a usuarios con una cierta calidad de servicio (QoS). Típicamente, una CDN operará uno o más centros de datos para prestar soporte a la distribución de contenido. Tal como es bien sabido, un centro de datos es una gran instalación que aloja, típicamente, una amplísima variedad de hardware informático (por ejemplo, servidores, rúters, unidades de disco, etcétera) implicado en el almacenamiento y la distribución de contenido digital a usuarios. La CDN puede ser propietaria del(de los) centro(s) de datos, o puede contratar uno o más centros de datos independientes para facilitar la distribución de contenido.

Tal como se apreciará, las CDN pueden ser muy costosas económicamente y complicadas de operar. Además, con la demanda siempre creciente de contenido digital, se está recurriendo a las CDN para mantener o soportar centros de datos de mayor tamaño y más numerosos con el fin de satisfacer la demanda y cumplir sus obligaciones de calidad de servicio. Existe una necesidad de técnicas, sistemas y dispositivos que se puedan usar para reducir los costes y los requisitos de hardware asociados a la distribución de contenido digital.

El documento US 2012/072526, Kling, da a conocer una red de distribución de contenido en la cual un nodo virtual representa en términos lógicos un conjunto de nodos caché, presentando el conjunto de nodos caché un coste de comunicación común entre los nodos caché. Se recopilan estadísticas referentes a solicitudes de contenido, y, sobre la base de estas estadísticas, se determina si se va a almacenar o no contenido específico en memoria caché, en el nodo virtual.

Sumario

En varias formas de realización descritas en la presente, se proporcionan técnicas y sistemas para potenciar el rendimiento y el diseño de redes de distribución de contenido (CDN) usando un almacenamiento en memoria caché de contenido distribuido (DCC). Algunas de las características descritas en la presente se pueden usar, por ejemplo, para potenciar la eficiencia de funcionamiento de CDN, y reducir el coste de implementación de tales sistemas. En algunas formas de realización, se proporcionan técnicas de optimización para optimizar el rendimiento del DCC en una CDN. Se proporcionan también técnicas de codificación de red para almacenar contenido en una CDN con el fin de permitir, por ejemplo, una distribución, sin fisuras, de archivos a través de diferentes pasarelas o dispositivos de borde en la red. El uso de la codificación de red puede permitir sortear algunos problemas asociados a la partición de archivos individuales, permitiendo, así, la distribución, sin fisuras, de archivos a través de la red sin tener que preocuparse por la ubicación específica de diferentes partes de un archivo dado.

De acuerdo con un aspecto, se proporciona un método implementado por máquina para su uso en una red de distribución de contenido (CDN) que distribuye contenido a una pluralidad de usuarios de acuerdo con la reivindicación 1.

En una forma de realización, la recopilación, el ensamblaje, la determinación y el envío se llevan a cabo en el servidor central.

En una forma de realización, el método comprende, además, repetir continuamente la recopilación, el ensamblaje, la determinación y el envío para operar la CDN de una manera eficiente.

En una realización, la función de coste incluye un término para tener en cuenta retardos de entrega del servidor.

En una forma de realización, la función de coste incluye una restricción sobre la varianza de la carga del servidor.

En una forma de realización, la determinación del contenido que se va a almacenar en la pluralidad de dispositivos de pasarela incluye usar por lo menos uno de entre los siguientes procedimientos de optimización

para minimizar la función de coste: el Esquema de Gradiente de Proximidad General, el método de punto interior, y solucionadores numéricos, tales como GUROBI y CVX.

En una forma de realización, el envío de contenido a la pluralidad de dispositivos de pasarela para que sea almacenado en memoria caché por ellos incluye el envío del contenido por medio de Internet.

En una forma de realización, la pluralidad de dispositivos de pasarela asociados a la CDN se comunica con Internet a través de un Proveedor de Servicios de Internet (ISP) común, en donde el servidor central está conectado a Internet.

En una forma de realización, el envío de contenido a la pluralidad de dispositivos de pasarela incluye hacer que por lo menos parte del contenido se envíe a los dispositivos de pasarela desde uno o más centros de datos que no están situados en la ubicación del servidor central.

En una forma de realización, el envío de contenido a la pluralidad de dispositivos de pasarela para que sea almacenado en memoria caché por ellos, incluye enviar segmentos de archivo codificados en red a dispositivos de pasarela.

De acuerdo con otro aspecto, se proporciona un servidor de distribución de contenido para gestionar servicios de distribución de contenido para la CDN según la reivindicación 9.

En una forma de realización, la pluralidad de dispositivos de pasarela desplegados en o cerca de las ubicaciones de usuario están bajo el control de la CDN.

En una forma de realización, el servidor de distribución de contenido está configurado para actualizar ocasionalmente el esquema de almacenamiento en memoria caché con el fin de tener en cuenta cambios en la CDN con el paso del tiempo.

En una forma de realización, el servidor de distribución de contenido está acoplado a la pluralidad de dispositivos de pasarela a través de Internet.

En una forma de realización, la pluralidad de dispositivos de pasarela está asociada, en su totalidad, a un proveedor de servicios de Internet (ISP) común.

En una forma de realización, el servidor de distribución de contenido está configurado para: recopilar información sobre una configuración actual de la CDN; ensamblar estadísticas de funcionamiento de la CDN en correspondencia con requisitos de contenido de usuario; y usar la información recopilada y las estadísticas ensambladas para desarrollar el esquema de almacenamiento en memoria caché para la CDN.

En una forma de realización, la función de coste usa un coste para la transmisión, por parte del servidor de distribución de contenido, de una unidad de contenido a una pasarela, y un coste de almacenamiento en memoria caché de una unidad de contenido en una pasarela.

En una forma de realización, la función de coste incluye un término para tener en cuenta retardos de entrega en el servidor.

En una forma de realización, la función de coste incluye una restricción sobre la varianza de carga del servidor.

En una forma de realización, el servidor de distribución de contenido está configurado para distribuir contenido codificado en red a la pluralidad de dispositivos de pasarela con el fin de que sea almacenado en memoria caché en estos últimos.

De acuerdo todavía con otro aspecto de los conceptos, sistemas, circuitos y técnicas descritos en la presente, se proporciona un artículo que comprende uno o más soportes legibles por ordenador, no transitorios, que tienen almacenadas, en ellos, instrucciones que, cuando son ejecutadas por un sistema informático, llevan a cabo un método para su uso en una red de distribución de contenido (CDN) que distribuye contenido a una pluralidad de usuarios.

En una forma de realización, la función de coste incluye un término para tener en cuenta retardos de entrega en el servidor.

En una forma de realización, la función de coste incluye una restricción sobre la varianza de carga del servidor.

Breve descripción de los dibujos

Las anteriores características se pueden entender más exhaustivamente a partir de la siguiente descripción de los dibujos, en los cuales:

la Fig. 1 es un diagrama de bloques que ilustra una red de distribución de contenido (CDN) de acuerdo con una forma de realización;

la Fig. 2 es una gráfica que ilustra valores de coste óptimos para diversos números de pasarelas (N) dentro de una CDN en función del número de archivos de película (M) que están disponibles dentro de la CDN;

la Fig. 3 es una gráfica que ilustra la convergencia de un proceso de optimización de acuerdo con una forma de realización;

la Fig. 4 es una gráfica que ilustra diversos costes en función de la capacidad del servidor para una CDN de acuerdo con una forma de realización; y

la Fig. 5 es un diagrama de flujo que ilustra un método ejemplificativo para operar una CDN que usa DCC de acuerdo con una forma de realización.

Descripción detallada

En la presente se describen técnicas, sistemas y dispositivos que permiten distribuir contenido digital a usuarios dentro de redes de una manera eficiente y rentable. Estas técnicas, sistemas y dispositivos pueden ser usados por proveedores de contenido, operadores de redes de distribución de contenido (CDN), y operadores de centros de datos para, por ejemplo, reducir los requisitos de coste, complejidad y hardware asociados a la distribución de contenido. En diversas formas de realización, se usan técnicas de almacenamiento, en memoria caché, de contenido distribuido (DCC) para reducir requisitos de hardware dentro de centros de datos asociados a una CDN. Tal como se describirá de forma más detallada, el DCC implica almacenar parte o la totalidad del contenido que se va a distribuir cerca de los puntos extremos de la red (es decir, cerca de las ubicaciones de usuario) en lugar de en una o más ubicaciones centralizadas (por ejemplo, un centro de datos, etcétera). De este modo, en algunas formas de realización que se describen en la presente, se almacena contenido en dispositivos de borde o pasarelas asociados a usuarios de una red que están situados en o cerca de las ubicaciones de usuario. Además, en varias formas de realización, se proporcionan técnicas para seleccionar de forma inteligente el contenido que se almacenará en memoria caché dentro de los dispositivos de pasarela de la red, con el fin de lograr una distribución eficiente del contenido, teniendo en cuenta el coste de almacenamiento y las restricciones de comunicación en la red. Las técnicas, sistemas y dispositivos descritos en la presente se pueden usar para distribuir contenido en tipos diferentes de redes, incluyendo redes tanto grandes como pequeñas y redes tanto privadas como públicas.

En algunas formas de realización, se proporcionan una o más funciones de coste para su uso en la mejora de la eficiencia de funcionamiento del DCC. Se pueden utilizar técnicas de optimización para minimizar las funciones de coste con el fin de hallar valores para una serie de variables de decisión asociadas al sistema. A continuación, los valores de las variables de decisión se pueden usar para seleccionar el contenido que se va a almacenar en las pasarelas de usuario de la red. Para codificar el contenido que se va a almacenar en las ubicaciones de borde se usa codificación de red. Usando este planteamiento, pueden evitarse problemas asociados a la partición y la secuenciación de partes de un archivo de contenido (por ejemplo, partes separadas de un archivo de una película en vídeo, etcétera). Usando técnicas de codificación de red, un archivo de contenido se puede separar en numerosos segmentos codificados que no requieren números de secuencia para su reensamblaje en un archivo utilizable. Por el contrario, un usuario final debe simplemente recopilar un número suficiente de dichos segmentos codificados de diferentes ubicaciones fuente para permitir que se produzca la descodificación. Las fuentes concretas de los segmentos codificados no serán relevantes siempre que se reciba un número suficiente de segmentos codificados linealmente independientes para permitir la descodificación.

Tal como se ha descrito previamente, las CDN usan, típicamente, centros de datos para facilitar la distribución de contenido a usuarios finales. Los centros de datos son grandes instalaciones que centralizan gran parte de los equipos requeridos para una distribución eficaz de contenido. Esta centralización de equipos permite reducir los costes globales, incluyendo los costes tanto de mantenimiento como de comunicación. No obstante, dicha centralización tiene también problemas asociados a ella. Estos problemas pueden incluir aquellos asociados al sobredimensionamiento, la disipación de energía y la distancia al usuario final. Cualquier empresa que proporcione servicios en línea, con independencia de su tamaño, debe tener en consideración los costes asociados a centros de datos como componente crítico de su modelo de negocio.

El sobredimensionamiento está relacionado con el hecho de que la mayoría de centros de datos está diseñada para igualar la demanda de pico en lugar de la demanda media, lo cual hace que aumente considerablemente el número de servidores requerido. Debido a sus arquitecturas físicas (por ejemplo, un número elevado de

procesadores que funcionan simultáneamente, etcétera), los centros de datos pueden generar una gran cantidad de calor no deseado que es necesario eliminar para garantizar un funcionamiento continuado. En ocasiones, la eliminación de este calor puede constituir un desafío. Además, típicamente los centros de datos están alejados de las ubicaciones de usuario en la red. Estas grandes distancias pueden requerir un aumento del suministro de ancho de banda para la comunicación, y también pueden presentar un problema para aplicaciones con restricciones de retardo estrictas, tales como el flujo continuo de vídeo.

El uso de almacenamiento en memoria caché de contenido distribuido (DCC) es una de las maneras de reducir o eliminar algunos de los problemas asociados a centros de datos cuando se lleva a cabo la distribución de contenido digital. Tal como se ha descrito anteriormente, el DCC conlleva el almacenamiento de parte o la totalidad del contenido que se va a distribuir a usuarios, en dispositivos de borde o pasarela situados en o cerca de las ubicaciones de usuario. Con el precio de los medios de almacenamiento en constante disminución, las pasarelas se pueden equipar con grandes cantidades de medios de almacenamiento digital para su uso en el almacenamiento de contenido digital. Estos dispositivos también se pueden equipar con cierto nivel de poder computacional digital que se puede usar para prestar soporte al almacenamiento de contenido en memoria caché en algunas formas de realización. Además, típicamente dichos dispositivos se conectan a Internet (u otra red) a través de enlaces de banda ancha, ofreciéndose, así, a los dispositivos, la capacidad de actuar como servidores para la distribución de contenido a pequeña escala. En algunos sistemas, dispositivos de pasarela asociados a usuarios diferentes pueden comunicarse directamente entre sí, o comunicarse uno con otro a través de un ISP correspondiente, de una manera relativamente económica. En algunas formas de realización, se puede aprovechar esta capacidad de comunicación económica cuando se implementa el DCC en una CDN, con pasarelas que prestan servicio a otras pasarelas al estilo de una comunicación entre pares.

Para ilustrar la idea antes descrita, considérese el desafío al que se enfrenta una compañía de cable que desea proporcionar a los usuarios, servicios de vídeo bajo demanda, de alta resolución. Por un lado, la compañía de cable puede usar su CDN para distribuir contenido a usuarios con un cierto coste, el cual dependerá de la cantidad de recursos demandados (por ejemplo, volumen de tráfico, costes de los centros de datos, etcétera). Como alternativa, la compañía de cable puede sacar provecho de la memoria y la conectividad de los miles de pasarelas de su red para liberar, a la CDN, de por lo menos parte de su funcionalidad. Si las pasarelas se usan de esta manera, la compañía de cable tendrá que determinar cómo distribuir los archivos de contenido a través de las diferentes pasarelas teniendo en cuenta, por ejemplo, el coste de almacenamiento, las restricciones de comunicación de la red, y la fiabilidad de esta última. Tal como se ha descrito previamente, en la presente se describen técnicas, sistemas y dispositivos que se pueden usar para determinar cómo distribuir el contenido entre los dispositivos de pasarela de una manera eficiente y rentable.

La Fig. 1 es un diagrama de bloques que ilustra una red de distribución de contenido (CDN) 10 de acuerdo con una forma de realización. Tal como se ilustra, la CDN 10 incluye un servidor de distribución de contenido 12 que se usa para gestionar la distribución de contenido digital a una pluralidad de usuarios 14a a 14n por medio de Internet 20 u otra red. El contenido digital puede incluir cualquier tipo de contenido que se pueda distribuir a usuarios finales como respuesta a solicitudes, incluyendo, por ejemplo, archivos de vídeo, archivos de audio, descargas de software, medios de flujo continuo, archivos de datos, archivos de texto, contenido de noticias, videojuegos, juegos en línea, y/u otros. En la exposición que se ofrece seguidamente, se describirán varias técnicas y sistemas de distribución de contenido en el contexto de archivos de vídeo, tales como películas distribuidas como parte de una aplicación de tipo películas bajo demanda. No obstante, debe entenderse que las técnicas y sistemas descritos se pueden usar con otros tipos de contenido digital en otras implementaciones.

Tal como se muestra en la Fig. 1, cada uno de los usuarios 14a a 14n asociados a la CDN 10 está acoplado a un dispositivo de pasarela 16a a 16n correspondiente que actúa como punto de entrada a la red de mayor tamaño (por ejemplo, Internet 20). Los dispositivos de pasarela 16a a 16n pueden incluir, cada uno de ellos, cualquier tipo de dispositivo que permita que un usuario se conecte a una red de mayor tamaño. Por ejemplo, un dispositivo de pasarela puede incluir: una caja de adaptación del televisor, una pasarela residencial, una pasarela WiMax, una pasarela celular, un módem de cable, un módem DSL, una picocélula asociada a un sistema de comunicaciones celulares, un rúter, una consola de videojuegos, o cualquier otro dispositivo que tenga medios de almacenamiento y suficiente poder computacional para prestar soporte a la aplicación deseada (tal como un ordenador de sobremesa o portátil). En algunas formas de realización, uno o más de los dispositivos de pasarela 16a a 16n también pueden proporcionar cierto nivel de conversión de protocolos o formatos de señales entre redes dispares, aunque esto no es un requisito. Adicionalmente, en algunas formas de realización, tipos diferentes de dispositivos de pasarela 16a a 16n pueden ser usados por usuarios diferentes 14a a 14n. En algunas formas de realización, los dispositivos de pasarela 16a a 16n pueden ser equipos que son proporcionados a los usuarios por un ISP, un proveedor de contenido, o un operador de CDN, aunque también se pueden usar dispositivos de pasarela 16a a 16n que sean propiedad del usuario y/o controlados por el mismo.

En la forma de realización ilustrada, todos los usuarios 14a a 14n asociados a la CDN 10 se conectan a Internet 20 a través de un Proveedor de Servicios de Internet (ISP) 18 común. Tal como se describirá de forma más detallada, al limitar la CDN 10 a usuarios que comparten un ISP, se puede adoptar una comunicación de coste relativamente bajo entre los dispositivos de pasarela 16a a 16n. No obstante, en algunas formas de realización,

dentro de una CDN puede haber presencia de usuarios que se comunican con Internet (u otra red) por medio de múltiples proveedores diferentes. En referencia a la Fig. 1, los dispositivos de pasarela 16a a 16n se acoplan al ISP 18 por medio de una pluralidad de enlaces de comunicación 22. Estos enlaces 22 pueden incluir enlaces inalámbricos y/o por cable. En algunas formas de realización, las pasarelas de usuario se pueden conectar directamente a una red de mayor tamaño sin necesidad de un proveedor de servicios intermedio (es decir, un ISP, etcétera).

Cada uno de los usuarios 14a a 14n mostrados en la Fig. 1 puede tener un equipo de usuario asociado a él que le permita solicitar, recibir y utilizar el contenido digital que les está siendo distribuido. El equipo de usuario puede incluir, por ejemplo, un ordenador, un televisor, un reproductor de medios, un dispositivo de audio, y/u otros. En una aplicación de flujo continuo de vídeo, por ejemplo, puede usarse un equipo de usuario que tenga la capacidad de reproducir archivos de vídeo en flujo continuo. Pueden usarse, de manera adicional o alternativa, muchos otros tipos de equipo de usuario. En algunas implementaciones, el equipo de usuario asociado a un usuario puede incluir, por ejemplo, un equipo de red local. Por ejemplo, uno o más de los usuarios 14a a 14n pueden mantener una red de área local (LAN) dentro de un edificio residencial o de oficinas correspondiente. Un rúter o punto de acceso inalámbrico (WAP) asociado a la LAN se puede conectar a un dispositivo de pasarela correspondiente para conectar la LAN a Internet. Tal como se apreciará, una LAN de este tipo puede permitir que múltiples usuarios compartan un dispositivo de pasarela.

En algunas formas de realización, los usuarios 14a a 14n pueden representar, cada uno de ellos, usuarios en ubicaciones de usuario fijas, tales como dentro de edificios, residencias u otras estructuras estacionarias correspondientes. En otras implementaciones, parte o la totalidad de los usuarios 14a a 14n puede ser usuarios móviles que tengan equipos de usuario móviles. En estas implementaciones, las pasarelas que incluyen los medios de almacenamiento local para almacenar contenido en memoria caché pueden incluir, por ejemplo, una estación base inalámbrica (por ejemplo, una estación base celular en un sistema celular, una estación base WiMax en una red WiMax, etcétera) que proporcione acceso a una red para los usuarios móviles. En algunas otras implementaciones, una estación base inalámbrica puede actuar como servidor central, y los propios dispositivos móviles pueden actuar como pasarelas locales que almacenan el contenido localmente.

El servidor de CDN 12 puede ser operativo para, entre otras cosas, recibir y procesar solicitudes de contenido digital provenientes de usuarios. De este modo, el servidor de CDN 12 puede recibir una solicitud de distribución de una película particular, desde, por ejemplo, el usuario 14h. A continuación, el servidor de CDN 12 puede determinar si la película solicitada está almacenada en la CDN 10, y hacer que la película sea distribuida desde esa ubicación al usuario solicitante. En otras implementaciones, las propias pasarelas 16a a 16n pueden recibir parte o la totalidad de las solicitudes provenientes de usuarios, determinar las ubicaciones de la película solicitada en los dispositivos de pasarela cercanos, y facilitar la distribución del archivo de película solicitado al usuario solicitante. En algunas formas de realización, un archivo solicitado único se puede dividir entre múltiples ubicaciones de almacenamiento diferentes dentro de la CDN 10. En este caso, el servidor de CDN 12 y/o la pasarela pueden conseguir que la totalidad de las diferentes partes del archivo sea distribuida al usuario. En una implementación en la que se use codificación de red para almacenar contenido digital, segmentos de archivo codificados correspondientes a un archivo particular se pueden almacenar en numerosos lugares dentro de la CDN 10. En este escenario, el servidor de CDN 12 o pasarela puede conseguir que paquetes codificados sean distribuidos al usuario solicitante desde diversas fuentes diferentes. En algunas implementaciones, la transferencia de segmentos codificados puede continuar hasta que se haya enviado un número predeterminado de segmentos codificados independientes al usuario o hasta que el usuario envíe un mensaje de acuse de recibo indicando que se han recibido suficientes paquetes codificados (es decir, suficientes grados de libertad) para posibilitar que tenga lugar la descodificación. Tal como se ha descrito previamente, en algunas implementaciones, las diversas pasarelas 16a a 16n podrán comunicarse entre sí. Esta comunicación puede ser una comunicación por cable o inalámbrica. Cuando se soporta una comunicación inalámbrica, la comunicación puede incluir tanto una comunicación de un solo salto como comunicaciones de múltiples saltos. En algunas formas de realización, las pasarelas se pueden comunicar entre sí a través del ISP 18 o por medio de alguna otra ruta de bajo coste (por ejemplo, una ruta que no sea Internet).

En algunas formas de realización, un archivo de película solicitado puede no estar almacenado, o puede no estar disponible en ese momento, en uno de los dispositivos de pasarela 16a a 16n. En este escenario, el servidor de CDN 12 puede decidir distribuir el archivo solicitado al usuario desde sus propios medios de almacenamiento local, o desde uno o más centros de datos 26, 28 asociados a la CDN 10. En algunas formas de realización, el servidor central 12 tendrá todos los archivos de película posibles almacenados localmente en la ubicación del servidor. En algunas otras formas de realización, los archivos de película se pueden almacenar en uno o más centros de datos que no estén ubicados conjuntamente con el servidor central. Todavía en otras formas de realización, algunos de los archivos de película pueden estar almacenados en la ubicación del servidor y otros pueden estar almacenados en otra u otras ubicaciones. Si un archivo solicitado no está disponible en ninguna fuente asociada a la CDN 10, el servidor de CDN 12 puede denegar la solicitud.

De acuerdo con algunas formas de realización descritas en la presente, se proporcionan técnicas y sistemas para determinar cómo disponer contenido dentro de una CDN de una manera eficiente, con el fin de proporcionar

una distribución eficiente de contenido de datos dentro de la CDN. Más específicamente, se proporcionan técnicas que permiten que una CDN determine qué contenido se debería almacenar dentro de las diversas pasarelas de la red para generar un funcionamiento de coste relativamente bajo. En al menos una implementación, el servidor de CDN 12 se puede configurar para determinar qué contenido se almacenará dentro de qué dispositivo de pasarela 16a a 16n con el fin de lograr un funcionamiento eficiente y de bajo coste. Debe apreciarse que esta función se puede llevar a cabo alternativamente en otra u otras ubicaciones dentro de la CDN 10. En algunas formas de realización, esta función se puede realizar de una manera distribuida en numerosas ubicaciones dentro de la CDN 10.

Tal como se describirá de manera más detallada, en algunas formas de realización, se definen una o más funciones de coste que calibran los diversos costes implicados en el almacenamiento de contenido dentro de una CDN. Para determinar cómo disponer contenido dentro de la CDN, se pueden usar uno o más programas de optimización con el fin de minimizar una función de coste. Se definen variables de decisión que se pueden variar durante el proceso de optimización para alcanzar una función de coste mínimo. A continuación, los valores finales de las variables de decisión se pueden usar para determinar qué contenido se debe almacenar en qué dispositivo de pasarela de la CDN 10. A continuación se describirá un marco matemático destinado a usarse en la descripción de técnicas de optimización ejemplificativas que se pueden usar de acuerdo con formas de realización.

En primer lugar, se definirá una pluralidad de variables de sistema de una CDN. Las variables de sistema pueden incluir, por ejemplo:

$\mathcal{X} = \{1, \ldots \mathrm{N}\}$: el conjunto de índices de pasarelas domésticas, donde el número total de pasarelas es *N*.

$\mathcal{M} = \{1, \ldots \mathrm{M}\}$: el conjunto de índices de archivos de película disponibles en la CDN, donde el número total de archivos de película independientes viene dado por *M*. El índice de cada película es también el rango de popularidad de la película.

$y = (y_1, \ldots, y_M)$: un vector que tiene el tamaño de cada uno de los *M* archivos de película. Se puede formar una matriz duplicando y como $Y = [y, \ldots, y]$.

$C = \{C_1, \ldots, C_N\}$: la capacidad de transmisión de la pasarela *i* para transmisiones salientes.

$P_{m,i}$: la probabilidad de que haya una demanda de película *m* en la pasarela *i*. *P* puede ser la matriz *M x N* cuya entrada en la posición *(m, i)* viene dada por $P_{m,i}$ (es decir, la columna $i^{ésima}$ de la matriz *P* proporciona la probabilidad de demandar cada película en la pasarela *i*). En algunas formas de realización, se puede considerar que la demanda de películas sigue una distribución de Zipf, la cual se ha establecido como una buena aproximación para medir la popularidad de películas de vídeo [véase, por ejemplo, "I Tube, You Tube, Everybody Tubes: Analyzing the World's Largest User Generated Content Video System" de Cha et al., *Proc. of the 7th ACM SIGCOMM Conference on Internet Measurement, ser. IMC '07,* Nueva York, NY, Estados Unidos*; ACM, 2007, págs. 1-14*]. Específicamente, una película con un rango de popularidad *j* (indexada *j*) tiene una probabilidad de demanda de:

$$P_{m,i} = \frac{1/j^{\gamma}}{\sum_{k=1}^{M} 1/k^{\gamma}}$$

donde γ caracteriza la distribución. Puesto que $\gamma \rightarrow 0$, la distribución se aproxima a uniforme. A no ser que se especifique, en la presente se considerará que $P_{m,i}$ es igual para toda *i*.

c_s: el coste para que el servidor central transmita una unidad de contenido a cualquier pasarela.

c_g: el coste de almacenar en memoria caché una unidad de contenido en cualquier pasarela.

S_c: la capacidad del servidor. Este es el volumen total medio máximo de contenido que se puede transmitir desde el servidor a todos los usuarios.

δ_s: el límite superior sobre la varianza de la carga del servidor.

α: el coste de retardo en el servidor.

Tal como se ha descrito previamente, puesto que todos los usuarios en la CDN 10 de la Fig. 1 se conectan a Internet por medio de un ISP común, puede considerarse que el tráfico entre las pasarelas en la CDN 10 tiene un coste reducido o inexistente. Esta consideración es práctica desde el punto de vista del ISP en la medida en la

que la mayor parte del coste en un ISP se genera por tráfico inter-ISP. Siempre que la carga de tráfico entre las pasarelas esté por debajo de la capacidad, el coste para el ISP no será significativo.

A continuación se describen las variables de decisión de un proceso de optimización ejemplificativo de acuerdo con una forma de realización:

$x_i = [x_{i,1}, \ldots, x_{i,M}]^T, i \in \chi$: la fracción de cada archivo de película *M* almacenada en memoria caché en la pasarela doméstica *i*.

$X = [x_1, x_2, \ldots, x_N]$: la matriz de vectores x_i.

$a_{i,j,m}$: la demanda desde la pasarela *i*, de contenido *m* de la pasarela *j*, en caso de que haya una solicitud de *m* en la pasarela *i*. La unidad de $a_{i,j,m}$ está en términos de la fracción de contenido *m*. La variable A_m se puede usar para indicar la "matriz de conectividad" para el archivo *m*. Es decir, la entrada $a_{i,j,m}$ representa el porcentaje de archivo *m* demandado por la pasarela *i* del usuario *j*, en el caso en el que la pasarela *i* solicite el archivo *m*.

$L_{m,i}$: la fracción del archivo *m* solicitada al servidor por una pasarela *i*, $i \in \chi$, $1 \leq m \leq M$. La matriz correspondiente es *L*.

g_i: el espacio de almacenamiento usado en la pasarela $i \in \chi$ para almacenar en memoria caché cualquier contenido de las *M* películas.

$g = \{g_1, \ldots, g_N\}$: el vector de valores de g_i.

s_m: el volumen total medio de demanda del archivo *m* desde el servidor.

$s = \{s_1, \ldots, s_N\}$: el vector de valores de s_i.

$S = \sum_{j=1}^{M} s_j$: el volumen total medio de contenido servido por el servidor central a los usuarios.

A continuación se proporcionan varias formulaciones diferentes del problema de optimización de acuerdo con formas de realización ejemplificativas. Una primera formulación es una formulación de programa lineal (LP) de línea de base del problema. A continuación, se agregan objetivos y restricciones no lineales encima de la formulación de LP para generar soluciones deseadas para diferentes necesidades de la CDN (es decir, formulaciones no lineales). Tal como se describirá de forma más detallada, en cada una de las formulaciones, se minimizará una función de coste al mismo tiempo que planteando diversas restricciones y condiciones sobre las variables de decisión. Debe apreciarse que las técnicas y estrategias descritas en la presente no se limitan a su uso con las diversas formulaciones descritas en este documento. Es decir, alternativamente se pueden desarrollar otras formulaciones para materializar una o más de las funciones, resultados, o beneficios de aquellas descritas en la presente.

A continuación se describirá una formulación de programa lineal (LP) de línea de base. En al menos una forma de realización, la formulación de LP se puede expresar de la manera siguiente:

Formulación 1. (P1: Formulación de LP)

minimizar $c_s \cdot 1^T s + c_g \cdot 1^T g$

sujeto a $y^T x_i \leq g_i, \ i \in \mathcal{X}$

$0 \leq x_i \leq 1$

$0 \leq a_{i,j,m} \leq x_{j,m}$

$a_{i,i,m} = x_{i,m}$

$L_{m,i} \geq 1 - \sum_{k=1}^{N} a_{i,k,m} \qquad m \in \mathcal{M}, i \in \mathcal{X}$

$L_{m,i} \geq 0$

$\sum_{m=1}^{M} \sum_{\substack{j=1 \\ j \neq i}}^{N} y_m a_{j,i,m} P_{m,j} \leq C_i$

$s_m = \sum_{i=1}^{N} y_m L_{m,i} P_{m,i}, \quad m \in \mathcal{M}$

Las diversas variables se definieron anteriormente. La función de coste que es minimizada por el proceso de optimización es $c_s \cdot 1^T s + c_g \cdot 1^T g$. Tal como se pone de manifiesto, esta función incluye un término de coste asociado a distribuciones de contenido desde el servidor central, y un término de coste asociado al almacenamiento de contenido en los dispositivos de pasarela. Las diversas restricciones enumeradas anteriormente capturan las limitaciones de almacenamiento y de transmisión que se pueden plantear sobre la red y los nodos en una forma de realización. Estas restricciones se pueden generalizar para capturar diferentes limitaciones de dispositivos de red y de demandas de usuario, y se pueden adaptar a casos diferentes. El proceso de optimización hallará valores para los diversos parámetros de decisión, que minimizarán esta función de coste. A continuación, estos valores se pueden usar para determinar qué contenido se debe almacenar en qué pasarela. En algunas implementaciones, se puede suponer que el número de pasarelas, la distribución de la demanda de películas para usuarios detrás de cada pasarela, y el número y el tamaño de archivos son conocidos. Por lo tanto, la finalidad es hallar una estrategia de almacenamiento en memoria caché óptima (o casi óptima) en las pasarelas, con el fin de minimizar el coste de la diseminación de las películas sobre la red.

En la formulación de LP antes descrita, no hay ninguna restricción impuesta sobre la capacidad de la carga media del servidor. Esto es debido a que, en esta implementación, es suficiente con penalizar el uso de carga del servidor en la función objetivo usando $c_s \cdot 1^T s$. La formulación de LP ignora el coste de retardo en el servidor, lo cual, en algunos escenarios, puede ser no deseable en un nivel elevado. Por lo tanto, en una formulación modificada, a la función de coste se le puede añadir un término de penalización para tener en cuenta el retardo.

A partir de la teoría de colas, se sabe que el retardo de servicio de un servidor se expresa como donde $\frac{1}{1-\rho}$, es $\rho = \frac{S}{S_C}$ el factor de carga del servidor. La formulación nueva es:

Formulación 2: (P2, Penalización de Retardo de Servicio)

$$\textit{minimizar} \quad c_s \cdot 1^T s + c_g \cdot 1^T g + \frac{\alpha}{1 - \frac{S}{S_C}}$$

$$\textit{sujeto a} \quad \sum_{m=1}^{M} s_m = S$$

$$S \leq S_C$$

X, L, g, s, A_m son factibles para P1

Las restricciones mostradas en la anterior Formulación 2 son una versión simplificada de aquellas descritas anteriormente en relación con la Formulación 1. Igual que antes, estas restricciones se pueden adaptar a diferentes ajustes y demandas del usuario.

En un centro de datos o servidor de CDN de gran escala, con frecuencia es deseable limitar la variación de la carga del servidor con el fin de mantener, por ejemplo, estabilidad en el sistema y/o estabilidad de red eléctrica. Considerando dichos factores, la varianza de la carga del servidor se puede restringir por δ_s, y se puede resolver una versión modificada de la formulación de LP usando técnicas de programación no lineal (NLP) de acuerdo con algunas formas de realización. Esta formulación se puede expresar como:

Formulación 3. (P3, varianza de carga del servidor restringida)

$$\textit{minimizar} \quad c_s \cdot 1^T s + c_g \cdot 1^T g$$

$$\textit{sujeto a} \quad \sum_{m=1}^{M} s_m = S$$

$$S \leq S_C$$

$$\sum_{i=1}^{N} \sum_{m=1}^{M} P_{m,i}(1 - P_{m,i})(y_m L_{m,i})^2 \leq \delta_s$$

X, L, g, s, A_m son factibles para P1

En algunas implementaciones, se puede suponer que, para la misma pasarela, las solicitudes de diferentes archivos de película son independientes. Como consecuencia, la varianza de la carga del servidor se puede expresar como:

$$\sum_{i=1}^{N}\sum_{m=1}^{M} P_{m,i}\left(1-P_{m,i}\right)\left(y_m L_{m,i}\right)^2 - S^2.$$

A continuación se describirán técnicas para resolver las formulaciones antes descritas, comenzando con la formulación de LP. En el planteamiento de una solución, se desarrolló un algoritmo para convertir la formulación de LP en un problema de LP convencional en forma de:

$$\min c^T x$$

$$\text{sujeto a } Ax = b$$

$$x \geq 0.$$

El número de variables de decisión crece rápidamente con N y M usando este planteamiento. Es decir, existen aproximadamente $N^2M + 2MN + M + N$ variables en la formulación original, y otras $2N^2M + MN + N$ variables de holgura son necesarias para convertir el problema en la forma convencional. Como consecuencia, la matriz *A* crece rápidamente y resulta difícil de calcular y manipular en la medida en la que el tamaño del problema crece.

La formulación de LP de línea de base del problema se puede resolver con al menos dos solucionadores diferentes. El método de interpunto se puede utilizar para la LP con funciones de barrera logarítmicas con el fin de resolver problemas relativamente más pequeños (*N = 10* y *N = 15, M < 60*). Para problemas de tamaño mayor, puede usarse el solucionador CVX (véase, por ejemplo, *CVX: Matlab Software for Disciplined Convex Programming*, versión 2.0, julio de 2013, http://cvxr.com/cvx/; y "*Graph Implementations for Nonsmooth Convex Programs*," de Grant et al., *Recent Advances in Learning and Control, ser. Lecture Notes in Control and Information Sciences*, Springer Verlag Limited, 2008 págs. 95-110, http://www.stanford.edu/~boyd/papers/pdf/graph_dcp.pdf). Los problemas se solucionaron con *N* variando de 10 a 40 y *M* en un intervalo de 25 a 200. El número de variables implicadas en el problema de optimización va desde aproximadamente 5.000 a 680.000. Aunque el tamaño de estas soluciones ejemplificativas es relativamente pequeño en comparación con aquellos a los que se hace frente en ajustes prácticos, se cree que estos escenarios analizados aportan una valiosa percepción de cara a implementaciones futuras de gran escala.

La Fig. 2 es una gráfica que ilustra los valores de coste óptimos para diversos valores de *N* a medida que el número de archivos de película aumenta de 25 a 200, para γ = 0,25. En este escenario, los tamaños de los archivos de película se seleccionan aleatoriamente entre 800 y 900 Megabytes, mientras se permite que cada pasarela preste servicio a un volumen de tráfico de aproximadamente 1/3 a 1/2 del tamaño de una película para la totalidad del resto de pasarelas en la red. Se supone que el coste de almacenar en memoria caché una unidad de contenido en una pasarela es considerablemente inferior al coste de servir una unidad de contenido desde el servidor. Tal como se esperaba, el coste óptimo es una función creciente tanto de *N* como de *M*. Por otro lado, tiende a almacenarse más contenido en pasarelas con una mayor capacidad para prestar servicio a otras entidades pares.

Con la restricción de varianza, el problema es un programa cuadrático simple. El problema se resuelve usando técnicas de programación cuadrática convencionales. No obstante, debido al tamaño de la matriz de entrada, se sometieron a prueba únicamente ajustes con dimensiones relativamente pequeñas. En la medida en la que la varianza de la carga del servidor está restringida de forma ajustada, se observa que la carga del servidor se reduce rápidamente. Los resultados son muy intuitivos, puesto que la cantidad de contenido almacenado en memoria caché en las pasarelas aumenta y la mayoría de la distribución de contenido se produce entre las pasarelas pares.

En al menos una forma de realización, se puede usar un método de primer orden generalizado para resolver la segunda formulación antes descrita (es decir, la Formulación 2 con la penalización de retardo de servicio). Este problema es convexo, aunque la función objetivo es ilimitada dentro de la región factible. En uno de los planteamientos, puede usarse un Esquema de Gradiente de Proximidad General acelerado. Las funciones de proximidad pueden incluir, por ejemplo, la norma al cuadrado (es decir, $D(x,y) = \frac{1}{2}\|x-y\|^2$), considerando la norma Euclídea convencional (que es autodual). Puede considerarse que $S \leq 0{,}95S_c$ con el fin de solucionar cualquier cuestión que pudiera surgir debido a la ilimitación de la función objetivo. Debe indicarse que el punto inicial del algoritmo debe cumplir $S \leq 0{,}95S_c$.

Simplificando la notación usada, la segunda formulación tendrá la forma general:

$$\textit{minimizar} \quad c^T x + \frac{1}{1 - \frac{S}{S_c}}$$

$$\textit{sujeto a} \; a^T x - S$$

$$x \in \chi$$

$$S \leq S_c$$

donde χ es la región factible de la Formulación 1. Obsérvese que la anterior función objetivo se puede expresar como *P(S) + f(S)*, donde *P(·)* es la solución óptima de un programa lineal y *f(S)* es la penalización no lineal. Suponiendo $0 \leq S \leq 0{,}95S_c$, se observa que $||f'(y) - f'(x)|| \leq L||y - x||$ para $L = 400/S_c$.

En cada iteración del esquema de descenso por gradiente, se resuelve un programa cuadrático convencional. A continuación se ofrece el paso principal de cada iteración:

$$y^i \leftarrow (1 - \theta_i)x^i + \theta_i z^i,$$

$$z^{i+1} \leftarrow args\min\left\{c^T x + \frac{S_c}{y^i - S_c}(S - y^i) + \frac{L}{2}\|S - z^i\|_2^2, \text{ sujeto a } a^T x = S, x\epsilon\mathcal{X}, S \leq S_c\right\}$$

$$x^{i+1} \leftarrow (1 - \theta_i)x^i + \theta_i z^{i+1}.$$

El problema previo es un problema convexo convencional y se puede resolver usando cualquier herramienta de programación convexa. En una implementación, se usó CVX debido a las restricciones de velocidad. La Fig. 3 muestra la convergencia de la evolución del método para 500 iteraciones. El algoritmo claramente converge a la solución óptima, tal como se esperaba.

La Fig. 4 es una gráfica que ilustra los diferentes costes $c_s \cdot 1^T s + c_g \cdot 1^T g$ para diferentes valores de S_c. Se observa que el coste no incluye la penalización de retardo. Claramente, a medida que S_c aumenta, el coste total converge al coste sin restricción de penalización. Además, el componente del coste debido a la carga del servidor se incrementa con un ritmo menor que la reducción del coste de almacenamiento en las pasarelas. Esto indica que, para diseños prácticos, el servidor juega un papel importante en la liberación de los requisitos de almacenamiento en las pasarelas, incluso cuando se considera una penalización significativa para el retardo. Además, pueden lograrse costes que son como mucho el 10% del valor óptimo, al mismo tiempo que garantizando restricciones de retardo razonables.

La Fig. 5 es un diagrama de flujo que ilustra un método ejemplificativo 100 para operar una CDN que usa DCC de acuerdo con una forma de realización.

Los elementos rectangulares (tipificados por el elemento 102 en la Fig. 5) se indican en la presente como “bloques de procesado” y pueden representar instrucciones o grupos de instrucciones de software informático. Debe indicarse que el diagrama de flujo de la Fig. 5 representa una forma de realización ejemplificativa del diseño descrito en la presente, y se considera que las variaciones de dicho diagrama, que siguen de manera general el proceso esbozado, se sitúan dentro del alcance de los conceptos, sistemas y técnicas descritos y reivindicados en el presente documento.

Alternativamente, los bloques de procesado pueden representar operaciones llevadas a cabo por circuitos funcionalmente equivalentes, tales como un circuito de procesado de señal digital, un circuito integrado de aplicación específica (ASIC), o una matriz de puertas programables in situ (FPGA). Algunos bloques de procesado se pueden llevar a cabo manualmente mientras que otros bloques de procesado pueden ser llevados a cabo por un procesador u otro circuito. El diagrama de flujo no representa la sintaxis de ningún lenguaje de programación particular. Por el contrario, el diagrama de flujo ilustra la información funcional que necesita alguien con conocimientos habituales en la técnica para fabricar circuitos y/o para generar software informático con el fin de ejecutar el procesado requerido del aparato en particular. Debe indicarse que muchos elementos de programas de rutinas, tales como la inicialización de bucles y variables y el uso de variables temporales, pueden no mostrarse. Aquellos con conocimientos habituales en la técnica apreciarán que, a no ser que se indique lo contrario en la presente, la secuencia particular descrita es solamente ilustrativa y se puede variar sin desviarse con respecto al espíritu de los conceptos descritos y/o reivindicados en la presente. Por lo tanto, a no ser que se establezca lo contrario, los procesos descritos a continuación no tienen ningún orden lo cual significa que, cuando sea posible, las secuencias mostradas en la Fig. 5 se pueden llevar a cabo en cualquier orden conveniente o deseable.

El método 100 de la Fig. 5 supone que una CDN de interés tiene una serie de pasarelas situadas en o cerca de ubicaciones de usuario que tienen un espacio de almacenamiento digital abundante para su uso en el almacenamiento de contenido en memoria caché. En referencia a continuación a la Fig. 5, la CDN puede recopilar información que se refiere a la composición global de la CDN (bloque 102). Esta información puede incluir, por ejemplo, información sobre el número y las identidades de pasarelas de red asociadas a la CDN, información sobre el número, la identidad, y el tamaño de archivos de película (y/u otros tipos de archivos) asociados a la CDN, información sobre la capacidad de transmisión de las pasarelas asociadas a la CDN, información sobre la capacidad del servidor, información de coste relacionada con la distribución de contenido desde el servidor central, información de coste relacionada con el almacenamiento de contenido en memoria caché en las pasarelas, y/u otra información. También se puede ensamblar información que se refiere a ciertas estadísticas de funcionamiento de la CDN (bloque 104). Parte o la totalidad de esta información estadística puede referirse a estadísticas asociadas a preferencias y requisitos de los usuarios en la CDN. Esta información puede incluir, por ejemplo, probabilidades relacionadas con la demanda de películas en las diversas pasarelas asociadas a la CDN, rango de popularidad de los diversos archivos de película, información relacionada con la demanda de películas específicas en una pasarela, que se almacena en memoria caché en una pasarela diferente, información relacionada con el espacio de almacenamiento usado en diversas pasarelas para almacenar contenido de películas en memoria caché, volumen total medio de demanda para diferentes películas y/u otra información.

A continuación se puede desarrollar una estrategia para almacenar contenido en memoria caché en las pasarelas de las CDN que dará como resultado una distribución eficiente de contenido a usuarios, minimizando una función de coste para el funcionamiento de la CDN que tiene en cuenta costes asociados tanto a la distribución de contenido basada en servidores como a la distribución de contenido basada en pasarelas (bloque 106). Típicamente, el proceso de minimización de la función de coste generará valores para una o más variables de decisión que, a continuación, se pueden usar para determinar qué elementos del contenido se deberían almacenar en memoria caché dentro de qué pasarela. En algunas formas de realización, para minimizar la función de coste se pueden usar una o más técnicas de optimización bien conocidas. Estas técnicas de optimización pueden incluir, por ejemplo, el Esquema de Gradiente de Proximidad General, el Método de Punto Interior, y solucionadores numéricos, tales como GUROBI y CVX. Alternativamente pueden desarrollarse procedimientos de optimización personalizados. En algunas implementaciones, una de las formulaciones descritas en la presente (es decir, la Formulación 1, 2 ó 3) se puede usar para la función de coste. Alternativamente pueden usarse otras formulaciones. La formulación particular que se usa puede depender de la característica específica de la CDN que se esté implementado. Por ejemplo, en una CDN en la que los retardos del servidor pudieran ser significativos, un operador puede decidir usar la Formulación 2 antes descrita, que incluye una penalización por retardo del servidor. En una CDN en la que la varianza en la carga del servidor pueda presentar un problema, un operador puede decidir usar la Formulación 3 que impone una restricción sobre la varianza de carga del servidor.

Después de que se hayan generado valores para las diferentes variables de decisión, parte o la totalidad de estos valores se puede transmitir a las pasarelas de la CDN (bloque 108). Tal como se ha descrito anteriormente, los valores de las variables de decisión serán usados por la CDN (por ejemplo, el servidor central) para determinar cómo se almacenará en memoria caché el contenido dentro de las pasarelas. A continuación, el contenido se distribuirá a los diversos dispositivos de pasarela según la manera deseada, para su almacenamiento en memoria caché (bloque 110). En algunas formas de realización, cuando una pasarela recibe un nuevo contenido de caché, la pasarela puede descartar contenido de caché almacenado previamente antes de almacenar el contenido nuevo. En algunas formas de realización, el servidor central puede determinar qué contenido almacenado en ese momento dentro de una pasarela particular se debería eliminar, y qué contenido nuevo se debería añadir, y puede dar instrucciones a la pasarela para que lleve a cabo esto. Usando este planteamiento, únicamente es necesario distribuir el contenido nuevo a la pasarela.

En algunas formas de realización, el método 100 de la Fig. 5 se puede llevar a cabo de una manera repetitiva durante la vida de una CDN. Por ejemplo, la CDN se puede configurar para repetir el método 100 de manera periódica, continua, o en momentos fijados. Alternativamente, o de forma adicional, la CDN se puede configurar para repetir el método 100 siempre que se detecte una condición predeterminada. En algunas implementaciones, la CDN se puede configurar para que un operador pueda iniciar manualmente el método 100 cuando el primero perciba que puedan existir ineficiencias en la CDN. El objetivo del método 100 puede ser el de lograr un esquema de memorización de almacenamiento de contenido en memoria caché en las pasarelas de una CDN, que dé como resultado una distribución eficiente de contenido a los usuarios. Esta distribución eficiente de contenido puede permitir que usuarios accedan a contenido de una manera puntual con una calidad de servicio (QoS) potenciada. Idealmente, se desarrollará una estrategia de almacenamiento óptimo en memoria caché que minimice el coste de la diseminación de películas en la CDN, pero también se pueden generar esquemas eficientes de almacenamiento en memoria caché que sean inferiores a los óptimos.

En uno de los planteamientos, el método 100 se ejecutará principalmente en un servidor central de una CDN correspondiente, no obstante, también se puede usar una ejecución en otras ubicaciones, incluyendo una

ejecución distribuida en múltiples ubicaciones. En algunas formas de realización, se puede suponer que cualquier fragmento de un archivo de película almacenado en memoria caché y transmitido en una red ya está codificado en red.

Tal como se ha descrito anteriormente, para codificar el contenido que se almacenará dentro de la CDN se usa la codificación de red. Cuando, posteriormente, un usuario solicita el contenido, el contenido codificado se puede distribuir al usuario, y este último tendrá que descodificar el contenido antes de usarlo. Por ejemplo, en una CDN que usa la codificación de red para codificar archivos de película almacenados, cada archivo de película se puede dividir en una serie de segmentos diferentes. A continuación, se pueden generar coeficientes aleatorios para cada uno de los segmentos. Se puede generar, entonces, una combinación lineal de los diferentes segmentos, ponderados por los coeficientes aleatorios, para formar un segmento codificado de la manera siguiente:

$$segmento\ codificado = \sum_{i=1}^{N} a_i S_i$$

donde a_i son los coeficientes aleatorios, S_i son los segmentos de archivo, y N es el número de segmentos de archivo. A continuación, se puede generar una serie de segmentos codificados adicional usando los mismos segmentos de archivo con diferentes coeficientes aleatorios. Para generar los coeficientes aleatorios se puede usar un generador de números aleatorios. Los coeficientes concretos usados para generar un segmento codificado se pueden añadir a cada segmento codificado para su uso eventual en la descodificación. A continuación, los segmentos codificados se pueden almacenar en diversas ubicaciones dentro de una CDN. Debe apreciarse que la técnica antes descrita para implementar la codificación de red en una CDN representa una forma posible de usar la codificación de red. Son también posibles otros planteamientos.

Cuando, posteriormente, un usuario solicita un archivo de película particular, la CDN puede distribuir paquetes codificados al usuario que se corresponden con ese archivo de película, desde cualquier ubicación en la que los mismos estén almacenados. Es importante que, debido a que los segmentos están codificados, no habrá ninguna secuenciación implicada en la distribución de los segmentos al usuario. Es decir, se pueden recuperar segmentos de cualquier ubicación y los mismos se pueden distribuir al usuario sin tener que realizar un seguimiento de los números de secuencia. El usuario tendrá que recibir satisfactoriamente un cierto número de segmentos codificados para poder descodificar el contenido. Típicamente, el número de segmentos codificados requerido será igual que el número de segmentos de archivo N en el que se dividió originalmente el archivo. Además, los segmentos codificados que se reciben deben ser linealmente independientes entre sí para que sean útiles en el proceso de descodificación. Típicamente, el uso de coeficientes generados aleatoriamente, dará como resultado que cada segmento codificado, almacenado, sea linealmente independiente de los otros segmentos codificados. Típicamente, el proceso de descodificación implica la resolución de N ecuaciones lineales para N incógnitas.

En un planteamiento posible, una CDN puede continuar enviando segmentos codificados a un usuario solicitante hasta que se reciba un mensaje de acuse de recibo (ACK) del usuario indicando que se han recibido suficientes segmentos. Alternativamente, una CDN puede enviar, inicialmente, un número fijo de segmentos codificados (por ejemplo, N o mayor) al usuario, y únicamente enviar más si el usuario indica que ello es necesario. Tal como se apreciará, pueden usarse alternativamente otras técnicas para gestionar la distribución de segmentos codificados.

Puesto que es necesario descodificar N paquetes codificados, la CDN puede generar y almacenar más de N paquetes codificados para un archivo particular. El número de segmentos usados y el tamaño de los segmentos pueden variar. En el método 100 antes descrito, y en métodos similares, el proceso de minimización de la función de coste puede tener en cuenta el uso de la codificación de red en la CDN para determinar un esquema de almacenamiento en memoria caché para la CDN. A continuación, se puede desarrollar un esquema de almacenamiento en memoria caché que identifique qué pasarelas van a almacenar en memoria caché segmentos codificados asociados a archivos de contenido particulares.

REIVINDICACIONES

1. Método implementado por máquina para su uso en una red de distribución de contenido, CDN (10), que distribuye contenido a una pluralidad de usuarios, en el que la CDN comprende un servidor central (12) y una pluralidad de dispositivos de pasarela (16), estando el servidor central configurado para gestionar servicios de distribución de contenido, estando la pluralidad de dispositivos de pasarela ubicados en o próximos a unas ubicaciones de usuario (14) e incluyen una capacidad de almacenamiento de datos para su uso en el almacenamiento en memoria caché de por lo menos parte del contenido que se va a distribuir en la CDN, comprendiendo el contenido unos segmentos de archivos codificados en red que se forman generando una combinación lineal de diferentes segmentos de archivo ponderados mediante unos coeficientes de codificación aleatorios, y siendo el método llevado a cabo por el servidor central, comprendiendo el método:

recopilar (102) información sobre una configuración actual de la CDN;

ensamblar (104) estadísticas de funcionamiento de la CDN en correspondencia con requisitos de contenido del usuario;

determinar (106) el contenido que se va a almacenar en memoria caché en la pluralidad de dispositivos de pasarela minimizando una función de coste asociada a la distribución de contenido, teniendo en cuenta la función de coste unos costes asociados a la distribución de contenido desde el servidor y costes asociados a la distribución de contenido desde los dispositivos de pasarela, incluyendo la determinación del contenido usar la información recopilada y las estadísticas ensambladas, y teniendo en cuenta el proceso de minimizar la función de coste el uso de codificación de red en la CDN; y

hacer que el contenido sea enviado a la pluralidad de dispositivos de pasarela para que sea almacenado en memoria caché por ellos de acuerdo con los resultados de dicha determinación.

2. Método según la reivindicación 1, que además comprende que el servidor central:

envíe (110) contenido a la pluralidad de dispositivos de pasarela para que sea almacenado en memoria caché por ellos de acuerdo con los resultados de dicha determinación.

3. Método según la reivindicación 2, que además comprende:

repetir continuamente la recopilación, el ensamblaje, la determinación y el envío.

4. Método según la reivindicación 1, en el que:

la función de coste incluye un término para tener en cuenta retardos de distribución en el servidor.

5. Método según la reivindicación 1, en el que:

la función de coste incluye una restricción sobre la varianza de carga del servidor.

6. Método según la reivindicación 1, en el que:

la determinación de contenido que se va a almacenar en memoria caché en la pluralidad de dispositivos de pasarela incluye usar por lo menos uno de entre los siguientes procedimientos de optimización para minimizar la función de coste: el Esquema de Gradiente de Proximidad General, el método de punto interior, y solucionadores numéricos, tales como GUROBI y CVX.

7. Método según la reivindicación 2, en el que:

el envío de contenido a la pluralidad de dispositivos de pasarela para que sea almacenado en memoria caché por ellos incluye enviar el contenido por medio de Internet.

8. Método según la reivindicación 7, en el que:

el servidor central (12) se comunica con la pluralidad de dispositivos de pasarela a través de un Proveedor de Servicios de Internet, ISP (18), común.

9. Servidor de distribución de contenido (12) para una red de distribución de contenido, CDN (10) que proporciona unos servicios de distribución de contenido para una pluralidad de usuarios, estando el servidor de distribución de contenido configurado para gestionar servicios de distribución de contenido para la CDN, estando el servidor de distribución de contenido configurado para llevar a cabo las etapas según cualquiera de las reivindicaciones 1 a 8.

10. Red de distribución de contenido, CDN, que comprende el servidor de distribución de contenido de la reivindicación 9 y una pluralidad de dispositivos de pasarela, en el que:

la pluralidad de dispositivos de pasarela está bajo el control de la CDN.

11. Red de distribución de contenido, CDN, según la reivindicación 10, que además comprende una pluralidad de centros de datos no ubicados en la ubicación de un servidor central de distribución de contenido, en el que:

la consecución de que el contenido sea enviado a la pluralidad de dispositivos de pasarela incluye hacer que por lo menos parte del contenido sea enviado a los dispositivos de pasarela desde uno o más de entre dicha pluralidad de centros de datos.

12. Servidor de distribución de contenido según la reivindicación 9, en el que:

la función de coste usa un coste para que el servidor de distribución de contenido transmita una unidad de contenido a una pasarela y un coste de almacenamiento en memoria caché de una unidad de contenido en una pasarela.

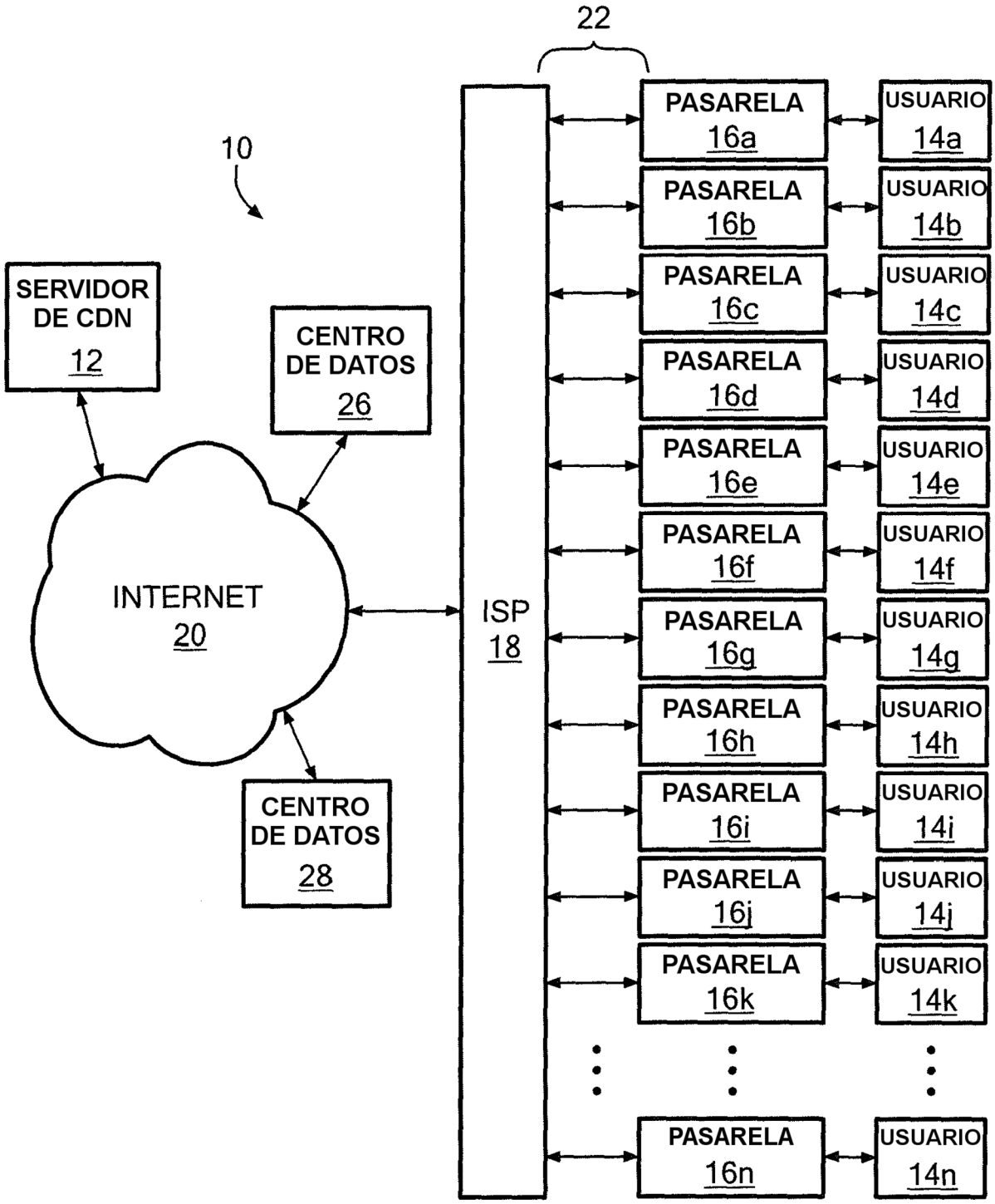
22
10
SERVIDOR DE CDN 12
CENTRO DE DATOS 26
INTERNET 20
CENTRO DE DATOS 28
ISP 18
PASARELA 16a
PASARELA 16b
PASARELA 16c
PASARELA 16d
PASARELA 16e
PASARELA 16f
PASARELA 16g
PASARELA 16h
PASARELA 16i
PASARELA 16j
PASARELA 16k
PASARELA 16n
USUARIO 14a
USUARIO 14b
USUARIO 14c
USUARIO 14d
USUARIO 14e
USUARIO 14f
USUARIO 14g
USUARIO 14h
USUARIO 14i
USUARIO 14j
USUARIO 14k
USUARIO 14n

FIG. 1

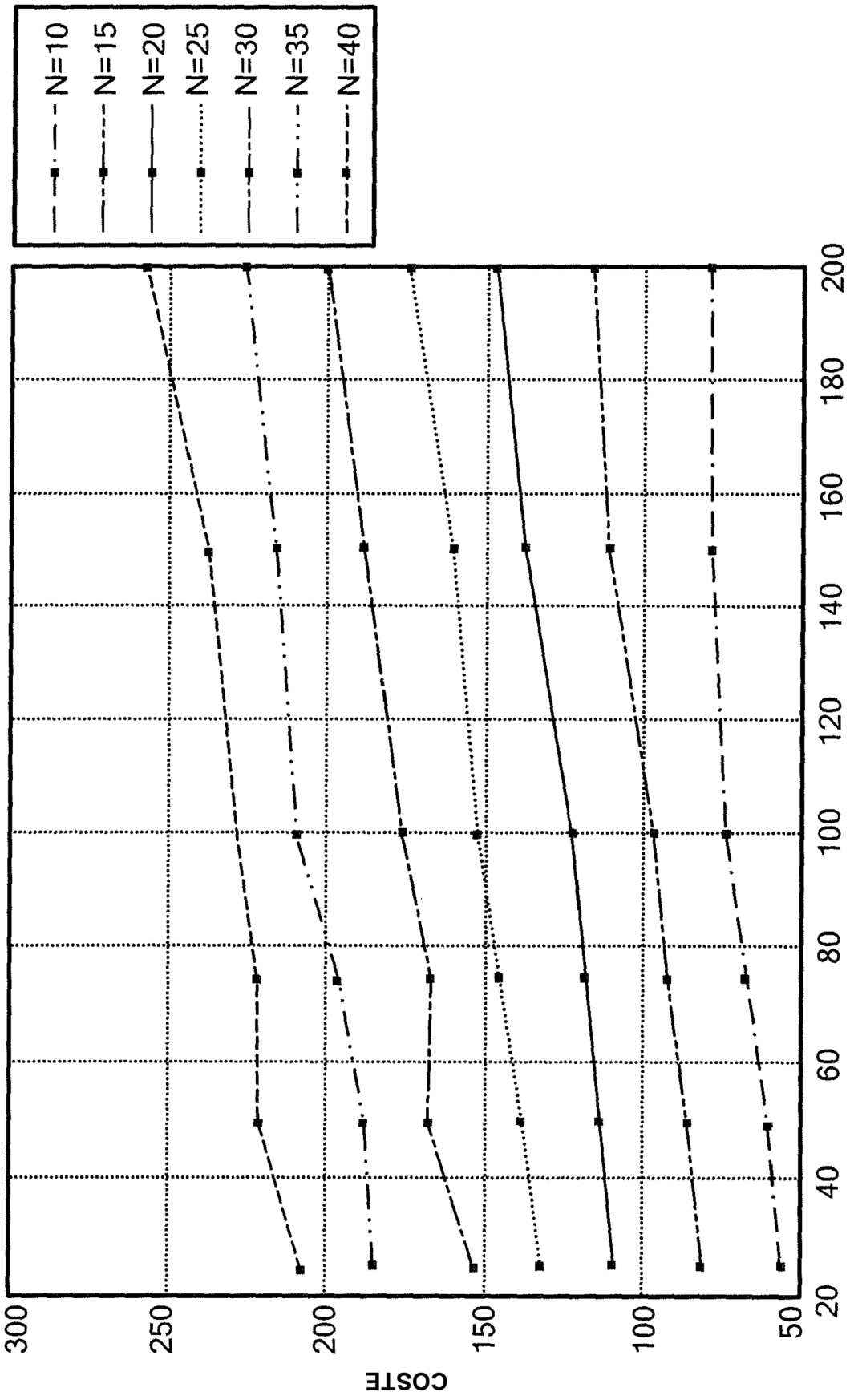
N=10
N=15
N=20
N=25
N=30
N=35
N=40
COSTE
300
250
200
150
100
50
20
40
60
80
100
120
140
160
180
200
M

FIG. 2

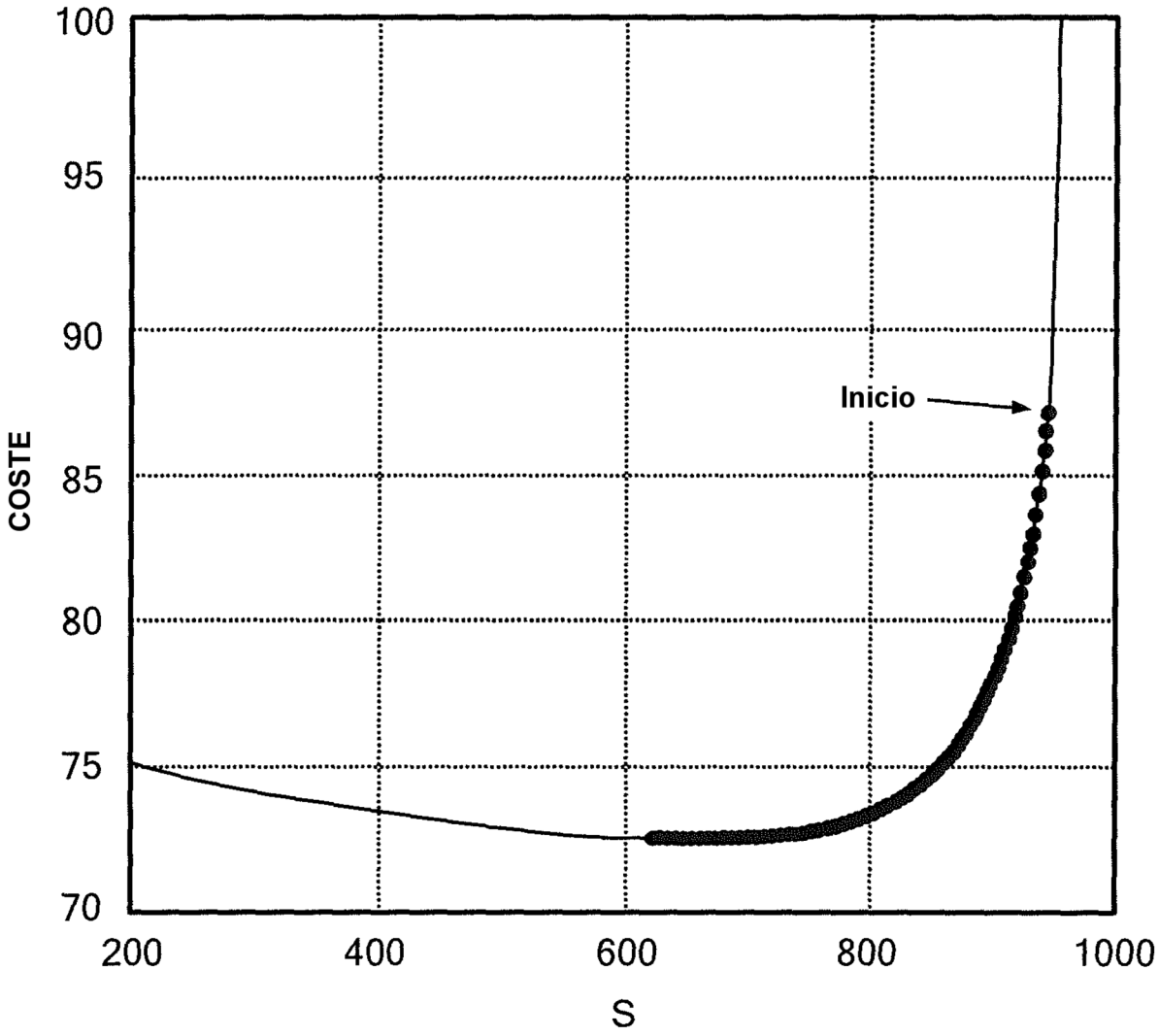

FIG. 3

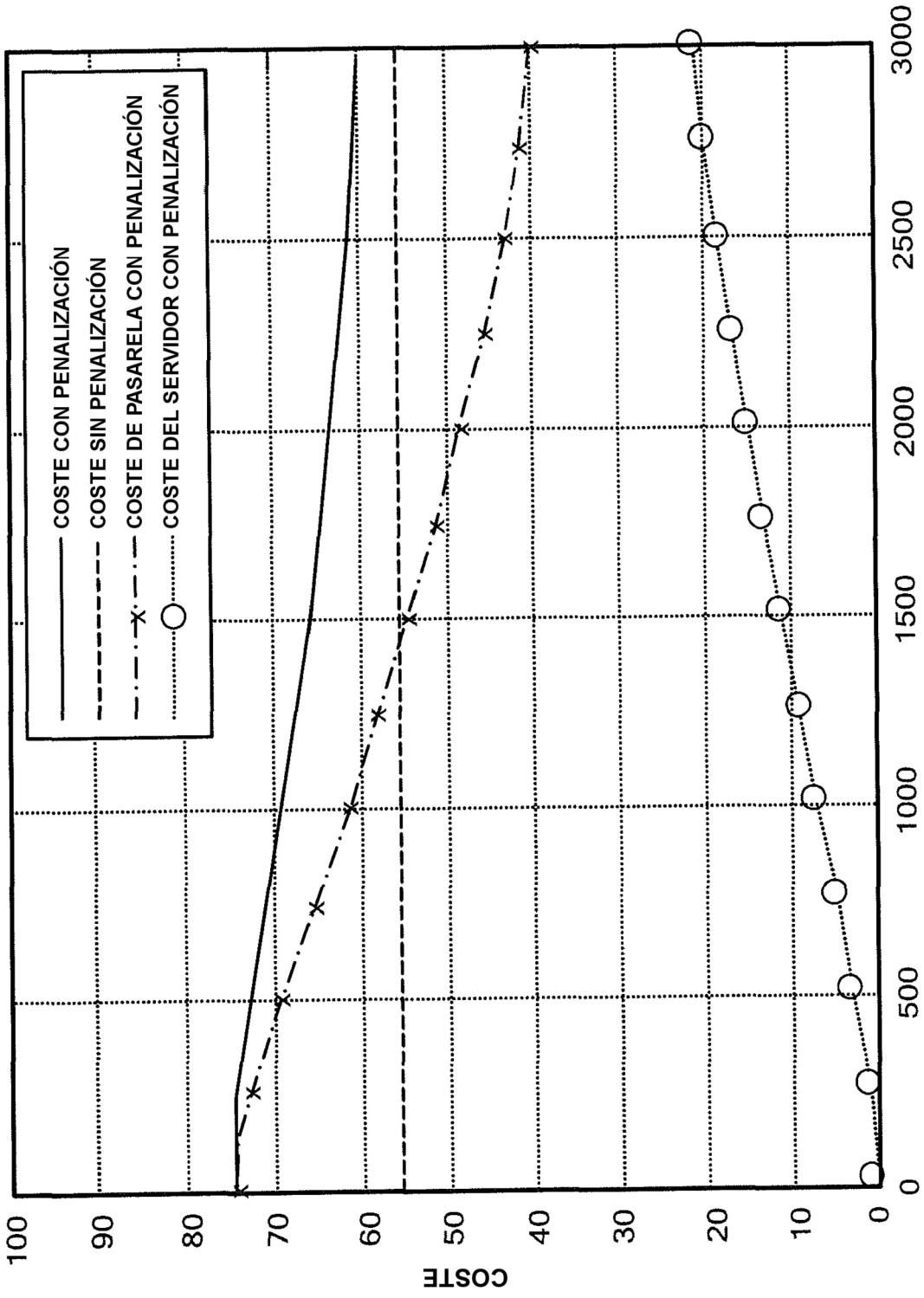
COSTE CON PENALIZACIÓN
COSTE SIN PENALIZACIÓN
COSTE DE PASARELA CON PENALIZACIÓN
COSTE DEL SERVIDOR CON PENALIZACIÓN
COSTE
100
90
80
70
60
50
40
30
20
10
0
0
500
1000
1500
2000
2500
3000
S_c

FIG. 4

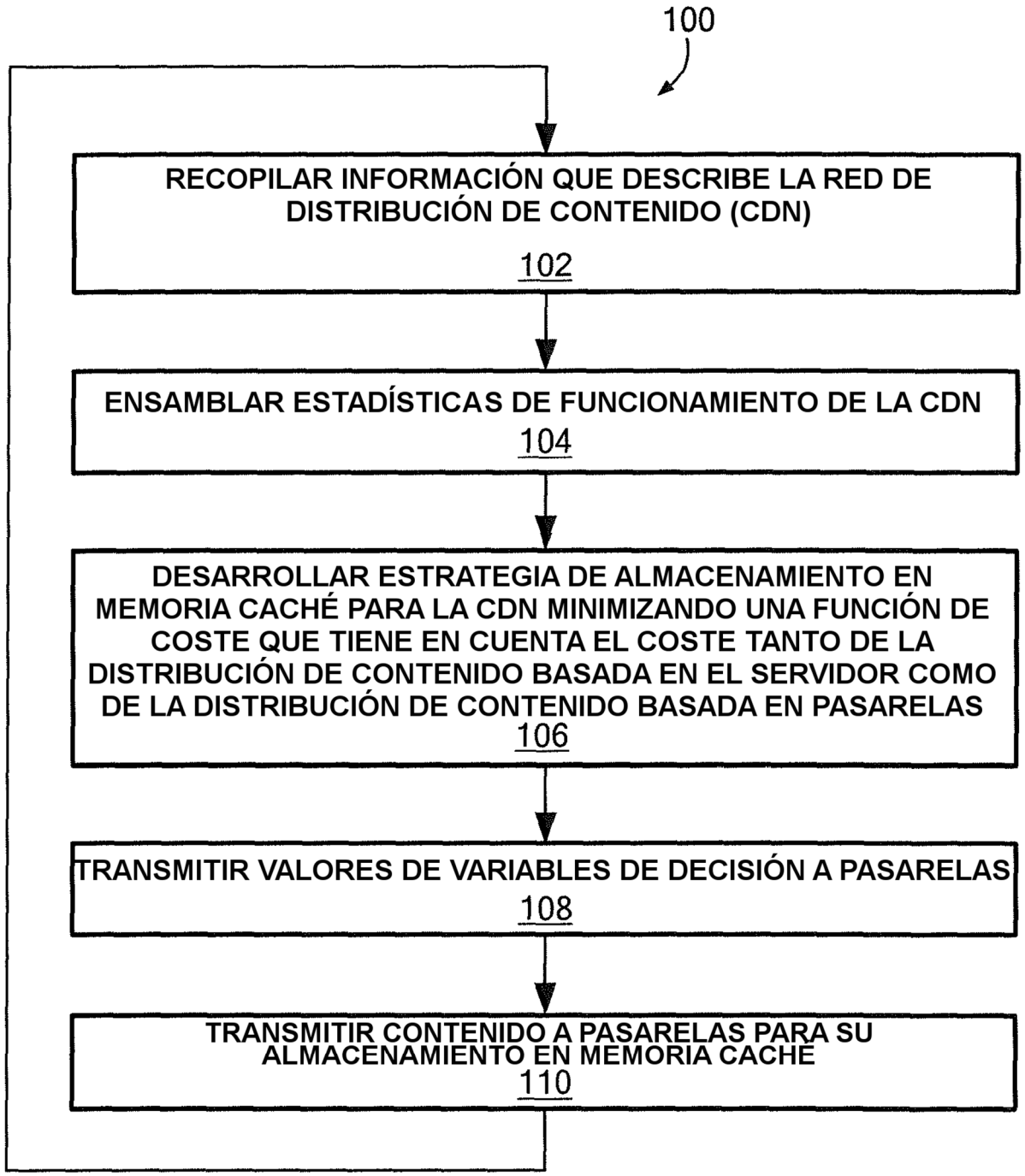
100
RECOPILAR INFORMACIÓN QUE DESCRIBE LA RED DE DISTRIBUCIÓN DE CONTENIDO (CDN)
102
ENSAMBLAR ESTADÍSTICAS DE FUNCIONAMIENTO DE LA CDN
104
DESARROLLAR ESTRATEGIA DE ALMACENAMIENTO EN MEMORIA CACHÉ PARA LA CDN MINIMIZANDO UNA FUNCIÓN DE COSTE QUE TIENE EN CUENTA EL COSTE TANTO DE LA DISTRIBUCIÓN DE CONTENIDO BASADA EN EL SERVIDOR COMO DE LA DISTRIBUCIÓN DE CONTENIDO BASADA EN PASARELAS
106
TRANSMITIR VALORES DE VARIABLES DE DECISIÓN A PASARELAS
108
TRANSMITIR CONTENIDO A PASARELAS PARA SU ALMACENAMIENTO EN MEMORIA CACHÉ
110

FIG. 5