

REPUBLICA DE COLOMBIA
SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO
DIRECCION DE NUEVAS CREACIONES

SOLICITUD DE PATENTE DE INVENCION

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO



No. 16-029128- -00000-0000

Fecha: 2016-02-08 12:28:57 Dep. 2020 DIR NUEVASCR
Tra. 2 PATENTES Eve: 1 REGDEPOSITO
Act. 411 PRESENTACION Folios 33

EXPEDIENTE

:

62

TITULO:

:

Procedimiento, Sistema y Dispositivo para Gestionar
la Congestión en Servicios de Red

CLASE

:

H04L29/08

INVENTORES

:

José Alfredo Picón Cabezudo / Paulo Villegas Nuñez
Juan José Andrés Gutiérrez

SOLICITANTE

:

Telefónica Digital España, S.L.U.

DOMICILIO

:

Madrid, España

APODERADO

:

Juan Carlos Cuesta Quintero

FECHA

:

Febrero de 2016

PETITORIO

Industria y Comercio
SUPERINTENDENCIA

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO
No. 16-029128-00000-0000
Fecha 2016-02-08 12:28:57 Dep. 2020 DIR NUEVASCR
Tra 2 PATENTES Eve 1 REGDEPOSITO
Act 411 PRESENTACION Folios 33

FORMULARIO UNICO DE SOLICITUD DE PATENTE

GL

①

SOLICITUD DE:

☒ X

Patente de Invención

☐

Patente de Modelo de Utilidad

②

SOLICITANTE

(71)

Nombre: Telefónica Digital España, S.L.U.

Dirección: C/Gran Vía 28, 28013, Madrid, España

Nacionalidad o Domicilio: España

Lugar de Constitución España

Teléfono:

Fax:

E-mail:

IDENTIFICACION

C.C.

☐

NIT

☐

C.E.

☐

Otro

☒ X

Número

Cual

Empresa Extranjera

③

APODERADO

Nombre Juan Carlos Cuesta Quintero

Dirección Calle 72 No. 10-07 Of. 603

Teléfono: 2113097 Fax: 2554538

E-mail: juancuesta@cuestalawyers.com

IDENTIFICACION

C.C.

☒ X

NIT

☐

C.E.

☐

Otro

☐

Número 79.339.809 TP 43.273

④

INVENTORES

(72)

Nombres: José Alfredo Picón Cabezudo / Paulo Villegas Núñez / Juan José Andrés Gutiérrez

Dirección: C/Gran Vía 28, 28013, Madrid, España

Nacionalidad o Domicilio: Españoles

⑤ TITULO (54)

Procedimiento, Sistema y Dispositivo para Gestionar la Congestión en Servicios de Red.

⑥

Clasificación Internacional (51)

H04L29/08

⑦

Prioridad

Si ☒ X

No ☐

(33) País de Origen

Unión Europea

(32) Fecha

10/02/2015

(31) Número de Solicitud

15382046.9

⑧

Para publicar a partir de la fecha de la presente solicitud a los :

6 meses

☐

12 meses

☒ X

18 meses

☐

Otro

☐

Cual

⑨

Comprobante de pago No.

16-

Fecha

8 de Febrero de 2016

Instrucciones para el diligenciamiento del presente formulario, ver reverso de esta página.

10

ANEXOS

2

- ☒ Comprobante de pago de la tasa de presentación de la solicitud
- ☐ Comprobante de pago de la tasa por concepto de excedente de palabras en la publicación
- ☒ Comprobante de pago de la tasa por concepto reivindicación de prioridad
- ☒ Poderes, si fuere el caso
- ☐ Documento que acredite la existencia y representación legal cuando el solicitante sea persona jurídica
- ☒ Poderes, si fuere el caso
- ☒ Certificado de la fecha de presentación de la solicitud prioridad expedida por la autoridad correspondiente y una copia certificada de la primera solicitud, si se reivindica prioridad.
- ☒ Traducción simple de la primera solicitud, si se reivindica prioridad
- ☒ Documento de cesión del inventor al solicitante o a su causante.
- ☒ Resumen
- ☒ Descripción de la invención
- ☒ Una o más reivindicaciones
- ☒ Dibujos y/o planos necesarios.
- ☐ De ser el caso, copia del contrato de acceso.
- ☐ De ser el caso, documento que acredite la licencia o autorización de uso de conocimientos tradicionales de las comunidades indígenas.
- ☐ De ser el caso, certificado de depósito del material biológico
- ☒ Arte final 12 X 12
- ☐ De ser el caso, información sobre otras solicitudes de patente o títulos obtenidos en el extranjero por el mismo titular o su causante, relacionadas parcial o totalmente con la invención de esta solicitud.

11

FIGURA CARACTERISTICA:

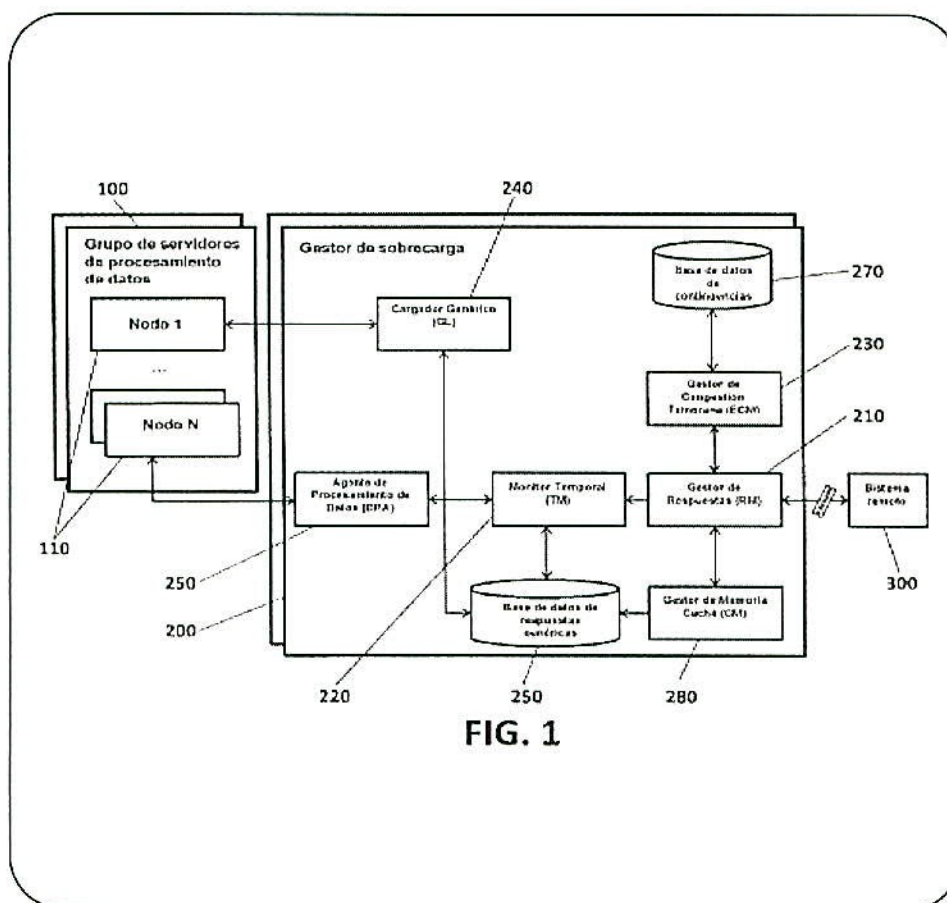


FIG. 1

12

Solicito la concesión de la patente,

NOMBRE: Juan Carlos Cuesta Quintero

FIRMA:

Juan Carlos Cuesta Quintero

C.C. 79.339.809

T.P. 43.273 del C.S. de la J.

		EXTRACTO PARA PUBLICACIÓN	
		☒ Patente de Invención ☐ Patente modelo de utilidad ☐ PCT	
(21) No. De solicitud		(51) int. Cl: H04L29/08	
(22) Fecha de solicitud		(71) Solicitante (s) Telefónica Digital España, S.L.U.	
(30) Prioridad:		(72) Interventor (es)	
(31) No. Prioridades: 15382046.9		José Alfredo Picón Cabezudo; (ES)	
(32) Fecha: 10/02/2015		Paulo Villegas Núñez; (ES)	
(33) País: Unión Europea (EP)		Juan José Andrés Gutiérrez; (ES)	
		(74) Apoderado (s) Juan Carlos Cuesta Quintero	
(54) Título Procedimiento, Sistema y Dispositivo para Gestionar la Congestión en Servicios de Red.			
<i>Las casillas identificadas con 85, 86 y 87 sólo aplican para PCT</i>			
(85) Fecha límite inicio Fase Nacional			
(86) Datos relativos a la presentación de la solicitud PCT Fecha de presentación de la solicitud Fecha: No. de solicitud PCT:		(87) Publicación internacional Fecha No. Publicación	

Resumen reivindicación (es):

Un procedimiento, un sistema y un dispositivo para gestionar estados de congestión / sobrecarga al nivel de aplicación, experimentados en un servicio de red proporcionado mediante una red de telecomunicaciones, que permite proporcionar respuestas de servicio rápidas, válidas y completas al cliente del servicio incluso cuando los servidores del servicio están sobrecargados. La gestión de congestión es proporcionada por un nodo intermedio comunicado con el dispositivo cliente y con el servidor, o servidores, que proporciona(n) el servicio de red. A fin de controlar la sobrecarga de uno o más servidores, este nodo usa un doble mecanismo de congestión, un mecanismo de detección de congestión temprana y un mecanismo de limitación del tiempo de respuesta.

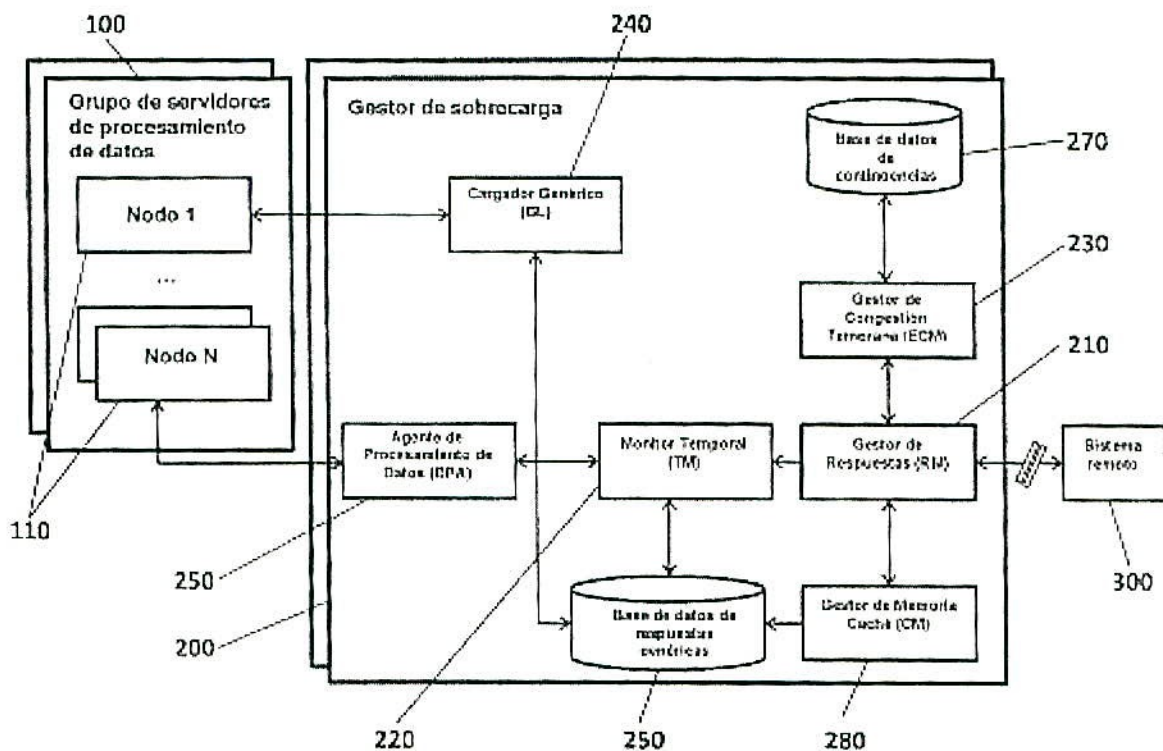


FIG. 1

PROCEDIMIENTO, SISTEMA Y DISPOSITIVO PARA GESTIONAR LA
CONGESTIÓN EN SERVICIOS DE RED

DESCRIPCIÓN

5

CAMPO TÉCNICO DE LA INVENCION

10

La presente invención se refiere a servicios de telecomunicaciones y, más específicamente, a un procedimiento, un sistema y un dispositivo para gestionar estados de congestión (sobrecarga), a nivel de aplicación, experimentados en servicios de red proporcionados a través de una red de telecomunicación.

15

ANTECEDENTES DE LA INVENCION

20

La cantidad de servicios proporcionados a nivel de aplicación, a través de redes de telecomunicaciones (servicios de red o en red), aumenta todos los días, ofreciendo a los usuarios finales (o nodos de red) toda clase posible de funciones o aplicaciones. Los servicios de red están creciendo aceleradamente en popularidad y en capacidades, por lo que la congestión en los servicios de red está tornándose un problema recurrente. Puede aparecer en muchos niveles en una pila de red, desde la congestión a nivel de paquete hasta cuellos de botella de aplicaciones que retardan los servicios.

25

30

A nivel de paquete, la congestión de red se produce sencillamente cuando un nodo encaminador o un enlace recibe tráfico que supera su capacidad. En el peor de los casos, puede llegar al colapso de la red (el sector afectado entra en un punto muerto y se torna incapaz de encaminar paquetes). A este bajo nivel, el control de congestión del TCP (Protocolo de Control de Transmisión) ya establece mecanismos mediante los cuales los clientes ajustan su transmisión a fin de controlar el caudal y evitar la congestión. No obstante, incluso con tales mecanismos disponibles, la congestión aún puede ocurrir si hay picos repentinos en el número de clientes que intentan

35

enviar tráfico. La gestión activa de colas brinda una alternativa a nivel de encaminamiento: por ejemplo, un nodo de red sobrecargado puede usar técnicas tales como la Detección Aleatoria Temprana y sus variantes, que monitorizan el tamaño de la cola o la información de carga para descartar paquetes cuando sea necesario.

Algunas patentes publicadas que describen soluciones para la congestión son, por ejemplo:

Patente US 8762567 B2: Describe un procedimiento mediante el cual un procesador controla la congestión estableciendo una fracción de aceptación de mensajes (descartando los mensajes que caen fuera de esa fracción). La fracción de aceptación está determinada por la velocidad de llegada, de una manera no especificada. Usa una capacidad normalizada de procesador para prever la congestión.

Patente US 7187651 B1: Propone un sistema al nivel de la red, que almacena un conjunto de factores de carga y asocia a cada uno de ellos una razón de rechazo de paquetes, que se aplica cuando se alcanza el factor de carga. El mecanismo parece estar predefinido y fijado. Puede diferenciar entre paquetes para la inicialización de mensajes y paquetes para comunicaciones establecidas, a fin de descartar preferiblemente paquetes para nuevas llamadas.

Patente US 7266612 B1: Propone un mecanismo para controlar la sobrecarga de la red en la capa de paquetes TCP, combinando la invención con el uso de normas de QoS (calidad de servicio). El objetivo de esta invención es controlar el tráfico a nivel de red y no a nivel de aplicación. Usa múltiples umbrales para distintas clases de tráfico.

Sin embargo, las soluciones existentes para la congestión presentan distintos inconvenientes. Las soluciones de bajo nivel (por ejemplo, en la capa de paquetes de la red) solamente

pueden funcionar a ciegas, dado que no pueden evaluar el contenido del mensaje. Por lo tanto, su política de descarte perjudicará a los clientes de manera indiscriminada. Algunas de ellas pueden establecer distintas clases de prioridad, para mantener un mayor caudal para paquetes de alta prioridad, pero esto todavía es bastante a ciegas, dado que la semántica de los datos se mantiene fuera de la competencia del mecanismo de control de congestión. Además, dado que carecen de información sobre el servicio de procesamiento que produce los mensajes, la detección de congestión solamente puede hacerse analizando al nivel de la red parámetros tales como la probabilidad de pérdida de paquetes de la red, o la longitud de colas de paquetes. Finalmente, la única medida que pueden tomar esas soluciones es descartar paquetes de la red para evitar la congestión, afectando por ello al servicio que lleva esos paquetes de una manera desconocida, pero posiblemente perjudicial.

Las soluciones de nivel superior, funcionando a nivel de aplicación, tienen un mejor comportamiento, ya que pueden identificar comunicaciones de mensajes y discernir el tipo de servicio que representan; por lo tanto, potencialmente, podrían tomar una decisión más informada tanto para (a) identificar la congestión como para (b) tomar contramedidas. Sin embargo, las soluciones existentes identifican la congestión, mayormente, rastreando velocidades de origen, en lugar de monitorizar la capacidad de procesamiento. Y la contramedida es descartar mensajes y, dado que la medida perjudicará al servicio, se demora hasta que sea absolutamente necesario (cuando el servicio ya está sobrecargado).

Las realizaciones propuestas de la invención, expuestas más adelante, superarán al menos algunos de los inconvenientes de las soluciones de la técnica anterior, proporcionando una solución de alto nivel para la congestión que, en lugar de descartar mensajes, devuelve, toda vez que sea posible, una respuesta sub-óptima, pero todavía útil (que requiere bajo poder de cálculo y que está generada de antemano) y, además,

detecta activamente la tendencia hacia la congestión y proporciona medidas mitigadoras (pero no perjudiciales) con antelación.

5 SUMARIO DE LA INVENCION

Los problemas hallados en las técnicas de la tecnología anterior son generalmente resueltos o eludidos, y son logradas, en general, ventajas técnicas por las realizaciones descritas, que proporcionan un procedimiento, un sistema y un dispositivo para gestionar la congestión en servicios de red al nivel de aplicación.

Algunos sistemas de información (p. ej., servicios web) pueden tener una carga de trabajo sumamente variable. La solución tradicional para este problema es construir sistemas de información que pueden ser ajustados a escala horizontalmente. Sin embargo, la carga de trabajo puede aumentar inesperadamente, más allá de la capacidad del sistema en un momento dado y, si la carga de trabajo recibida es demasiado alta, puede ocurrir una caída del sistema. Lo mismo vale para otros servicios de red que puedan tener una carga de trabajo sumamente variable.

La presente invención resuelve este problema de una forma distinta a la tradicional. La invención presenta un sistema (nodo) intermedio, entre un sistema remoto (el dispositivo o nodo cliente, es decir, la entidad a la cual se proporciona el servicio) y un sistema de procesamiento de datos (en términos generales, el servidor o grupo de servidores, que brindan el servicio de red), que pretende generar respuestas rápidas, válidas y completas incluso cuando el sistema de procesamiento de datos está sobrecargado (congestionado). De tal modo, es capaz de mantener la capacidad de respuesta del sistema remoto en cualquier situación de la carga de trabajo.

La invención está parcialmente basada en el siguiente principio: En situaciones de alta carga de trabajo es mejor

generar una respuesta que contenga datos no completamente exactos, que generen datos exactos que provoquen una degradación progresiva del sistema. La presente invención pre-carga un mensaje de respuesta genérica para todos los usuarios (clientes), previamente recuperada del servidor, y cuando el servidor esté acercándose a situaciones de carga alta, las solicitudes de servicio no serán descartadas, sino que serán respondidas usando dicha respuesta genérica (por lo que las solicitudes no serán procesadas por el servidor y, por lo tanto, la situación de carga alta del servidor no se degrada). La respuesta genérica será válida para todos los usuarios (o al menos para la mayoría de ellos), pero sub-óptima en alguna medida con respecto a la que podrían recibir siguiendo la ruta estándar (a través de un servidor totalmente funcionando). Esta respuesta genérica se actualiza periódicamente.

La invención propuesta incluye dos mecanismos simultáneos para controlar la sobrecarga del servidor y para generar respuestas genéricas en lugar de las normales:

Un umbral de tiempo absoluto para limitar el tiempo de respuesta. Toda solicitud desde el cliente al servidor es monitorizada y un contador interno acumula el tiempo consumido desde que fuera recibida la solicitud. Si se supera un umbral configurable, el procesamiento actual es abortado y se devuelve una respuesta genérica.

Una detección temprana de congestiones para prever sobrecargas del servidor. El tiempo medio de respuesta del servidor es calculado continuamente (como un promedio móvil) y, cuando comienza a hacerse alto, una parte de las solicitudes no es remitida a (y, por consiguiente, no procesada por) el servidor y, en cambio, se devuelve una respuesta normal, una genérica (no específica para dicho usuario). La parte de solicitudes no remitidas al servidor (pero contestadas con una respuesta genérica) depende del tiempo de respuesta monitorizado con una función escalonada; el objetivo es mantener el servidor en carga máxima (cerca de la saturación) pero sin sobrecargarlo.

En un primer aspecto, se propone un procedimiento para gestionar la congestión en un servicio de red proporcionado por un grupo de servidores (uno o más servidores), estando caracterizado el procedimiento por comprender las siguientes etapas, realizadas en un primer nodo de la red (gestor de sobrecarga) en comunicación con el grupo de servidores:

a) Recibir una solicitud de servicio desde un dispositivo cliente del servicio, mediante una primera red de telecomunicación;

b) Después de recibir la solicitud de servicio, determinar si hay o no un estado de sobrecarga en el grupo de servidores;

c) Decidir si se activa un mecanismo de contingencia basado al menos en la determinación;

d) Si el mecanismo de contingencia es activado para dicha solicitud de servicio, no enviar la solicitud de servicio al grupo de servidores;

e) Si el mecanismo de contingencia no ha sido activado (en la etapa c):

e1) Enviar la solicitud de servicio al grupo de servidores;

e2) Si una respuesta de servicio, resultante del procesamiento de la solicitud de servicio por parte del grupo de servidores, es recibida desde el grupo de servidores antes de que haya transcurrido un primer periodo de tiempo (desde que la solicitud de servicio fuera enviada al grupo de servidores), enviar dicha respuesta de servicio al dispositivo cliente y, si no es así (la respuesta de servicio no llega antes de que se agote el primer periodo de tiempo), activar el mecanismo de contingencia para dicha solicitud de servicio;

f) Si se activa el mecanismo de contingencia (en las etapas c) o e2)) para dicha solicitud de servicio, obtener desde una base de datos una respuesta genérica, no resultante del procesamiento por el grupo de servidores de la solicitud de servicio, y enviar dicha respuesta genérica al dispositivo (terminal) cliente.

Si el primer nodo de red no recibe una respuesta de servicio desde el grupo de servidores antes de que haya transcurrido el primer periodo de tiempo, el nodo de red puede enviar un mensaje al grupo de servidores para abortar el procesamiento de la solicitud de servicio.

La etapa c) de decidir si se activa un mecanismo de contingencia puede comprender:

- si se determina que no hay estado de sobrecarga en el grupo de servidores, no activar el mecanismo de contingencia;

- si se determina que hay estado de sobrecarga, el mecanismo de contingencia es activado, no para todas las solicitudes de servicio, sino solamente para algunas de ellas. A fin de hacerlo, por ejemplo, se calcula una probabilidad de activar el mecanismo de contingencia para dicha solicitud de servicio, y el mecanismo de contingencia es activado según dicha probabilidad. Dicha probabilidad puede ser calculada en base a parámetros de diseño y a los valores del tiempo de ejecución de solicitudes para las anteriores M solicitudes de servicio recibidas previamente en el primer nodo de red, donde M es un parámetro de diseño.

La determinación del estado de sobrecarga en el grupo de servidores puede estar basada en los valores de un tiempo de ejecución de solicitud para las M solicitudes de servicio precedentes, recibidas en el primer nodo de red, donde M es un parámetro de diseño (por ejemplo, un promedio de todos los valores, una mediana, una varianza o cualquier otro parámetro estadístico basado en dichos valores).

La determinación de si hay o no estado de sobrecarga temprana en el grupo de servidores puede comprender:

5 - calcular el promedio de los valores del tiempo de ejecución de solicitud para las M solicitudes anteriores recibidas en el primer nodo de red, donde M es un parámetro de diseño;

10 - si dicho promedio calculado está por encima de un primer umbral, determinar que hay estado de sobrecarga temprana en el grupo de servidores.

15 Si se activa el mecanismo de contingencia, puede ser estimado un tiempo (hipotético) de ejecución de solicitud para dicha solicitud de servicio, y almacenarse en una base de datos en el primer nodo de red (por ejemplo, para su uso para la detección de condiciones de sobrecarga temprana para futuras solicitudes de servicio).

20 Si el mecanismo de contingencia no es activado para dicha solicitud de servicio, puede calcularse un tiempo de ejecución de solicitud (tiempo consumido por el grupo de servidores en el procesamiento de la solicitud y el envío de la respuesta al primer nodo de red) como el tiempo transcurrido desde que la solicitud de servicio es enviada al grupo de servidores, hasta
25 que la respuesta de servicio es recibida por el primer nodo de red, y dicho tiempo de ejecución de solicitud es almacenado en una base de datos en el primer nodo de red (por ejemplo, para su uso para la detección de condiciones de sobrecarga temprana para futuras solicitudes de servicio).

30 El primer nodo de red puede estar co-situado con el grupo de servidores, o puede no estar co-situado con el grupo de servidores. En este último caso, se comunica con el grupo de servidores a través de las primeras redes de
35 telecomunicaciones, o a través de una red de telecomunicaciones distinta. El grupo de servidores puede ser un grupo de servidores de procesamiento de datos.

En una realización, la solicitud de servicio y la respuesta de servicio (y la respuesta de servicio genérica) son mensajes de la capa (o el nivel) de aplicación de red.

5 El dispositivo cliente puede ser uno entre el grupo siguiente:
un teléfono móvil, un ordenador portátil, un ordenador
personal, un teléfono inteligente, un televisor inteligente,
un televisor, un televisor de IPTV (TeleVisión del Protocolo
10 de Internet), una tableta o cualquier otro dispositivo
electrónico.

En una realización, la etapa e1) de envío de la solicitud de servicio al grupo de servidores comprende:

- 15 - Descomponer la solicitud de servicio en varias sub-solicitudes (también llamadas solicitudes individuales),
- Enviar la primera sub-solicitud al grupo de servidores y
- 20 - Enviar cada una del resto de sub-solicitudes al grupo de servidores cuando se recibe una respuesta parcial, resultante del procesamiento por el grupo de servidores de la sub-solicitud precedente (es decir, cada sub-solicitud es enviada al grupo de servidores solamente cuando ha sido recibida la
- 25 respuesta a la sub-solicitud previa desde el grupo de servidores);
- y la etapa e2) comprende:

- 30 - Si todas las respuestas de servicio parciales, resultantes del procesamiento por el grupo de servidores de cada sub-solicitud, son recibidas desde el grupo de servidores antes de que haya transcurrido el primer periodo de tiempo, componer una respuesta de servicio basada en todas las respuestas parciales recibidas, y enviar dicha respuesta de servicio al
- 35 nodo cliente y, si no es así, activar el mecanismo de contingencia.

El servicio de red puede ser un servicio de información (el servicio proporcionado consiste en información incluida en la respuesta de servicio). En este caso, la respuesta de servicio resultante del procesamiento de la solicitud de servicio por el grupo de servidores incluye información seleccionada en base al contenido de la solicitud de servicio y / o al perfil del cliente, y la respuesta genérica consiste en información no seleccionada en base al contenido de la solicitud de servicio y / o al perfil del cliente (puede incluso ser seleccionada antes de recibir la solicitud de servicio). Por ejemplo, el servicio puede ser un servicio de recomendación de contenidos o un servicio de noticias, y la respuesta de servicio consiste en un contenido recomendado o noticias seleccionadas sobre la base de al menos el perfil del cliente y / o el contenido de la solicitud. Y la respuesta genérica puede consistir en un contenido recomendado, o en noticias, válidos para todos los clientes (no seleccionados para un cliente específico).

La respuesta de servicio genérica puede ser pre-almacenada en una base de datos (p. ej., en una memoria caché del primer nodo de red); es decir, la respuesta genérica puede estar almacenada antes de que la solicitud de servicio haya llegado al primer nodo de red. O bien la respuesta genérica puede estar basada en datos genéricos recibidos por el primer nodo de red desde el grupo de servidores, antes de que la solicitud de servicio haya llegado al primer nodo de red. O bien la respuesta genérica puede estar basada en datos genéricos recibidos por el primer nodo de red desde el grupo de servidores, antes de que la solicitud de servicio haya llegado al primer nodo de red; la respuesta genérica no es una respuesta específica generada para dicha solicitud de servicio y / o perfil de cliente, sino que es independiente del perfil del cliente que solicita el servicio (o bien, en otras palabras, la respuesta genérica puede ser válida para todos los clientes del servicio).

En un segundo aspecto, se proporciona un nodo de red para gestionar la congestión de un servicio de red, estando el nodo de red caracterizado por comprender:

- Medios de recepción para recibir una solicitud de servicio desde un dispositivo cliente del servicio, mediante una primera red de telecomunicación;

5

- Medios para determinar si hay o no un estado de sobrecarga y decidir si se activa un mecanismo de contingencia basado al menos en la determinación;

10

- Medios para, si el mecanismo de contingencia es activado para dicha solicitud de servicio, no enviar la solicitud de servicio a un grupo de servidores que proporcionan el servicio de red y, si el mecanismo de contingencia no es activado, para:

15

- Enviar la solicitud de servicio al grupo de servidores que proporcionan el servicio de red;

20

- Si una respuesta de servicio, resultante del procesamiento de la solicitud de servicio por el grupo de servidores, es recibida desde el grupo de servidores antes de que haya transcurrido un primer periodo de tiempo, enviar dicha respuesta de servicio al dispositivo cliente y, si no es así, activar el mecanismo de contingencia para dicha solicitud de servicio;

25

- Medios para, si el mecanismo de contingencia es activado para dicha solicitud de servicio, obtener desde una base de datos una respuesta genérica pre-almacenada, no resultante del procesamiento por el grupo de servidores de la solicitud de servicio, y enviar dicha respuesta genérica al dispositivo cliente.

30

En un tercer aspecto, se proporciona un sistema para gestionar la congestión de un servicio de red, estando el sistema caracterizado por comprender:

35

- Un grupo de servidores configurados para proporcionar el servicio de red, recibiendo solicitudes de servicio de

clientes, procesando las solicitudes y generando respuestas de servicio basadas al menos en el contenido de las solicitudes de servicio de clientes;

5 - Un nodo de red que comprende:

10 - Medios de recepción para recibir una solicitud de servicio desde un dispositivo cliente del servicio, a través de una primera red de telecomunicación;

15 - Medios para determinar si hay o no un estado de sobrecarga en el grupo de servidores, y para decidir si se activa un mecanismo de contingencia basado al menos en la determinación;

20 - Medios para, si el mecanismo de contingencia es activado para dicha solicitud de servicio, no enviar la solicitud de servicio al grupo de servidores;

 - Medios, si el mecanismo de contingencia no es activado, para:

 - Enviar la solicitud de servicio al grupo de servidores;

25 - Si una respuesta de servicio, resultante del procesamiento de la solicitud de servicio por el grupo de servidores, es recibida desde el grupo de servidores antes de que haya transcurrido un primer periodo de tiempo, enviar dicha respuesta de servicio al dispositivo cliente y, si no es así, activar el mecanismo de contingencia para dicha solicitud de servicio;

30 - Medios para, si el mecanismo de contingencia es activado para dicha solicitud de servicio, obtener desde una base de datos una respuesta genérica pre-almacenada, no resultante del procesamiento por el grupo de servidores

de la solicitud de servicio, y enviar dicha respuesta genérica al dispositivo cliente.

En una realización, los medios indicados en lo que antecede pueden ser un procesador en el nodo de red configurado para realizar las acciones precitadas.

De acuerdo a otro aspecto de la invención, se proporciona un producto de programa de ordenador, que comprende instrucciones ejecutables por ordenador para realizar cualquiera de los procedimientos previamente descritos, cuando el programa es ejecutado en un ordenador, y también se proporciona un medio de almacenamiento de datos digitales, que codifica un programa, ejecutable por máquina, de instrucciones para realizar cualquiera de los procedimientos descritos.

En consecuencia, de acuerdo a la invención, se proporcionan un procedimiento, un sistema, un dispositivo, un programa de ordenador y un medio de almacenamiento, de acuerdo a las reivindicaciones independientes. Las realizaciones favorables están definidas en las reivindicaciones dependientes. Estos y otros aspectos de la invención serán evidentes a partir de, y esclarecidos con referencia a, las realizaciones descritas a continuación en la presente memoria.

La invención propuesta proporciona, entre otras, las siguientes ventajas:

- Es una solución ligera para controlar la saturación en un servidor, o grupo de servidores, que asegura una respuesta rápida a los clientes. Esto significa que, aunque es un intermediario que figura entre un cliente y un servidor, el procesamiento no es complejo.

- Reduce la cantidad de nodos requeridos en un grupo de servidores para poder dar soporte a una carga de trabajo anormal, recibida

esporádicamente. Significa que el grupo de servidores puede estar calculado para una carga de trabajo normal, pero no para una carga de trabajo excepcional.

• El umbral de tiempo absoluto asegura un tiempo máximo para ejecutar (procesar) una solicitud en el sector servidor. Esto significa que el cliente obtendrá una respuesta válida y completa en un tiempo máximo.

• La detección temprana de la congestión se hace sobre la capa de aplicación del protocolo de red. Significa que las solicitudes desde el cliente al servidor no son rechazadas en la capa del TCP; por lo tanto, el sistema remoto no percibe ninguna situación excepcional.

• La detección temprana de la congestión es un mecanismo adaptativo automático. Su objetivo es mantener el servidor en la carga máxima (cerca de la saturación) pero sin sobrecargarlo. Adicionalmente, la detección temprana de la congestión asegura que al menos una tasa mínima de solicitudes va al servidor para monitorizar su estado continuamente.

• El servidor queda totalmente funcional incluso cuando se aproxima a la capacidad máxima. En otras palabras, puede mantener el servidor en la carga alta sin sobrecargarlo.

• La respuesta genérica es actualizada de tanto en tanto para evitar respuestas obsoletas, y es almacenada en una memoria de acceso rápido para asegurar que la respuesta genérica sea devuelta inmediatamente incluso cuando se recibe una alta carga de trabajo.

DESCRIPCIÓN DE LOS DIBUJOS

Para completar la descripción que se está haciendo, y con el objeto de asistir para una mejor comprensión de las características de la invención, de acuerdo a un ejemplo preferido de realizaciones prácticas de la misma, y acompañando a dicha descripción como parte integral de la misma, hay un conjunto de dibujos en los que, a modo de ilustración y no de manera restrictiva, ha sido representado lo siguiente:

Figura 1: Muestra un diagrama de bloques esquemático del sistema propuesto, de acuerdo a una realización de la invención.

Figura 2: Muestra un diagrama de flujo del control de congestión (sobrecarga) de servicios, de acuerdo a una realización de la invención.

Figura 3: Muestra un diagrama de flujo de la verificación de la sobrecarga de un Sistema de Procesamiento de Datos, de acuerdo a una realización de la invención.

Figura 4: Muestra un diagrama de flujo de la monitorización del tiempo de ejecución (procesamiento), de acuerdo a una realización de la invención.

DESCRIPCIÓN DETALLADA DE LA INVENCION

La presente invención puede ser realizada en otros dispositivos, sistemas y / o procedimientos específicos. Las realizaciones descritas han de ser consideradas, en todo respecto, como solamente ilustrativas, y no restrictivas. En particular, el ámbito de la invención está indicado por las reivindicaciones adjuntas, antes que por la descripción y las figuras en la presente memoria. Todos los cambios que queden dentro del significado y gama de equivalencia de las reivindicaciones han de ser acogidas dentro de su ámbito. En la siguiente descripción, con fines de explicación, se

estipulan numerosos detalles específicos, a fin de brindar una comprensión exhaustiva de la presente invención. Sin embargo, será evidente a un experto en la técnica que la presente invención puede ser puesta en práctica sin estos detalles específicos. En otros casos, estructuras y dispositivos bien conocidos se muestran en forma de diagrama de bloques, a fin de evitar oscurecer innecesariamente la presente invención.

La presente invención propone un procedimiento, un sistema y un dispositivo capaces de gestionar la congestión (sobrecarga de uno o más servidores) en un servicio de red (un servicio de procesamiento de datos, un servicio de la Red o cualquier otro tipo de servicio de red). En este contexto, un servicio de red es un servicio proporcionado en la capa de aplicación de red. Se implementa a menudo usando una arquitectura de cliente-servidor, basada en protocolos de la capa de aplicación de red (también pueden ser usadas arquitecturas de igual a igual). Los servicios de red pueden proporcionar almacenamiento de datos, manipulación, personalización de la información, presentación, comunicación u otras muchas capacidades a nivel de aplicación. Cada servicio es proporcionado usualmente por un servidor (en términos generales, un grupo de servidores ejecutándose en uno o más ordenadores), al que se accede mediante una red de telecomunicaciones desde dispositivos clientes. El servicio de red puede ser proporcionado por una aplicación o un conjunto coordinado de aplicaciones ejecutándose en los servidores, y comunicándose mensajes de alto nivel entre ellos, y con las aplicaciones clientes, a través de la red.

La Figura 1 muestra un diagrama de bloques del sistema propuesto, de acuerdo a una realización de la invención. El sistema propuesto por la presente invención presenta un sistema o nodo intermedio (llamado Gestor de Congestión o de Sobrecarga 200) conectado con uno o más servidores que proporcionan el servicio de red. En términos generales, estos servidores se encargarán de realizar el procesamiento de datos necesario para responder a la solicitud de servicio (y generar la respuesta

de servicio a enviar al cliente que solicita el servicio), y es por eso que este servidor, o estos servidores (que proporcionan el servicio), es (son) llamado(s) Servidor(es) de Procesamiento de Datos o, más en general, Grupo de Servidores de Procesamiento de Datos (DPS, 100), que es el grupo de uno o más servidores que proporcionan el servicio. El Gestor puede estar conectado con un único grupo de servidores, o con un conjunto de distintos grupos de servidores. Cada grupo puede estar compuesto por un número indefinido de nodos (N nodos (110) en la figura 1, donde $N \geq 1$), que se encargarán del proceso de procesamiento (cálculo) final y que serán responsables de proporcionar las respuestas a nivel de aplicación que espera el cliente al cual se proporciona el servicio (el sistema remoto en la figura 1).

En una realización, el resultado del servicio de red es un mensaje de respuesta enviado por el proveedor del servicio al cliente del servicio (la respuesta de servicio). Por ejemplo, el servicio puede ser un servicio de información (p. ej., un servicio de recomendación, un servicio de noticias...) en el cual el cliente solicita una información, el servidor (o servidores) de procesamiento de datos procesa(n) la solicitud y obtiene, selecciona y prepara la información (de acuerdo al contenido de la solicitud y / o al perfil del cliente), y el resultado del servicio es un mensaje de respuesta enviado al cliente con la información solicitada. Pero esto es solamente un ejemplo; en términos generales, la presente invención puede ser aplicada a cualquier tipo de servicios de red que den una respuesta al cliente que solicita el servicio.

El gestor de sobrecarga también está conectado con un dispositivo cliente, llamado sistema remoto (300), que será el dispositivo (terminal o, en términos generales, un nodo) que efectúa la solicitud (solicitud a nivel de aplicación) original (de servicio / información), y al cual se proporciona el servicio. Según el servicio proporcionado, el sistema remoto puede ser, por ejemplo, un terminal de comunicación abonado al servicio, u otro nodo de red al cual se proporciona el servicio.

El sistema remoto estará conectado con el gestor de sobrecarga mediante una red de telecomunicaciones. El gestor de sobrecarga puede estar co-situado con el servidor, para un rápido acceso a los nodos del grupo, o puede estar situado en una ubicación distante del servidor, y estar conectado con el servidor mediante la misma red de telecomunicaciones, o una distinta. Usualmente, en este último caso (el gestor de sobrecarga no está co-situado con el servidor), la velocidad y el retardo del enlace entre el gestor de sobrecarga y el servidor son suficientes para mantener una veloz comunicación entre ellos. La red de telecomunicaciones puede ser una red cableada, una red inalámbrica, una red de comunicaciones móviles (LTE, 3G, 4G, GSM...) o cualquier otro tipo de red de telecomunicación. El servicio proporcionado puede pertenecer al mismo operador propietario de la red de telecomunicación, o a otro operador. El gestor de sobrecarga puede estar situado en un nodo de red de la red de telecomunicaciones, o en un nodo de red perteneciente a otra red.

El sistema remoto (dispositivo cliente) puede ser un teléfono móvil, un ordenador portátil, un sistema de ordenador, un teléfono inteligente, un televisor inteligente, un televisor de IPTV (Televisión del Protocolo de Internet), una tableta o cualquier otro dispositivo electrónico capaz de comunicarse a través de una red de telecomunicaciones.

El gestor de congestión o sobrecarga (200) representa el núcleo del sistema. Puede usarse más de un gestor de sobrecarga (por ejemplo, uno para cada servicio proporcionado por la red, o uno para cada grupo de servicios, o bien, si el número de clientes de un servicio es muy alto, puede haber varios gestores para el mismo servicio, uno para cada grupo de clientes). Cada gestor de sobrecarga (200) realiza distintas funciones; para mayor simplicidad y claridad, el gestor de sobrecarga puede ser visto como un nodo con distintos elementos o módulos funcionales, cada uno a cargo de al menos una función (en una realización, el gestor incluirá un procesador configurado para realizar algunas de, o todas, las funciones).

En la realización mostrada en la figura 1, el gestor de sobrecarga está compuesto por los siguientes módulos (esto es solamente un ejemplo, y algunos de los módulos son optativos y, en consecuencia, no esenciales para la invención):

5

10

Gestor de Respuestas (RM) (210): Encargado de recibir la solicitud original desde el cliente (sistema remoto). Si se activa el mecanismo de contingencia (por parte del Gestor de Congestión Temprana o por el Monitor Temporal), la respuesta genérica será entregada al sistema remoto. Si no es así, la solicitud será enviada al Grupo de Servidores de Procesamiento de Datos, DPS; el DPS generará una respuesta adecuada, específica para dicha solicitud, y el RM enviará el mensaje de respuesta al cliente.

15

20

Monitor Temporal (TM) (220): Se ocupa de asegurar un máximo umbral en el tiempo de respuesta (lapso absoluto de umbral); monitoriza el tiempo global de procesamiento de la solicitud (el tiempo consumido por el DPS en el procesamiento de la solicitud y la generación de una respuesta, llamado también lapso de ejecución de solicitud, tiempo de ejecución o tiempo de respuesta) y, si detecta que la operación está llegando a un umbral (es decir, ha transcurrido un cierto umbral cronológico y la respuesta no ha sido todavía generada por el DPS), aborta el procesamiento y activa el mecanismo de contingencia.

30

35

Gestor de Congestión Temprana (ECM) (230): Aplica la Detección de Congestión Temprana monitorizando los tiempos de respuesta en curso y, toda vez que el tiempo de respuesta en curso (por ejemplo, un promedio de los tiempos de respuesta para las últimas M solicitudes, donde M es un parámetro de diseño) crece con un patrón arriesgado, activa el mecanismo de contingencia, que comienza a no remitir solicitudes al servidor (a una tasa dinámica) para mantener los servidores de procesamiento de datos un poco por debajo de la congestión.

Cargador Genérico (GL) (240): Periódicamente (el periodo puede depender de la aplicación y de los datos) solicita al servidor de procesamiento de datos una respuesta genérica, o un conjunto de respuestas genéricas (o, en términos generales, datos genéricos para componer una respuesta genérica), válidos como mecanismo de repliegue (en situaciones de contingencia) para la gama de posibles solicitudes que el sistema pueda recibir.

Agente de Procesamiento de Datos (DPA) (250): En algunos servicios, o para algunas solicitudes específicas, un mensaje individual de solicitud de un sistema remoto es descompuesto (en este módulo) en un cierto número de operaciones individuales a enviar a los servidores de procesamiento de datos. La respuesta final (también llamada global) será creada componiendo entre sí las respuestas individuales a las distintas operaciones individuales en las cuales ha sido descompuesta la solicitud. También es posible que un mensaje de solicitud de sistema remoto corresponda a un único mensaje de solicitud para los servidores de procesamiento de datos (por lo que no hay ningún proceso de descomposición / composición).

Base de Datos de Respuestas Genéricas (GRDB) (260): Base de datos encargada de almacenar respuestas genéricas. Estas respuestas son calculadas previamente por el Cargador Genérico.

Base de Datos de Contingencias (CDB) (270): Base de datos encargada de almacenar tiempos de ejecución y parámetros de configuración requeridos por el Gestor de Congestión Temprana para calcular si el mecanismo de contingencia debe ser activado.

Gestor de Memoria Caché (CM) (280): Su propósito es mantener la respuesta genérica en una memoria local y rápida (memoria caché) para mejorar las prestaciones (el acceso a la respuesta genérica es más rápido) cuando se activa el mecanismo de contingencia.

Ahora se va a explicar el funcionamiento detallado de la presente invención (es decir, las distintas etapas realizadas o, en otras palabras, el flujo del procesamiento de datos), de acuerdo a distintas realizaciones de la invención.

5

Cuando el gestor de sobrecarga detecta una situación de congestión en el servidor, o los servidores, de servicio (por ejemplo, en los servidores de procesamiento de datos), se activa un mecanismo de contingencia para generar una respuesta al sistema remoto, usando datos genéricos, a fin de devolver una respuesta rápida. El flujo completo del procesamiento de datos puede ser dividido en dos procesos distintos y desacoplados.

10

15

Carga previa de datos genéricos: La condición previa para hacer que el sistema funcione consiste en cargar datos genéricos, que serán usados para generar una respuesta cuando entre en acción el mecanismo de contingencia

20

Control de congestión de servidores: Cuando un nuevo mensaje de solicitud llega al gestor de sobrecarga desde el sistema remoto, debe ser gestionado para controlar la sobrecarga en el DPS, y para activar o no el mecanismo de contingencia (es decir, para determinar si está ocurriendo una situación de congestión, o si está por ocurrir). Este proceso se divide en tres sub-procesos:

25

30

Gestión de congestión temprana: Se lleva a cabo una verificación de congestión temprana a fin de detectar si los servidores de procesamiento de datos están sobrecargados (o están cerca de estarlo). Cuando se detecta una situación de congestión, algunas solicitudes (no todas) son respondidas usando el mecanismo de contingencia.

35

Monitorización del tiempo de ejecución de solicitud: Si la solicitud del sistema remoto no es respondida con datos genéricos, es enviada al DPS y el tiempo

consumido es controlado para asegurar que no supere un umbral específico. Cuando esto ocurre (se consume el tiempo de espera), la solicitud es respondida usando el mecanismo de contingencia.

Entrega de respuesta genérica: Cuando se detecta una situación de congestión temprana, o se consume el tiempo de espera de ejecución de una solicitud, el gestor de sobrecarga activa el mecanismo de contingencia, componiendo (si no está ya compuesta) y entregando la respuesta al sistema remoto, en base a los datos genéricos.

Carga previa de datos genéricos

El cargador genérico (GL) es el responsable de obtener datos genéricos del DPS y de almacenarlos en la base de datos de respuestas genéricas (GRDB). Dichos datos genéricos serán usados para construir el mensaje de respuesta genérica (que será enviado al cliente que solicita el servicio cuando sea activado el mecanismo de contingencia). A fin de tener una respuesta genérica actualizada, el cargador de datos genéricos es ejecutado periódicamente (el intervalo de ejecución depende de la aplicación y del servicio (los datos)). Cada vez que es ejecutado, se efectúa una solicitud a un servidor de procesamiento de datos, para obtener resultados genéricos, pero válidos, a devolver cuando el mecanismo de contingencia entra en acción. Los datos devueltos deberían ser lo bastante genéricos para satisfacer las expectativas de la mayoría de los clientes del sistema, pero no serán generados teniendo en cuenta un perfil de cliente o solicitud específicos. Por ejemplo, si es un sistema de recomendaciones televisivas, la respuesta genérica puede ser una guía general de los programas de televisión disponibles, o una lista de los programas de televisión más vistos, que no esté personalizada para un cliente específico, para que sea válida para todos los clientes (usuarios abonados). Los datos recuperados desde el DPS deben ser almacenados en la GRDB en el formato adecuado para poder

generar un mensaje de respuesta genérica tan rápidamente como sea posible.

Control de congestión de servicios

5

La Figura 2 muestra un diagrama de actividad de flujo del proceso para controlar la congestión de uno o más servidores en una realización de la invención. Este proceso está dividido en tres sub-procesos:

10

1) Gestión de congestión temprana: El objetivo principal de este sub-proceso es detectar congestiones (es decir, sobrecarga) en los servidores de procesamiento de datos. Este sub-proceso es llevado a cabo por el Gestor de Congestión Temprana (ECM) usando los datos almacenados en la Base de Datos de Contingencias (CDB). Es responsable de las siguientes tareas:

15

20

Verificar la sobrecarga de los servidores de procesamiento de datos: Después de que el RM recibe (211) una solicitud del sistema remoto (un mensaje de solicitud de servicio), el ECM verifica si el DPS está sobrecargado (212) y si debe ser activado (213) el mecanismo de contingencia. Estas decisiones pueden estar basadas en los tiempos de ejecución registrados (almacenados en la CDB) para las más recientes solicitudes procesadas, y en algunos parámetros de configuración. Un procedimiento específico debería ser usado para tomar esta decisión. Por ejemplo, si el tiempo de respuesta promediado para las últimas M solicitudes es más que un cierto umbral, se determina que el DPS está sobrecargado y, en consecuencia, el mecanismo de contingencia debe ser activado. Esto es solamente un ejemplo; otros ejemplos serán explicados más adelante. En términos generales, cualquier procedimiento de decisión conocido puede ser usado para tomar esta decisión. Si el mecanismo de contingencia no es activado, la solicitud es enviada al DPS para su procesamiento (214) por el DPA.

25

30

35

Estimar el tiempo de ejecución de solicitud (también llamado tiempo de procesamiento): Cuando el mecanismo de contingencia entra en acción, no es posible saber exactamente cuánto tiempo habría sido consumido por el DPS procesando la solicitud para preparar la respuesta, por lo que debe ser estimado (215). Más adelante se explicarán algunos ejemplos de cómo estimar este tiempo de ejecución.

Registrar el tiempo de ejecución de solicitud: Cada vez que una solicitud del sistema remoto es procesada por el gestor de sobrecarga, el tiempo de ejecución, estimado o real, de la solicitud es almacenado (216, 218) en la CDB. Estas mediciones se usan para determinar si el DPS está congestionado (sobrecargado) en futuras solicitudes de servicio.

2) Monitorización del tiempo de ejecución (217): Si el mecanismo de contingencia no ha sido activado, la solicitud es enviada al DPS para su procesamiento (214). El objetivo principal de este sub-proceso es controlar el tiempo consumido por una solicitud al DPS. Si este tiempo supera un límite temporal dado (es decir, un cierto límite temporal transcurrido desde que la solicitud ha sido enviada al DPS, sin haber recibido el gestor de sobrecarga una respuesta desde los servidores), la solicitud al DPS es abortada y se aplica el mecanismo de contingencia, realizando las acciones de estimar (215) y almacenar (216) el tiempo de ejecución, y obtener (221) y entregar (222) la respuesta genérica. Por ello, se evitan largas esperas para generar una respuesta a las solicitudes del sistema remoto. Este sub-proceso es llevado a cabo por el Monitor Temporal (TM), en colaboración con el Agente de Procesamiento de Datos (DPA) y el Gestor de Respuestas (RM).

3) Entrega de respuesta genérica: Al final del procesamiento de la solicitud del sistema remoto, debe ser generada una respuesta final. Si el proceso entero es ejecutado con éxito (es decir, el mecanismo de contingencia no es activado), el tiempo de ejecución real es almacenado (218) (en la CDB) y la respuesta generada por el DPS es encaminada, mediante el DPS,

al RM, y la respuesta de servicio (un mensaje de respuesta de servicio) es enviada al sistema remoto (219) mediante la red de telecomunicaciones.

5 Pero, si el proceso entero no es ejecutado con éxito (es decir, el mecanismo de contingencia es activado debido a la detección de congestión temprana, o debido a que el tiempo de ejecución de la solicitud supera el umbral), el gestor de respuestas debe ejecutar este mecanismo para devolver una respuesta genérica.
10 Debe señalarse que, si se detecta congestión temprana, el mecanismo de contingencia no es activado para todas las solicitudes entrantes, sino por lo común solamente para algunas de ellas (la tasa de las solicitudes entrantes para las cuales se aplica el mecanismo de contingencia puede ser actualizada para cada solicitud entrante, y dependerá de un parámetro de configuración). La entrega de la respuesta genérica puede comprender las siguientes etapas:

20 El RM llama al Administrador de Memoria Caché (CM) para obtener la respuesta genérica.

Si el CM contiene la respuesta genérica y estos datos son válidos, el CM devuelve la respuesta genérica al RM.

2 Si el CM no contiene la respuesta genérica, o si los datos en la memoria caché no son válidos, el CM obtiene la respuesta genérica de la GRDB, la almacena en memoria caché y devuelve la respuesta genérica al RM.

30 Una vez que el RM obtiene la respuesta genérica (221), es entregada al sistema remoto (222) mediante la red de telecomunicaciones.

35 El CM se usa para mejorar las prestaciones cuando se ejecuta el mecanismo de contingencia. Debido a que el mecanismo de contingencia se activa cuando el sistema está sobrecargado, es necesario minimizar los recursos usados para generar la respuesta genérica. Por lo tanto, el CM almacena los datos en memoria caché en un almacén rápido, local para el RM, para

evitar efectuar conexiones de red con la GRDB, que por lo común está alojada en un nodo dedicado dentro del Gestor de Sobrecarga, que puede no ser local para el RM (por ejemplo, la GRDB puede estar alojada en otro nodo dentro del mismo conjunto de nodos). Sin embargo, los datos en memoria caché no siempre serán válidos. Después de un periodo de tiempo, o cuando los datos en memoria caché hayan sido usados un cierto número de veces, son invalidados y deben ser regenerados obteniendo nuevamente la respuesta genérica desde la GRDB.

Pero la presencia del gestor de memoria caché no es obligatoria. En una realización alternativa, no hay ningún gestor de memoria caché, por lo que la respuesta genérica es siempre solicitada y obtenida directamente por el RM desde la GRDB.

Ahora serán explicadas realizaciones que materializan algunos de los procedimientos específicos mencionados en el control del proceso de congestión de servicios, previamente descrito. Son aplicables a servicios de red cualesquiera.

Estimación del tiempo de ejecución esperado (también llamado tiempo de respuesta) para los servidores de procesamiento de datos: Como se ha explicado anteriormente, la presente invención necesita estimar el tiempo de ejecución (tiempo consumido en el procesamiento de la solicitud y en la generación de una respuesta) para el servidor, o los servidores, de procesamiento de datos cuando se activa el mecanismo de contingencia (de hecho, como el DPS no da una respuesta a la solicitud, puede decirse que lo que se estima es un tiempo de ejecución hipotético, es decir, una estimación de cuánto tiempo habría sido consumido por el DPS para procesar la solicitud y preparar la respuesta). Esta estimación se hace a fin de activar las distintas fases del mecanismo de detección de congestión para solicitudes futuras. La medida más sencilla es usar el tiempo de la última ejecución (es decir, el tiempo de ejecución de la solicitud anterior). Sin embargo, esto no es fiable (los fallos y los fenómenos transitorios pueden

17
producir sobresaltos ocasionales en el tiempo de procesamiento, que no son relevantes en términos de congestión) y, por lo tanto, las realizaciones tenderán a realizar alguna clase de promedio sobre las últimas R solicitudes ejecutadas, a fin de allanar irregularidades a corto plazo. Algunas realizaciones también pueden escoger variantes adicionales:

Usar una mediana en lugar del valor medio, como forma de eliminar del cálculo elementos marginales sin interés y retener el verdadero estado de congestión subyacente.

En caso de que una ejecución sea abortada debido a la superación del umbral en el módulo de Control Temporal, el tiempo estimado (añadido a la ventana promediadora) sería el tiempo hasta el momento, más un incremento aleatorio que depende del número de solicitudes recientemente abortadas (para estimar el eventual tiempo final, incluyendo el efecto de los servidores congestionados).

Los anteriores son solamente algunos ejemplos de cómo estimar el tiempo de ejecución; podrían usarse otras variantes y procedimientos de estimación conocidos.

Comprobación de sobrecarga de servidores de procesamiento de datos (para la detección de congestión temprana): Un procedimiento específico debería ser aplicado para decidir si los DPS (el Servidor de Procesamiento de Datos, o el Grupo de Servidores) están sobrecargados y si el mecanismo de contingencia debe ser activado (una respuesta genérica debe ser enviada al cliente solicitante). Aquí, como ejemplo, un procedimiento que podría ser usado con este propósito va a ser explicado de acuerdo a una realización de la invención. En esta realización, los principales puntos de datos y parámetros requeridos por el procedimiento para tomar la decisión son:

- El tiempo medio de ejecución actual de las últimas solicitudes.

- El umbral para el tiempo medio de ejecución a partir del cual se considera que los DPS están sobrecargados.

- La probabilidad / tasa actual de no remisión de la solicitud al DPS (es decir, la tasa actual para la activación del mecanismo de contingencia y el uso de datos genéricos para la respuesta de la solicitud).

- La tasa máxima de solicitudes respondidas con datos genéricos (no remitidas a los DPS), para asegurar que al menos algunas solicitudes sean enviadas a los DPS a fin de medir si los DPS están aún sobrecargados o no.

Esta detección / gestión de congestión temprana puede usar, por ejemplo, los siguientes parámetros de configuración para operar (por lo común, estos parámetros son debidamente seleccionados y personalizados para cada aplicación específica, por ejemplo, usando pruebas de prestaciones, a fin de optimizar el funcionamiento y los resultados del sistema):

- `HABILITAR_GESTOR_SOBRECARGA` (booleano): Indicador para determinar si la detección de congestión temprana está habilitada o inhabilitada. Si está inhabilitada, todas las solicitudes son enviadas al DPS.

- `TAMAÑO_COLA_GESTOR_SOBRECARGA` (entero): Número de valores del tiempo de ejecución de solicitudes, usados para calcular el tiempo medio de ejecución del DPS.

- `UMBRAL_GESTOR_SOBRECARGA` (entero): Tiempo (p. ej., en milisegundos) a partir del cual, si el tiempo medio de ejecución de los DPS lo supera, se considera a los DPS como sobrecargados.

- `PASO_GESTOR_SOBRECARGA` (entero): Tiempo (p. ej., en milisegundos) para el paso de aumento de la tasa de solicitudes respondidas con datos genéricos.

• PASO_TASA_GESTOR_SOBRECARGA (número en formato de coma flotante): Tasa de solicitudes a responder con datos genéricos (o, en otras palabras, la tasa de solicitudes para las cuales va a ser activado el mecanismo de contingencia) por cada paso.

• MÁXIMA_TASA_GESTOR_SOBRECARGA (número en formato de coma flotante): Tasa máxima de solicitudes a responder con datos genéricos. Debe ser menor que uno para asegurar que algunas solicitudes sean enviadas al DPS cuando se considera como sobrecargado, a fin de medir el tiempo de ejecución y saber si los DPS están todavía sobrecargados o no.

Además, pueden ser usadas dos variables:

• cola_tiempos: Es una cola LIFO (Último en Entrar, Primero en Salir) para mantener el tiempo de ejecución de las últimas solicitudes. Cada vez que se adjunta a la cola una medición del tiempo de ejecución de una solicitud (una medición real si la última solicitud ha sido procesada con éxito por el DPS, o una estimación), otro elemento es eliminado de la cola desde el sector opuesto, para mantener un número máximo de mediciones de tiempo de ejecución de solicitudes en la cola. El tamaño de la cola está fijado en el valor del parámetro TAMAÑO_COLA_GESTOR_SOBRECARGA.

• tiempo_medio: Es el tiempo medio de ejecución de solicitud para las últimas solicitudes. Se calcula usando todas las mediciones en la cola_tiempos. Como medida de exactitud, en una realización, si la cola está llena hasta menos, por ejemplo, que la mitad de su tamaño (pero pueden usarse otros valores), no se calcula el tiempo medio y el control de sobrecarga (congestión) se considera como inhabilitado. Esto es solamente un ejemplo; realizaciones

alternativas podrían usar la mediana u otra estadística en lugar del promedio matemático.

Cada vez que llega una nueva solicitud al gestor de sobrecarga desde el sistema (cliente) remoto, el DPS actualiza el tiempo_medio y comprueba si este valor supera un umbral configurado (UMBRAL_GESTOR_SOBRECARGA). Si esto ocurre, significa que el DPS está considerado como sobrecargado (congestionado). Entonces, el gestor de sobrecarga debe decidir si la solicitud actual no se envía al DPS y se envía una respuesta genérica al sistema remoto (es decir, se activa el mecanismo de contingencia). Esta decisión está basada en una probabilidad. Esta probabilidad puede ser calculada, por ejemplo, según en cuántos pasos esté sobrepasado el umbral. Cada paso es una cantidad de tiempo (milisegundos) y cada paso tiene asociada una probabilidad.

En una realización ejemplar, están seleccionados los siguientes valores de parámetros de configuración (esto es solamente un ejemplo; pueden ser usados muchos otros valores):

```
UMBRAL_GESTOR_SOBRECARGA = 300 ms
PASO_GESTOR_SOBRECARGA = 50 ms
PASO_TASA_GESTOR_SOBRECARGA = 0,2 (20%)
```

Si el tiempo medio es, por ejemplo, de 360 ms, la probabilidad de responder con datos genéricos (de activar el mecanismo de contingencia) a la solicitud actual puede ser calculada como:

```
núm_pasos = int((360ms - 300ms + 50ms) / 50ms)
           = 2
probabilidad (tasa actual) = núm_pasos * 0,2 =
           0,4 (40%)
```

Esta probabilidad no puede ser mayor que el valor de MÁXIMA_TASA_GESTOR_SOBRECARGA; y, en cualquier caso, el límite máximo para este valor es, por ejemplo, 0,9. La probabilidad de activar el mecanismo de contingencia para la solicitud

actual en este caso será de 0,4. La probabilidad calculada es comparada, por ejemplo, con un número aleatorio en la gama de [0, 1) para decidir finalmente si la solicitud actual no se envía al DPS y se activa el mecanismo de contingencia para producir una respuesta genérica para las solicitudes actuales. En otra realización ejemplar, pueden ser seleccionados los siguientes valores de parámetros de configuración:

```

UMBRAL_GESTOR_SOBRECARGA = 300 ms
PASO_GESTOR_SOBRECARGA = 30 ms
PASO_TASA_GESTOR_SOBRECARGA = 0,3 (30%)

```

Si el tiempo medio es el mismo que antes, de 360 ms, la probabilidad de activar el mecanismo de contingencia para la solicitud actual es la siguiente:

```

núm_pasos = int((360ms - 300ms + 30ms) / 30ms)
= 3
probabilidad = núm_pasos * 0,3 = 0,9 (90%)

```

La probabilidad de activar el mecanismo de contingencia para una solicitud entrante en este caso será de 0,9. Es decir, con el mismo umbral y el mismo tiempo medio que en el ejemplo anterior, la probabilidad de activar el mecanismo de contingencia es mucho mayor. Esto es porque algunos de los parámetros de configuración (PASO_GESTOR_SOBRECARGA, PASO_TASA_GESTOR_SOBRECARGA) han sido seleccionados de modo que el sistema sea mucho más sensible a cualquier aumento en el tiempo medio de ejecución (hay más probabilidades de activar el mecanismo de contingencia). En otras palabras, según la situación (aplicación) específica a la cual se aplica la presente invención, puede usarse una detección de congestión más o menos sensible (estricta), según el valor seleccionado para los parámetros de configuración.

Para resumir y clarificar el proceso explicado en lo que antecede, se usará la figura 3. La figura 3 muestra un diagrama de actividad del flujo del proceso para comprobar la sobrecarga

del DPS (para la detección de congestión temprana), de acuerdo a la realización explicada anteriormente de la invención.

Cuando una nueva solicitud (un mensaje de solicitud) llega al gestor de sobrecarga desde el sistema (cliente) remoto, la primera etapa es comprobar (311) si está activada la detección de congestión temprana (por ejemplo, comprobando que el parámetro `HABILITAR_GESTOR_SOBRECARGA` es Verdadero). Si no está activada, todas las solicitudes son enviadas al DPS para su procesamiento (316) y el procedimiento se acaba (319). Si está activada, entonces se comprueba (312) si la cola de tiempos (que almacena el tiempo de ejecución de las últimas solicitudes) tiene valores suficientes para dar un promedio aceptable (en este caso, la mitad de los valores). Si no es así, la congestión temprana se considera inhabilitada, y todas las solicitudes son enviadas al DPS para su procesamiento (316). Si tiene valores suficientes, se calcula (313) el tiempo medio de ejecución actual de las últimas solicitudes (en este caso, los valores almacenados en la cola). Luego se comprueba (314) si el tiempo medio de ejecución actual supera un umbral configurado (p. ej., `UMBRAL_GESTOR_SOBRECARGA`). Si no supera el umbral, entonces la solicitud es enviada al DPS para su procesamiento (316). Si el umbral es superado, entonces se calcula (315) la tasa actual de las solicitudes a responder con una respuesta genérica o, en otras palabras, se calcula la tasa a aplicar a la solicitud actual. En este ejemplo específico, para calcular dicha tasa, se realizan las siguientes operaciones:

```
núm_pasos = int((tiempo medio - UMBRAL_GESTOR_SOBRECARGA +
PASO_GESTOR_SOBRECARGA) / PASO_GESTOR_SOBRECARGA)
probabilidad = núm_pasos * PASO_TASA_GESTOR_SOBRECARGA
tasa actual=min (probabilidad, MÁXIMA_TASA_GESTOR_SOBRECARGA)
```

Finalmente, se compara la tasa actual calculada, por ejemplo, con un número aleatorio (317) para decidir finalmente si la solicitud actual se envía al DPS (316) o si se activa (318) el

mecanismo de contingencia para producir y enviar al cliente una respuesta genérica para las solicitudes actuales.

Los anteriores son solamente algunos ejemplos de cómo comprobar la sobrecarga del DPS; podrían usarse otras variantes y procedimientos conocidos.

Monitorización del tiempo de ejecución: Debería usarse un procedimiento específico para controlar (monitorizar) el tiempo consumido por el DPS al procesar una solicitud (tiempo de ejecución de solicitud o solamente tiempo de ejecución). Como se ha explicado anteriormente, si se supera un tiempo de espera dado, debe ser activado el mecanismo de contingencia. Aquí, como ejemplo, va a explicarse un procedimiento que podría ser usado con este fin (monitorizar el tiempo de ejecución), de acuerdo a una realización de la invención. Este procedimiento se aplica generalmente cuando cada solicitud desde el sistema remoto debe ser descompuesta en varias (sub) solicitudes enviadas al DPS para obtener varios trozos de datos (uno para cada sub-solicitud) que el Agente de Procesamiento de Datos (DPA) debe combinar para generar la respuesta final. Si no es este el caso (la solicitud no es descompuesta en varias (sub) solicitudes), puede valer el mismo procedimiento, pero con la condición específica de que haya solamente una única sub-solicitud.

Este procedimiento puede usar, por ejemplo, el siguiente parámetro de configuración para operar (por lo común, este parámetro es debidamente seleccionado y personalizado para cada aplicación específica, por ejemplo, usando pruebas de prestaciones, a fin de optimizar el funcionamiento y los resultados del sistema):

- `LÍMITE_TEMPORAL_SOLICITUD` (entero): Se usa para fijar el tiempo máximo (p. ej., en milisegundos) para procesar completamente una solicitud desde el sistema remoto y para generar una respuesta final (es decir, el máximo tiempo de ejecución). Si este límite temporal es

superado mientras el procesamiento del DPS está en marcha, se activa el mecanismo de contingencia y se crea la respuesta final usando los datos genéricos. Si el parámetro `LÍMITE_TEMPORAL_SOLICITUD` se fija en cero o un valor nulo, no se aplica ningún límite temporal al procesar la solicitud para el DPS.

Además, puede ser usada una variable:

- `límite_temporal_pendiente` (entero): Se usa para controlar el tiempo restante para ejecutar las (sub) solicitudes pendientes para el DPS. Se inicializa con el valor de `LÍMITE_TEMPORAL_SOLICITUD` y, cada vez que es ejecutada una solicitud individual para el DPS, se mide el tiempo consumido y se reduce el `límite_temporal_pendiente` en este valor.

Para explicar este procedimiento va a usarse la figura 4. La figura 4 muestra un diagrama de actividad del flujo del procedimiento para monitorizar el tiempo de ejecución de solicitudes, de acuerdo a una realización de la invención.

Como se ha explicado antes, el DPA (Agente de Procesamiento de Datos) descompone una solicitud procedente del sistema remoto en un cierto número de operaciones individuales (llamadas subsolicitudes o solicitudes individuales). Realiza un bucle para enviar solicitudes individuales al DPS, para obtener trozos de datos para componer la respuesta final a la solicitud original. A fin de monitorizar el tiempo de ejecución, inicialmente, el `límite_temporal_pendiente` se fija en el valor de `LÍMITE_TEMPORAL_SOLICITUD` (411). Este es el tiempo total disponible para ejecutar todas las iteraciones del bucle (es decir, para procesar todas las solicitudes individuales a fin de generar la respuesta final). El `límite_temporal_pendiente` es actualizado al final de cada iteración del bucle, restando el tiempo consumido por la solicitud individual para el DPS. Por lo tanto, para cada iteración se conoce el tiempo restante para ejecutar todo el trabajo pendiente.

Primero, se comprueba (412) si todas las solicitudes individuales han sido procesadas o no (en otras palabras, si debe comenzar o no otro bucle). Si todas las solicitudes individuales han sido procesadas (no hay ninguna solicitud pendiente), se compone (418) la respuesta final con todas las respuestas parciales (las respuestas a cada solicitud individual) y el procedimiento termina (420). Si hay solicitudes individuales pendientes de ser procesadas, entonces comienza otro bucle para procesar otra solicitud individual. Al comienzo de cada iteración del bucle, se comprueba (413, 414) si el tiempo restante para ejecutar todo el trabajo pendiente (límite temporal pendiente) es menor o igual a cero. Si es menor o igual a cero (el límite temporal ha sido superado), entonces se aborta el proceso normal en el DPA (no se envían más solicitudes individuales al DPS) y la respuesta final es generada por el mecanismo de contingencia (419), es decir, la respuesta genérica es enviada al sistema remoto.

Si no es así (el tiempo pendiente es más que 0 y, en consecuencia, el límite temporal no ha sido superado), el DPA envía la siguiente solicitud individual al DPS para obtener un trozo de datos (una respuesta parcial) desde el DPS (415). El límite_temporal_pendiente se usa como el máximo tiempo disponible para la solicitud individual actual. Si este tiempo es consumido antes de que sea recibida la respuesta a la solicitud individual del DPS, se aborta el proceso normal en el DPA y la respuesta final es generada por el mecanismo de contingencia (419). Si no es así, se calcula (416) el tiempo consumido por el DPS para procesar (para generar la respuesta parcial para) la solicitud individual actual, se actualiza (417) el límite temporal pendiente, restando el tiempo consumido calculado al valor anterior del tiempo pendiente, y se comprueba (412) si todas las solicitudes individuales han sido procesadas o no (es decir, se comprueba si debe comenzar o no otro bucle).

Como se ha mencionado en la sección de Antecedentes de la Invención, hay muchos procedimientos establecidos para el control de la congestión en los servicios en red. Sin embargo, la mayoría de ellos actúan en un nivel muy bajo (p. ej., en la capa de red), descartando paquetes toda vez que detectan una situación de sobrecarga. Algunos de ellos contienen criterios específicos para la redistribución equitativa de los efectos, pero en cualquier caso el procedimiento actúa, necesariamente, bastante a ciegas. Por el contrario, la presente invención resuelve el problema a nivel de aplicación, y por tanto es consciente de los matices del servicio que está proporcionando y puede mejorar su comportamiento en contingencias. La presente invención puede ser aplicada a muchas clases de servicios de red, pero es particularmente adecuada para servicios que requieren un procesamiento intensivo al nivel del servidor, y en los cuales es posible preparar una respuesta más genérica sub-óptima (pero todavía válida) a entregar al agente solicitante en caso de sobrecarga. Un área objetivo especialmente aplicable es aquella en la cual el sistema ofrece servicios personalizados (es decir, suministra datos adaptados para cada usuario específico), posiblemente costosos de obtener, pero también puede devolver una versión (levemente menos óptima) despersonalizada (despersonalizada para un cliente específico) de los datos. Ejemplos de servicios donde la presente invención podría ser aplicada y sería muy útil son los sistemas (servicios) de recomendación y los servicios de personalización de noticias. Estos son solamente dos ejemplos (no limitadores) donde la presente invención es útil. Se entenderá que la presente invención puede ser aplicada a muchos otros tipos de servicios.

Servicios de recomendación:

En este caso los servidores de procesamiento de datos (los servidores que proporcionan el servicio) son motores de recomendaciones. Obtienen solicitudes desde las aplicaciones clientes (encaminadas a través del gestor de sobrecarga) para proporcionar listas de recomendaciones para los usuarios. Una

lista de recomendaciones consiste habitualmente en una lista clasificada de elementos, personalizada para cada usuario de acuerdo a su perfil de usuario. Existe un buen número de algoritmos distintos para motores de recomendaciones (Filtrado Colaborativo, Basado en Contenidos, de Base Social / Gráfica, etc.). El proceso por el cual se produce una recomendación puede ser intensivo en términos de necesidades de cálculo, y además la recomendación necesita, por lo general, ser compuesta en tiempo casi real (para proporcionar respuesta rápida a un usuario que navega por la aplicación cliente). Y puede cambiar de un momento a otro (pues el perfil del usuario está siendo continuamente actualizado, para aprovechar la retroalimentación generada por acciones en vivo); esto limita la utilidad del almacenamiento en memoria caché de las recomendaciones. Como resultado, los motores de recomendaciones deben ser adecuadamente cuantificados para proporcionar suficiente potencia de cálculo a fin de suministrar recomendaciones rápidas, y en los picos de tráfico (muchos clientes conectándose con el sistema de recomendaciones) pueden ocurrir cuellos de botella y congestiones. Esto hace de ellos un entorno perfecto para el mecanismo de gestión de sobrecarga (congestión) que esta invención implementa.

En este escenario de aplicación, la realización de la presente invención puede presentar, entre otras, las siguientes características:

- Un cierto número de motores de recomendaciones están conectados con un gestor de sobrecarga, que recibe las solicitudes y las encamina a uno de los motores, equilibrando la carga.

- El motor de recomendaciones también puede producir listas clasificadas genéricas, no personalizadas (p. ej., usando contenidos de máxima calificación, o populares); esto será usado como un mecanismo de recomendación reincidente (respuesta genérica). El cargador para el

contenido genérico (GL) usa esta modalidad del motor para pre-capturar resultados adecuados para todos los usuarios.

• El gestor de sobrecarga monitoriza el tiempo de respuesta del motor; si es necesario, el Monitor Temporal (RM) aborta la solicitud y entrega recomendaciones reincidentes.

• Análogamente, si la detección de congestión temprana resuelve una situación potencialmente peligrosa, activa el mecanismo de contingencia y comienza a entregar recomendaciones reincidentes a un subconjunto aleatorio de solicitudes.

Servicio personalizado de noticias:

En esta realización, de manera similar al caso de las recomendaciones, se entrega un resultado personalizado para cada usuario conectado con el sistema. Las noticias son seleccionadas de acuerdo al perfil del usuario y / o a la solicitud del usuario, y se compone una página de noticias tomando aquellos elementos de noticias seleccionadas y formateándolas según se necesite. El proceso también es intensivo en procesamiento de datos, ya que combina una fase de filtrado de contenidos y una representación del contenido. En una situación normal producirá contenidos especialmente adaptados para el usuario pero, en casos críticos, podría remitirse a conjuntos pre-generados de elementos de noticias. En este escenario de aplicación, la realización de la presente invención puede presentar las siguientes características:

• Los motores de noticias personalizados actúan como los servidores de procesamiento de datos. Están conectados con perfiles de usuario para cada uno de los usuarios de las plataformas.

- Un conjunto de resultados es pre-generado (p. ej., para unas pocas características distintas de clientes) y actualizado periódicamente para mantenerlo al día. Esto es el contenido reincidente (respuesta genérica) usado por el mecanismo de control de sobrecarga en el caso en que soluciona la necesidad de recortar demandas de procesamiento sobre los motores de noticias.

Una persona experta en la técnica reconocerá inmediatamente que las etapas de diversos procedimientos descritos en lo que antecede pueden ser realizadas por ordenadores programados. En la presente memoria, algunas realizaciones también están concebidas para abarcar dispositivos de almacenamiento de programas, p. ej., medios de almacenamiento de datos digitales, que son legibles por máquina u ordenador, y que codifican programas de instrucciones ejecutables por máquina o ejecutables por ordenador, en donde dichas instrucciones realizan algunas de, o todas, las etapas de dichos procedimientos anteriormente descritos. Los dispositivos de almacenamiento de programas pueden ser, p. ej., memorias digitales, medios de almacenamiento magnético tales como discos magnéticos y cintas magnéticas, controladores de discos rígidos o medios de almacenamiento de datos digitales legibles ópticamente. Las realizaciones también están concebidas para abarcar ordenadores programados para realizar dichas etapas de los procedimientos anteriormente descritos.

Incluso aunque muchas de las realizaciones presentadas se refieren a servicios de red, la presente invención no está limitada a esta clase de servicios, sino que puede ser aplicada a otros servicios proporcionados mediante redes de telecomunicaciones.

La descripción y los dibujos ilustran meramente los principios de la invención.

Aunque la presente invención ha sido descrita con referencia a realizaciones específicas, los expertos en la técnica deberán

entender que los precedentes, y otros varios, cambios, omisiones y agregados en la forma y detalle de las mismas pueden ser realizados en las mismas sin apartarse del ámbito de la invención, según lo definido por las siguientes reivindicaciones.

Además, todos los ejemplos expuestos en la presente memoria están principalmente concebidos de manera expresa para que sean solamente con fines pedagógicos, para ayudar al lector en la comprensión de los principios de la invención y de los conceptos aportados por el inventor, o los inventores, para fomentar la técnica, y han de ser interpretados como no limitadores a tales ejemplos y condiciones específicamente expuestos. Además, todas las afirmaciones en la presente memoria que expongan principios, aspectos y realizaciones de la invención, así como ejemplos específicos de la misma, están concebidas para abarcar los equivalentes de los mismos.

Los expertos en la técnica deberán apreciar que todos los diagramas de bloques en la presente memoria representan vistas conceptuales de circuitos ilustrativos que realizan los principios de la invención. De manera similar, se apreciará que todos los gráficos de flujo, diagramas de flujo, diagramas de transición de estados, pseudo-código y similares representan diversos procesos que pueden ser esencialmente representados en un medio legible por ordenador y por tanto ejecutados por un ordenador o procesador, ya sea que tal ordenador o procesador sea explícitamente mostrado o no.

24
24

REIVINDICACIONES

- 5 1. Un procedimiento para gestionar la congestión en un servicio de red, proporcionado por un grupo de servidores (100), estando el procedimiento caracterizado por comprender las siguientes etapas realizadas en un primer nodo de red (200), en comunicación con el grupo de servidores (100):
- 10 a) Recibir una solicitud de servicio desde un dispositivo cliente del servicio (300) mediante una primera red de telecomunicación;
- 15 b) Después de recibir la solicitud de servicio, determinar (212) si hay o no un estado de sobrecarga en el grupo de servidores (100);
- c) Decidir (213) si se activa un mecanismo de contingencia en base, al menos, en la determinación;
- 20 d) Si el mecanismo de contingencia es activado para dicha solicitud de servicio, no enviar la solicitud de servicio al grupo de servidores (100);
- e) Si el mecanismo de contingencia no es activado:
- 25 e1) Enviar (214) la solicitud de servicio al grupo de servidores (100);
- e2) Si una respuesta de servicio, resultante del procesamiento de la solicitud de servicio por el grupo de servidores, es recibida desde el grupo de servidores antes
- 30 de que haya transcurrido un primer periodo de tiempo, enviar dicha respuesta de servicio al dispositivo cliente (300) y, si no es así, activar el mecanismo de contingencia para dicha solicitud de servicio;
- 35 f) Si, en las etapas c) o e2), se activa el mecanismo de contingencia para dicha solicitud de servicio, obtener (221) desde una base de datos una respuesta genérica pre-almacenada, no resultante del procesamiento por el grupo de servidores de

la solicitud de servicio, y enviar (222) dicha respuesta genérica al dispositivo cliente.

5 2. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, en donde, en la etapa e), si el primer nodo de red (200) no recibe una respuesta de servicio desde el grupo de servidores antes de que haya transcurrido el primer periodo de tiempo, envía además un mensaje al grupo de servidores para abortar el procesamiento de la solicitud de servicio.

10

3. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde la etapa de decidir (213) si se activa un mecanismo de contingencia comprende:

15 - si se determina que no hay estado de sobrecarga, no activar el mecanismo de contingencia;

20 - si se determina que hay un estado de sobrecarga, calcular una probabilidad de activar el mecanismo de contingencia para dicha solicitud de servicio y activar el mecanismo de contingencia en base a dicha probabilidad.

25 4. Un procedimiento de acuerdo a la reivindicación 3, donde dicha probabilidad se calcula en base a parámetros de diseño y a los valores del tiempo de ejecución de solicitudes para las M solicitudes previas de servicio recibidas en el primer nodo de red, donde M es un parámetro de diseño.

30 5. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde la determinación del estado de sobrecarga en el grupo de servidores (100) se basa en los valores de un tiempo de ejecución de solicitudes para las M solicitudes de servicio precedentes, recibidas en el primer nodo de red, donde M es un parámetro de diseño.

35 6. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde, si se activa el mecanismo de contingencia, se estima un tiempo de ejecución de solicitud

25

para dicha solicitud de servicio (215) y se almacena (216) en una base de datos (270) en el primer nodo de red (100).

5 7. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde, si el mecanismo de contingencia no es activado para dicha solicitud de servicio, se calcula un tiempo de ejecución de solicitud como el tiempo transcurrido desde que la solicitud de servicio es enviada al grupo de servidores hasta que la respuesta de servicio es recibida por el primer nodo de red, y dicho tiempo de ejecución de solicitud es almacenado (218) en una base de datos (270) en el primer nodo de red (200).

15 8. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde el primer nodo de red no está co-situado con el grupo de servidores y se comunica con el grupo de servidores mediante la primera red de telecomunicaciones.

20 9. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde la solicitud de servicio y la respuesta de servicio son mensajes de la capa de aplicación de red.

25 10. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde el dispositivo cliente es uno del siguiente grupo: un teléfono móvil, un ordenador portátil, un ordenador personal, un teléfono inteligente, un televisor inteligente, un televisor, un televisor de IPTV, una tableta o cualquier otro dispositivo electrónico.

30 11. Un procedimiento de acuerdo a cualquiera de las reivindicaciones previas, donde la etapa e1) de enviar la solicitud de servicio (214) al grupo de servidores comprende:
35 - Descomponer la solicitud de servicio en varias sub-solicitudes,
- Enviar la primera sub-solicitud al grupo de servidores y

- Enviar cada una del resto de sub-solicitudes al grupo de servidores cuando se recibe una respuesta parcial, resultante del procesamiento por el grupo de servidores de la sub-solicitud precedente;

5 y la etapa e2) comprende:

- Si todas las respuesta de servicio parciales, resultantes del procesamiento por el grupo de servidores de cada sub-solicitud, son recibidas desde el grupo de servidores antes de que haya transcurrido el primer periodo de tiempo, componer la
10 respuesta de servicio en base a todas las respuestas parciales recibidas, y enviar dicha respuesta de servicio al nodo cliente, y si no es así, activar el mecanismo de contingencia.

12. Un procedimiento de acuerdo a cualquiera de las
15 reivindicaciones previas, donde el servicio de red es un servicio de información, la respuesta de servicio resultante del procesamiento de la solicitud de servicio por el grupo de servidores incluye información seleccionada en base al contenido de la solicitud de servicio y / o al perfil del
20 cliente, y la respuesta genérica consiste en información seleccionada antes de recibir la solicitud de servicio.

13. Un nodo de red (200) para gestionar la congestión de un servicio de red, estando el nodo de red caracterizado por
25 comprender:

- Medios de recepción para recibir (211) una solicitud de servicio desde un dispositivo cliente del servicio (300), mediante una primera red de telecomunicación;

30 - Medios para determinar si hay o no un estado de sobrecarga (212) y decidir (213) si se activa un mecanismo de contingencia en base, al menos, a la determinación;

- Medios para, si el mecanismo de contingencia es activado para dicha solicitud de servicio, no enviar la solicitud de servicio
35 a un grupo de servidores (100) que proporcionan el servicio de red;

- Medios, si el mecanismo de contingencia no es activado, para:

26/26

- Enviar (214) la solicitud de servicio al grupo de servidores (100) que proporcionan el servicio de red;
- Si una respuesta de servicio, resultante del procesamiento de la solicitud de servicio por el grupo de servidores, es recibida desde el grupo de servidores antes de que haya transcurrido un primer periodo de tiempo, enviar dicha respuesta de servicio al dispositivo cliente (300) y, si no es así, activar el mecanismo de contingencia para dicha solicitud de servicio;

- Medios para, si el mecanismo de contingencia es activado para dicha solicitud de servicio, obtener (221) desde una base de datos una respuesta genérica pre-almacenada, no resultante del procesamiento por el grupo de servidores de la solicitud de servicio, y enviar (222) dicha respuesta genérica al dispositivo cliente.

14. Sistema para gestionar la congestión de un servicio de red, estando el sistema caracterizado por comprender:

- Un grupo de servidores (100) configurados para proporcionar el servicio de red, por la recepción de solicitudes de servicio de clientes, el procesamiento de las solicitudes y la generación de respuestas de servicio, en base, al menos, al contenido de las solicitudes de servicios de clientes;

- Un nodo de red (200) que comprende:

- Medios de recepción para recibir (211) una solicitud de servicio desde un dispositivo cliente del servicio (300), mediante una primera red de telecomunicación;

- Medios para determinar si hay o no una condición de sobrecarga en el grupo de servidores (100) y decidir (213) si se activa un mecanismo de contingencia, en base, al menos, a la determinación;

- Medios para, si se activa el mecanismo de contingencia para dicha solicitud de servicio, no enviar la solicitud de servicio al grupo de servidores (100);

- Medios, si el mecanismo de contingencia no es activado, para:

- Enviar (214) la solicitud de servicio al grupo de servidores (100);

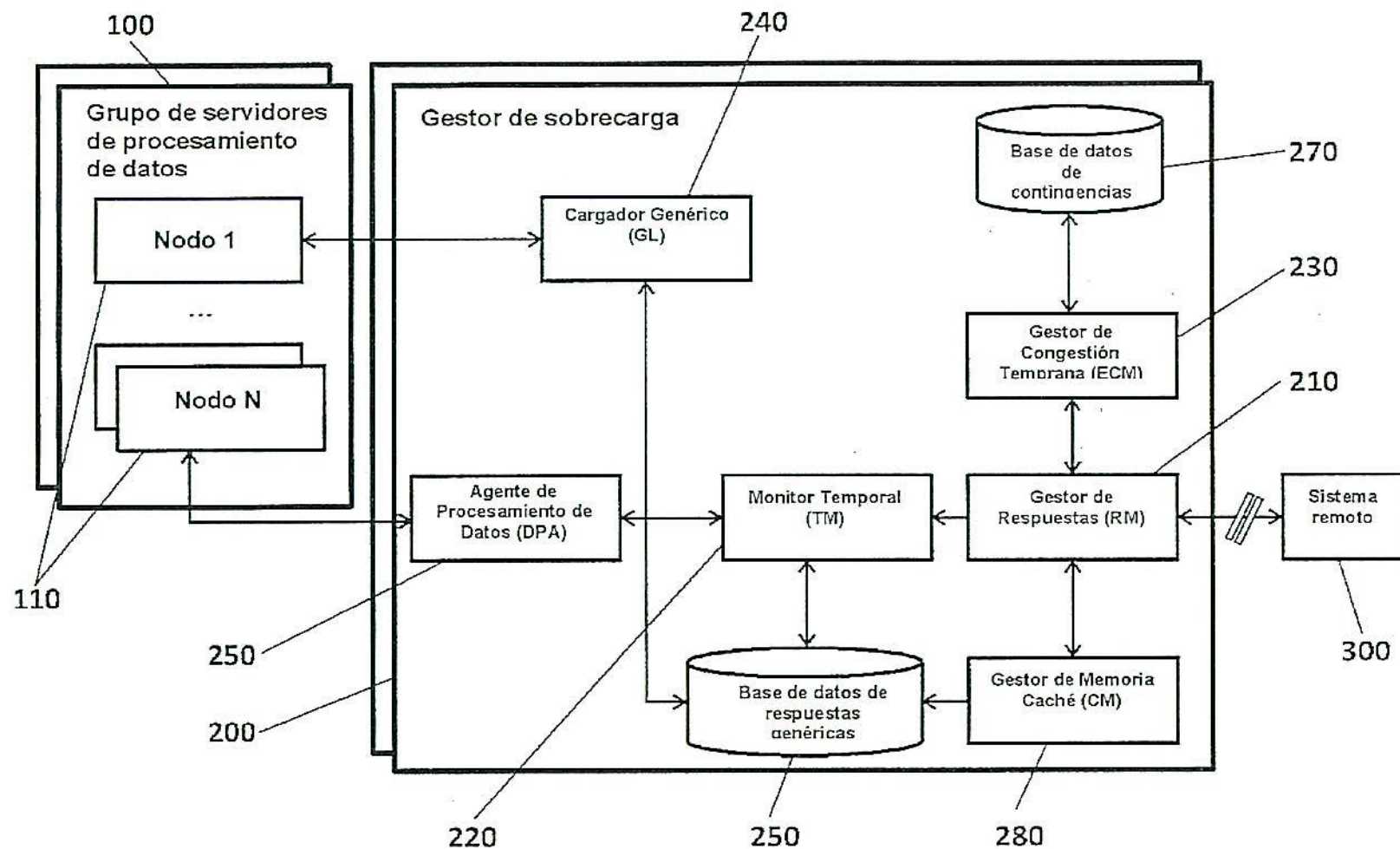
- Si una respuesta de servicio, resultante del procesamiento de la solicitud de servicio por el grupo de servidores, es recibida desde el grupo de servidores antes de que haya transcurrido un primer periodo de tiempo, enviar dicha respuesta de servicio al dispositivo cliente (300) y, si no es así, activar el mecanismo de contingencia para dicha solicitud de servicio;

- Medios para, si el mecanismo de contingencia es activado para dicha solicitud de servicio, obtener (221) desde una base de datos una respuesta genérica pre-almacenada, no resultante del procesamiento por el grupo de servidores de la solicitud de servicio, y enviar (222) dicha respuesta genérica al dispositivo cliente.

15. Un medio de almacenamiento de datos digitales que almacena un programa de ordenador que comprende instrucciones ejecutables por ordenador, para realizar el procedimiento de acuerdo a cualquiera de las reivindicaciones 1 a 12, cuando el programa es ejecutado en un ordenador.

RESUMEN

Un procedimiento, un sistema y un dispositivo para gestionar estados de congestión / sobrecarga al nivel de aplicación, experimentados en un servicio de red proporcionado mediante una red de telecomunicaciones, que permite proporcionar respuestas de servicio rápidas, válidas y completas al cliente del servicio incluso cuando los servidores del servicio están sobrecargados. La gestión de congestión es proporcionada por un nodo intermedio comunicado con el dispositivo cliente y con el servidor, o servidores, que proporciona(n) el servicio de red. A fin de controlar la sobrecarga de uno o más servidores, este nodo usa un doble mecanismo de congestión, un mecanismo de detección de congestión temprana y un mecanismo de limitación del tiempo de respuesta.



1/4

FIG. 1

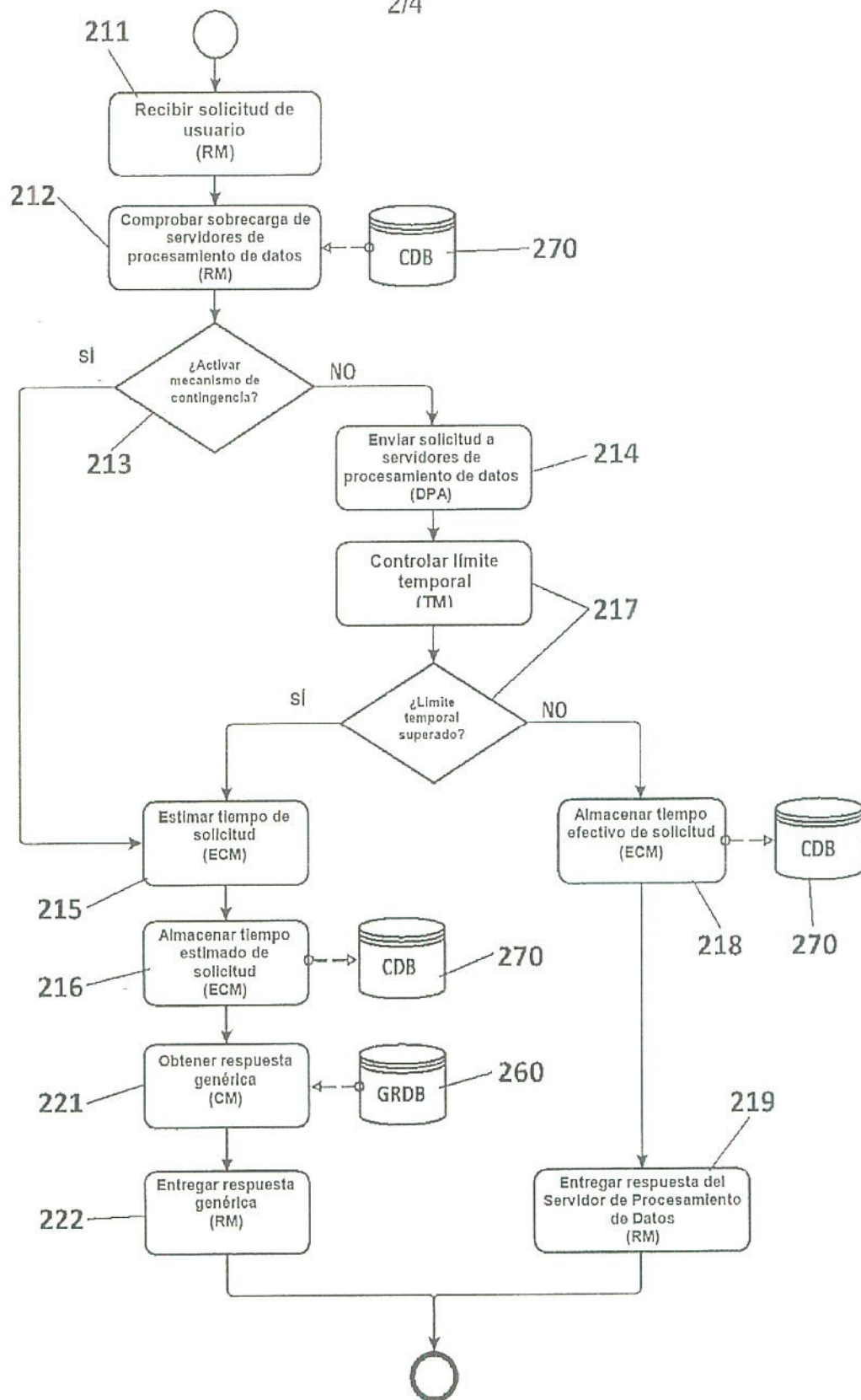


FIG. 2

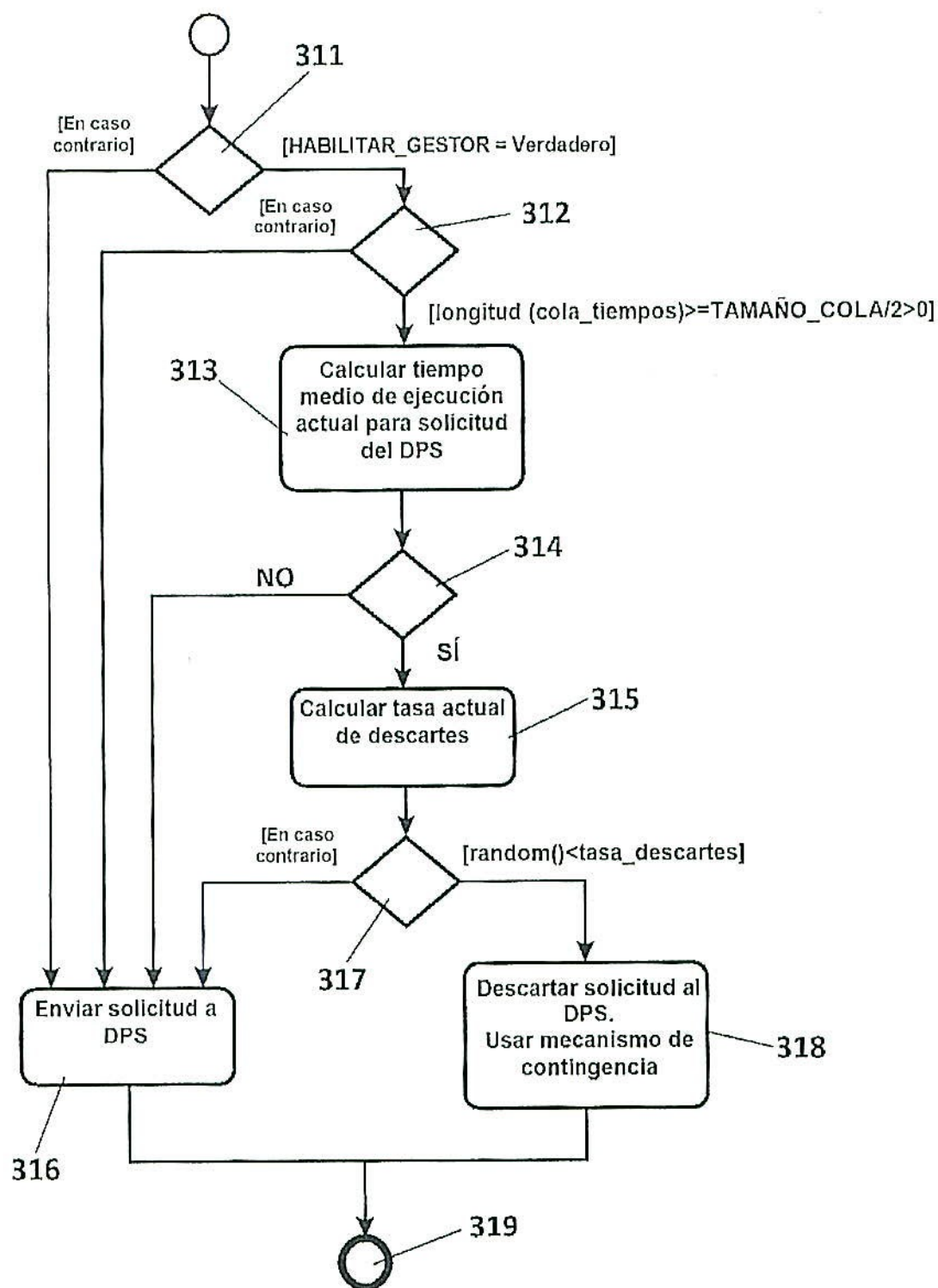


FIG. 3

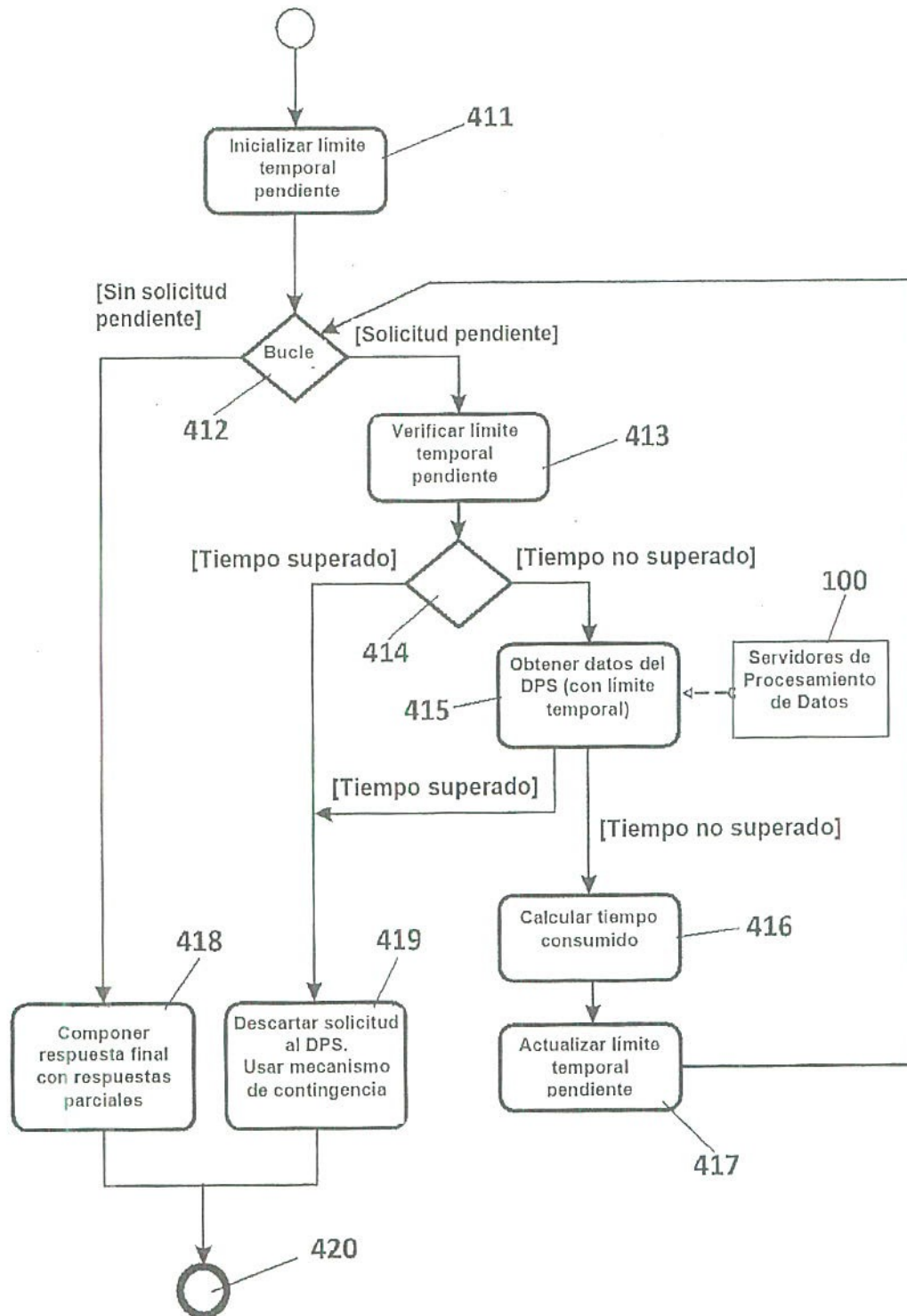


FIG. 4

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO

NIT : 800.176.089-2

- / -

30 *21*



Industria y Comercio
SUPERINTENDENCIA

RECIBO DE CAJA

No. 16 - 0010815

Bogotá D.C., Febrero 08 de 2016 - 12:21:28

RECIBIDO DE : CUESTA Y ASOCIADOS LTDA

NI 800.075.616

*** Soporte del Pago ***

TIPO PAGO	BANCO	CUENTA	No. PAGO	FECHA PAGO	VR. PAGO
CONSIGNACION	BANCO DE BOGOTA	062754387	769936845	08/02/2016	496.000.00

*** Conceptos Pagados ***

CANT. RENTISTICO	CONCEPTO	Vr.UNDITARIO	Vr.CONCEPTO
1 50005-01-01 SOLICITUDES	1 TRAMITES DE SOL. DE PATENTE DE INVENCIÓN	80.000.00	80.000.00
			\$80.000.00

SON: **OCHENTA MIL PESOS MONEDA CORRIENTE***

Responsable: _____

Recibo de Caja Aplicado al Expediente No. _____

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO



No. 16-029128- -00000-0000

Fecha: 2016-02-08 12:28:57 Dep: 2020 DIR.NUEVASCR
Tra: 2 PATENTES Eve: 1 REGDEPOSITO
Act: 411 PRESENTACION Folios: 33

Sede Centro: Carrera 13 No. 27 - 00 Pisos 3,4,5 y 10 Bogotá, D.C.- Colombia

Web: www.sic.gov.co e-mail: info@sic.gov.co Conmutador: (571) 5870000 Fax: (571) 5870284 Línea: 018000-910165 Call Center: (571) 6513240

Febrero 8 de 2016 - 12:21:29

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO

NIT : 800.176.089-2

- / -

31
27



RECIBO DE CAJA

No. 16 - 0010816

Bogotá D.C., Febrero 08 de 2016 - 12:21:28

RECIBIDO DE : CUESTA Y ASOCIADOS LTDA

NI 800.075.616

*** Soporte del Pago ***

TIPO PAGO	BANCO	CUENTA	No. PAGO	FECHA PAGO	VR PAGO
CONSIGNACION	BANCO DE BOGOTA	062754387	769936845	08/02/2016	496.000.00

*** Conceptos Pagados ***

CANT. RENTISTICO	CONCEPTO	Vr.UNDITARIO	Vr.CONCEPTO
5 50005-01-01 SOLICITUDES	1607 REIVINDICACION PATENTE UNITARIA ADICIONAL A LAS 10 INIC	40.000.00	200.000.00
			\$200.000.00

SON: **DOSCIENTOS MIL PESOS MONEDA CORRIENTE***

Responsable: _____

Recibo de Caja Aplicado al Expediente No. _____

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO



No. 16-029128- -00000-0000

Fecha: 2016-02-08 12:28:57 Dep. 2020 DIR.NUEVASCR
Tra 2 PATENTES Eve: 1 REGDEPOSITO
Act 411 PRESENTACION Folios 33

Sede Centro: Carrera 13 No. 27 - 00 Pisos 3,4,5 y 10 Bogotá, D.C.- Colombia

Web: www.sic.gov.co e-mail: Info@sic.gov.co Conmutador: (571) 5870000 Fax: (571) 5870284 Línea: 018000-910165 Call Center: (571) 6513240

Febrero 8 de 2016 - 12 21:29

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO

NIT : 800.176.089-2

- / -

32



Industria y Comercio
SUPERINTENDENCIA

RECIBO DE CAJA

No. 16 - 0010817

Bogotá D.C., Febrero 08 de 2016 - 12:21:28

RECIBIDO DE : CUESTA Y ASOCIADOS LTDA

NI 800.075.616

*** Soporte del Pago ***

TIPO PAGO	BANCO	CUENTA	No. PAGO	FECHA PAGO	VR. PAGO
CONSIGNACION	BANCO DE BOGOTA	062754387	769936845	08/02/2016	496.000.00

*** Conceptos Pagados ***

CANT.	RENTISTICO	CONCEPTO	Vr.UNDITARIO	Vr.CONCEPTO
1	50005-01-01 SOLICITUDES	82 INVOCACION DE UNA PRIORIDAD EN NUEVAS CREACIONES	216.000.00	216.000.00
				\$216.000.00

SON: **DOSCIENTOS DIECISEIS MIL PESOS MONEDA CORRIENTE***

Responsable: _____

Recibo de Caja Aplicado al Expediente No. _____

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO



No. 16-029128- -00000-0000

Fecha: 2016-02-08 12:28:57 Dep 2020 DIR NUEVASCR
Tra. 2 PATENTES Eve 1 REGDEPOSITO
Act. 411 PRESENTACION Folios 33

Sede Centro: Carrera 13 No. 27 - 00 Pisos 3,4,5 y 10 Bogotá, D.C.- Colombia

Web: www.sic.gov.co e-mail: info@sic.gov.co Conmutador: (571) 5870000 Fax: (571) 5870284 Línea: 018000-910165 Call Center: (571) 6513240

Febrero 8 de 2016 - 12:21:29

33

PATENTE DE INVENCION ☒

MODELO DE UTILIDAD ☐

Art 33 Decisión 486/00



Indicación que se solicita una patente.



Datos de identificación del solicitante o de la persona que presenta la solicitud



Descripción de la invención



Dibujos de ser estos pertinentes



Comprobante de pago de las tasas establecidas (De ser el caso formato de descuento)

Completa ☒

Incompleta ☐

PATENTE DE INVENCION PCT ☐

MODELO DE UTILIDAD PCT ☐

Art.33 Decisión 486/00, Circular Única.



Indicación que se solicita una PCT



Copia de la solicitud en español, tal como fue presentada inicialmente (capítulo descriptivo, reivindicatorio, resumen)



Dibujos de ser estos pertinentes



Comprobante de pago de las tasas establecidas (de ser el caso formato de descuento)

Completa ☐

Incompleta ☐

DISEÑO INDUSTRIAL ☐

(Art. 119 Decisión 486/00)



Indicación que se solicita Diseño industrial



Datos de identificación del solicitante o de la persona que presenta la solicitud



Representación gráfica y fotográfica del Diseño industrial o muestra del material que incorpora el diseño



Comprobante de pago de las tasas establecidas

Completa ☐

Incompleta ☐

ESQUEMA DE TRAZADO ☐

(Art. 92 Decisión 486/00)



Indicación que se solicita un esquema de trazado



Datos de identificación del solicitante o de la persona que presenta la solicitud



Representación gráfica de un esquema de trazado



Comprobante de pago de las tasas establecidas

Completa ☐

SUPERINTENDENCIA DE INDUSTRIA Y COMERCIO



No. 16-029128- -00000-0000

Fecha: 2016-02-08 12:28:57 Dep. 2020 DIR.NUEVASCR
Tra. 2 PATENTES Eve 1 REGDEPOSITO
Act 411 PRESENTACION Folios 33