(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 620 322**

(51) Int. CI.:

H04L 12/863 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **08.03.2013 PCT/US2013/029918**

(87) Fecha y número de publicación internacional: **26.09.2013 WO2013142099**

(96) Fecha de presentación y número de la solicitud europea: **08.03.2013 E 13763455 (6)**

(97) Fecha y número de publicación de la concesión europea: **21.12.2016 EP 2829029**

(54) Título: **Programador de entrada/salida de lectura estrangulada**

(30) Prioridad:

20.03.2012 US 201213506006

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
28.06.2017

(73) Titular/es:

DRW TECHNOLOGIES LLC (100.0%)
540 West Madison Street Suite 2500
Chicago, Illinois 60661, US

(72) Inventor/es:

KRAMNIK, ALEXANDER;
MALNICK, NICOLAUS P. MALNICK y
HAYHURST, LYLE

(74) Agente/Representante:

CARVAJAL Y URQUIJO, Isabel

DESCRIPCIÓN

Programador de entrada/salida de lectura estrangulada

Campo de la invención

La presente invención se refiere al procesamiento electrónico de datos.

Antecedentes de la invención

Con la proliferación de la potencia de procesamiento y del ancho de banda de comunicación electrónica, una multitud de fuentes están generando y transmitiendo una explosión de datos. Estos datos se recopilan y se almacenan para su análisis para una amplia diversidad de aplicaciones. Un área de este tipo en la que los mismos han visto un particular crecimiento explosivo es en la compraventa electrónica de productos financieros.

Hubo un tiempo en el que solo había bolsas con negociación a viva voz, en los que los corredores y los comerciantes, o, más en concreto, los compradores y los vendedores, se hubieran reunido para negociar en persona. Más recientemente, se han introducido unas bolsas electrónicas que procesan la equiparación automática y electrónica de pujas y ofertas. Por lo tanto, los métodos bursátiles han evolucionado de un proceso que requiere mucho trabajo manual a una plataforma electrónica posibilitada por la tecnología.

La compraventa electrónica se basa en general en ordenadores (de anfitrión) centralizados, una o más redes informáticas, y ordenadores (de cliente) participantes en bolsa. En general, la bolsa de anfitrión incluye uno o más ordenadores centralizados. Las operaciones de la bolsa de anfitrión incluyen, por lo general, equiparar órdenes, mantener posiciones y carteras de órdenes, información de cotizaciones, y gestionar y actualizar la base de datos durante el día bursátil hábil así como ejecuciones por lotes durante la noche. La bolsa de anfitrión también está equipada con interfaces externas que mantienen contacto con proveedores de cotizaciones y otros sistemas de información de cotizaciones.

Usando dispositivos de cliente, los participantes en el mercado enlazan con la bolsa de anfitrión a través de una o más redes. Una red es un grupo de dos o más ordenadores o dispositivos enlazados entre sí. Hay muchos tipos de redes cableadas e inalámbricas, tales como, por ejemplo, redes de área local y redes de área extensa. Las redes también se pueden caracterizar por la topología, el protocolo y la arquitectura. Por ejemplo, algunos participantes en el mercado pueden enlazar con el anfitrión a través de una conexión directa tal como una Red Digital de Servicios Integrados (ISDN, *Integrated Services Digital Network*) o T1. Algunos participantes enlazan con la bolsa de anfitrión a través de conexiones directas y a través de otros componentes de red comunes tales como servidores de alta velocidad, encaminadores y pasarelas. Existen muchos tipos diferentes de redes y combinaciones de tipos de red que pueden enlazar a los comerciantes con la bolsa de anfitrión. Internet se puede usar para establecer una conexión entre el dispositivo de cliente y la bolsa de anfitrión.

Internet es una red global de ordenadores. Los servidores de red soportan unas capacidades de hipertexto que permiten que Internet enlace webs de documentos entre sí. Las interfaces de usuario tales como las Interfaces Gráficas de Usuario (GUI, *Graphical User Interface*) se usan, por lo general, para explorar Internet para recuperar un documento relevante. Los Localizadores Uniformes de Recursos (URL, *Uniform Resource Locator*) se usan para identificar páginas web y sitios web específicos en Internet. Los URL también identifican la dirección del documento que se va a recuperar de un servidor de red.

Para transmitir la información digitalizada de una forma fiable, Internet usa un diseño de conmutación de paquetes. La conmutación de paquetes descompone bloques de información digitalizada en unos fragmentos más pequeños que se denominan paquetes. Estos paquetes se transmiten a través de la red, por lo general por diferentes rutas y, entonces, se vuelven a ensamblar en su destino. Véase, por ejemplo, Len Kleinrock, "*Information Flow in Large Communications Nets*", *RLE Quarterly Progress Report* (1960); Len Kleinrock, *Communication Nets* (1964). Véase también Paul Baren, "*On Distributed Communications Networks*", *IEEE Transactions on Systems* (marzo de 1964).

Se hizo referencia a la conmutación de paquetes original empleada por Internet como Protocolo de Control de Transmisión/Protocolo de Internet (TCP/IP, *Transmission Control Protocol/Internet Protocol*). El Protocolo de Control de Transmisión (TCP, *Transmission Control Protocol*) paquetiza la información y vuelve a ensamblar la información tras la llegada. El Protocolo de Internet (IP, *Internet Protocol*) encamina paquetes mediante la encapsulación de los paquetes entre redes. Véase, por ejemplo, Robert Kahn y Vincent Cerf, "*A Protocol for Packet Network Intercommunication*", *IEEE Transactions on Communications Technology* (mayo de 1974). Existen muchos protocolos adicionales que se usan para retransmitir el tráfico de Internet.

La conmutación de paquetes más reciente empleada por Internet incluye el Protocolo de Datagramas de Usuario/Protocolo de Internet (UDP/IP, *User Datagram Protocol/Internet Protocol*). Con frecuencia, las bolsas de

anfitrión transmiten información acerca del mercado bursátil a dispositivos de cliente que usan el protocolo UDP/IP. UDP/IP usa un modelo de transmisión simple sin diálogos de toma de contacto implícitos para proporcionar fiabilidad, ordenación o integridad de datos. El UDP proporciona un servicio poco fiable y los datagramas pueden llegar desordenados, mostrarse duplicados o extraviarse sin previo aviso; por lo tanto, el UDP asume que la comprobación y la corrección de errores o bien no es necesaria o bien se realiza en la aplicación, lo que evita la tara de tal procesamiento al nivel de interfaz de red. Las aplicaciones temporalmente sensibles usan con frecuencia el UDP debido a que descartar paquetes es preferible a aguardar paquetes retardados, lo que puede que no sea una opción en un sistema en tiempo real. Existen muchos otros protocolos de comunicación TCP/IP y UDP/TP que se pueden usar para transmitir o recibir datos a lo largo de una red.

TCP/IP y UDP/IP son ejemplos de los protocolos de comunicación que se pueden clasificar por medio del modelo de Interconexión de Sistemas Abiertos (OSI, *Open Systems Interconnection*). El modelo de OSI es una prescripción de caracterizar y normalizar las funciones de un sistema de comunicación en términos de capas. Funciones de comunicación similares se agrupan en capas lógicas, con una capa atendiendo a la capa por encima de la misma, y siendo atendida por la capa por debajo de ella. La capa más baja del modelo de OSI es la capa física, que también se conoce como “Capa 1”, y que se abrevia comúnmente como “PHY”. La capa física define las especificaciones eléctricas y físicas entre un dispositivo informático y un medio de transmisión, tal como cable de cobre o de fibra óptica. Los ejemplos de capas físicas incluyen DSL, ISDN, capa física de Ethernet (10BASE-T, 10BASE2, 100BASE-TX, 1000BASE-T, y otras variedades), Infiniband, T1, y similares.

Por encima de la capa física se encuentra la capa de enlace de datos, que también se conoce como “Capa 2”. La capa de enlace de datos proporciona los medios funcionales y procedimentales para transferir datos entre entidades de red y detectar y corregir los errores que pueden tener lugar en la capa física. Los ejemplos de capas de enlace de datos incluyen ARP, ATM, Ethernet, Frame Relay, 802.11, Infiniband, anillo con paso de testigo, y similares. Tanto para TCP/IP como para UDP/IP, la capa de enlace de datos es responsable de la encapsulación de paquetes de IP en tramas, de la sincronización de tramas, de la detección y manipulación de errores y del direccionamiento por medio de control de acceso a medios (MAC, *media access control*), entre otros.

Por encima de la capa de enlace de datos se encuentra la capa de red, que también se conoce como ‘Capa 3’. Esta capa se ocupa de llevar en la práctica los datos de un ordenador a otro, incluso si el ordenador es remoto o se encuentra en una red diferente. Las funciones de la capa de red incluyen comunicación sin conexión, direccionamiento de anfitrión y reenvío de mensajería. Los protocolos de capa de red a modo de ejemplo incluyen IPv4 e IPv6, ICMP, IGMP, PIM-SM y PIM-DM, Infiniband, IPX, y similares. Tanto para TCP/IP como para UDP/IP, el Protocolo de Internet, o “IP”, es la implementación de Capa 3 que se usa para encaminar paquetes a través de una o muchas redes.

Por encima de la capa de red se encuentra la capa de transporte, que también se conoce como ‘Capa 4’. La capa de transporte proporciona servicios de comunicación de extremo a extremo tales como soporte de flujo de datos orientado a la conexión, fiabilidad, control de flujo, evitación de congestión, orientación de bytes y multiplexión. Los ejemplos de capas de transporte incluyen TCP, UDP, RDP, IL, Infiniband, ATP, y similares.

Internet es un ejemplo de una implementación del modelo de OSI. Existen muchos otros ejemplos de redes, incluyendo redes personales, redes de área local, redes domésticas, redes de área de almacenamiento, redes en universidades, redes medulares, redes de área metropolitana, redes de área extensa, redes privadas virtuales, y similares. El uso, el nivel de confianza y los derechos de acceso difieren entre estos tipos diferentes de redes.

Un subconjunto de problemas de recopilación de datos se ocupa de los datos que se generan a partir de una diversidad de fuentes y que se transmiten a una tasa muy alta a través de una red informática que usa un protocolo de comunicación tal como, por ejemplo, UDP/IP. Cualquier solución que capture y dé persistencia a estos datos para su uso se encuentra sujeta a algunos requisitos básicos. Dar persistencia se refiere a la característica de las configuraciones de la información en un programa o máquina que sobrevive al proceso que la creó. Sin persistencia, esta información solo existiría en memoria de acceso aleatorio (RAM, *random access memory*), y se perdería cuando esta RAM se quedara sin alimentación, tal como el apagado de un ordenador. En la práctica, la persistencia se logra mediante el almacenamiento de la información como datos en un almacenamiento no volátil tal como una unidad de disco duro o una memoria flash.

En primer lugar, los datos a los que se ha dado persistencia han de ser una instantánea uno por uno de la totalidad de los datos que se difundieron a partir de la fuente. Si se transmiten datos, pero no se les da persistencia, esto se conoce como un ‘hueco’ o un ‘descarte’. Muchas categorías de uso se vienen abajo si los datos a los que se ha dado persistencia contienen huecos. Debido a que los datos transmitidos son, por lo general, no repetibles, la solución de captura de datos ha de garantizar que todos los paquetes transmitidos se capturarán y se les dará persistencia.

En segundo lugar, en unas condiciones de carga de red normales, los usuarios han de ser capaces de acceder a los datos a los que se ha dado persistencia incluso a medida que se están capturando y se está dando persistencia a

nuevos datos transmitidos. La incapacidad a la hora de acceder a los datos a los que se ha dado persistencia perjudicaría a muchos procesos comerciales debido a que los procesos comerciales serían incapaces de responder a los cambios en las condiciones en su entorno. Por ejemplo, el acto de detectar y de responder a la pérdida de paquetes en una red se vería gravemente obstaculizado si el administrador de red fuera incapaz de acceder a los datos de red a los que se ha dado persistencia hasta el final del día comercial. El acceso por parte del usuario tiene el potencial de interrumpir el proceso de dar persistencia a los datos, debido a que las dos actividades están, ambas, disputándose el mismo conjunto finito de recursos informáticos. Este potencial se vuelve más probable a medida que aumenta la carga de tráfico de red. A medida que la carga de red aumenta hacia la capacidad de captura y de persistencia, los sistemas existentes permitirán que la tasa de captura se deteriore como resultado de un acceso aumentado por parte del usuario, lo que conduce a huecos de datos.

Los sistemas existentes que intentan solucionar este problema comparten un diseño común. El enfoque habitual es el uso de un dispositivo de soporte físico al que se hace referencia como derivación de red pasiva, que proporciona una forma de acceder a los datos que fluyen a través de una red informática. La derivación de red pasiva duplica un flujo de paquetes y encamina el flujo de paquetes hacia una tarjeta de captura de protocolo de comunicación de red que está instalada en un aparato de captura de paquetes. La tarjeta de captura de protocolo de comunicación de red entrega los paquetes entrantes a un proceso en el sistema, que los da persistencia a en un medio de almacenamiento. La implementación de sistema habitual usa un ordenador que está ejecutando un sistema operativo y una unidad de disco magnético o de tipo flash para el almacenamiento.

Los sistemas existentes gestionan la lectura y escritura de información en almacenamiento a través de un proceso algorítmico sencillo. La lectura hace referencia al acceso por parte del usuario a una información almacenada, mientras que escritura se refiere al almacenamiento de la información. Un ejemplo es el ‘orden cíclico’, en el que un programador selecciona una lectura y escritura a la que apunta un contador a partir de una lista, después de lo cual el contador se incrementa y, si se alcanza el extremo, el programador vuelve al principio de la lista. Otro ejemplo es ‘primero en entrar - primero en salir’, en el que un programador almacena las lecturas y escrituras en una cola, siendo la primera lectura o escritura que se va a añadir a la cola la primera lectura o escritura sobre la que se va a actuar, en el que el procesamiento avanza de forma secuencial en el mismo orden.

La capacidad de las soluciones existentes para cumplir los criterios que se han bosquejado en lo que antecede recae en garantizar una tasa elevada y consistente de dar persistencia a (escribir) los datos entrantes en el medio de almacenamiento en unas condiciones de carga que consisten tanto en persistencia de transmisión como en acceso por parte del usuario. Con las cargas habituales que consisten en mezclas de solicitudes de lectura y de escritura, los medios de almacenamiento muestran una alta variabilidad en la latencia de las solicitudes de escritura individuales y un rendimiento total global más bajo. Para los medios magnéticos, esto es debido a la latencia del tiempo de búsqueda a medida que los cabezales de lectura - escritura pasan por los cilindros de almacenamiento. Para los medios de tipo flash, esto es lo que se ha denominado el ‘efecto bañera’: altos números para las lecturas y escrituras puras, pero unos mucho más bajos para las cargas de trabajo mixtas.

Esta variabilidad en el rendimiento de escritura eficaz del medio de almacenamiento deteriora la capacidad para mantener un rendimiento de escritura elevado y consistente. Esto, a su vez, crea una contrapresión sobre el resto del sistema. Si esta contrapresión alcanza la tarjeta de captura de protocolo de comunicación de red, la tarjeta no tiene otra opción que descartar los paquetes entrantes debido al desbordamiento de almacenamiento intermedio en memoria interna. Esto ocurre debido a que una solución de captura es de naturaleza pasiva, lo que quiere decir que la presencia de la solución de captura en la red es desconocida para otros dispositivos de red. Como resultado, el sistema no puede solicitar que se retransmitan paquetes si la tarjeta de captura de protocolo de comunicación de red no puede mantener el ritmo de la tasa de paquetes.

Las soluciones existentes usan un número de mecanismos para intentar evitar que esta ‘contrapresión’ (la acumulación de datos que tiene lugar cuando las memorias intermedias están llenas y no pueden recibir más datos) alcance la tarjeta de captura de protocolo de comunicación de red. Tales mecanismos usan una región de memoria - o bien en la propia tarjeta, o bien dentro del proceso o bien dentro del sistema operativo (que se conoce como la memoria caché de página) - para crear una o más memorias intermedias entre el protocolo de comunicación de red, la tarjeta de captura y el medio de almacenamiento. A pesar de que estos mecanismos pueden suavizar la variabilidad en el rendimiento de escritura del medio de almacenamiento, cuando la totalidad de las memorias intermedias se llenan, la contrapresión alcanza de forma inevitable la tarjeta de captura de protocolo de comunicación de red, lo que la fuerza a descartar los paquetes. Tales mecanismos también presentan la desventaja de llevar recursos de memoria lejos del resto del sistema.

El documento US 2012/066449 A1 divulga un sistema y método para programar de forma eficaz operaciones de lectura y de escritura entre una pluralidad de dispositivos de almacenamiento de estado sólido. Un sistema informático comprende ordenadores de cliente y agrupaciones de almacenamiento de datos que están acoplados entre sí por medio de una red. Una agrupación de almacenamiento de datos utiliza unidades de estado sólido y células de memoria flash para el almacenamiento de datos. Un controlador de almacenamiento dentro de una agrupación de almacenamiento de datos comprende un programador de E/S. El controlador de almacenamiento está

configurado para recibir una solicitud de lectura dirigida al medio de almacenamiento de datos, e identificar al menos un primer dispositivo de almacenamiento de la pluralidad de dispositivos de almacenamiento que contiene datos seleccionados como objetivo por la solicitud de lectura. En respuesta o bien a la detección o bien a la predicción de que el primer dispositivo de almacenamiento mostrará un rendimiento variable, el controlador está configurado para generar una solicitud de lectura de reconstrucción que está configurada para obtener los datos a partir de uno o más dispositivos de la pluralidad de dispositivos de almacenamiento que no sean el primer dispositivo de almacenamiento.

Sumario de la invención

De acuerdo con los principios de la presente invención, se proporcionan aplicaciones y métodos de programador de entrada/salida de lectura estrangulada. Un programador de entrada/salida con estrangulación de lectura toma solicitudes de escritura para los datos capturados de una red, proporciona estos datos a un sistema que da persistencia a los datos capturados, y toma solicitudes de lectura procedentes de sistemas de usuario externos. La tasa de solicitudes de lectura y de escritura se determina al mantener dos ventanas deslizantes a lo largo de solicitudes de escritura previas, siendo la segunda ventana más larga que la primera. El programador de entrada/salida con estrangulación de lectura está configurado de tal modo que, cuando la actividad de solicitudes de escritura supera un umbral según se determina a lo largo de la primera ventana, el programador de entrada/salida con estrangulación de lectura estrangula el flujo de solicitudes de lectura. Se proporciona un medio de almacenamiento en el cual se reenvían las solicitudes de lectura y de escritura. Cuando la actividad de solicitudes de lectura y de escritura cae por debajo de un umbral según se computa a lo largo de la segunda ventana, el programador de entrada/salida con estrangulación de lectura reanuda la atención a solicitudes de lectura.

El presente sumario presenta, de forma simplificada, conceptos que se describen adicionalmente en lo sucesivo en la Descripción detallada. El presente sumario no tiene por objeto identificar características clave o características esenciales de la materia objeto que se reivindica, ni el mismo tiene por objeto su uso como una ayuda en la determinación del alcance de la materia objeto que se reivindica.

Breve descripción del dibujo

La figura 1 es un dibujo conceptual que muestra una visión de conjunto de sistema de alto nivel que usa derivaciones de red pasivas.

La figura 2 es un dibujo conceptual que muestra una visión de conjunto de sistema de alto nivel sin usar derivaciones de red pasivas.

La figura 3 es una gráfica que ilustra lo que le ocurre a las tasas de lectura y de escritura a medida que sube la tasa de captura con un programador habitual de la técnica anterior.

La figura 4 es una gráfica que ilustra lo que le ocurre a las tasas de lectura y de escritura a medida que sube la tasa de captura.

La figura 5 es una gráfica que ilustra una ventana de retroceso en un escenario en el que la tasa de datos aumenta con mucha rapidez.

La figura 6 es una gráfica que ilustra una ventana de ascenso en un escenario en el que la tasa de captura es bastante plana y se encuentra justo por debajo del umbral, pero presenta algunos picos de tal modo que los valles se encuentran por debajo del umbral y los picos se encuentran justo por encima del umbral.

Descripción detallada de una realización preferida

Con las cargas habituales que consisten en mezclas de solicitudes de lectura y de escritura, los medios de almacenamiento muestran una alta variabilidad en la latencia de las solicitudes de escritura individuales y un rendimiento total global más bajo. Las soluciones existentes tienen una región de memoria para crear una o más memorias intermedias entre el protocolo de comunicación de red, la tarjeta de captura y el medio de almacenamiento. Cuando la totalidad de las memorias intermedias se llenan, no obstante, la contrapresión alcanza de forma inevitable la tarjeta de captura de protocolo de comunicación de red, lo que la fuerza a descartar los paquetes. Tales mecanismos también tienen la desventaja de llevar recursos de memoria lejos del resto del sistema. La presente invención elimina casi en su totalidad la variabilidad del medio de almacenamiento. Esto, a su vez, reduce la posibilidad de ejercer una contrapresión sobre la tarjeta de captura de protocolo de comunicación de red y de forzar a la misma a descartar los paquetes.

De acuerdo con los principios de la presente invención, se proporcionan aplicaciones y métodos de programador de entrada/salida de lectura estrangulada. De acuerdo con los principios de la presente invención, se proporcionan

aplicaciones y métodos de programador de entrada/salida de lectura estrangulada. Un programador de entrada/salida con estrangulación de lectura toma solicitudes de escritura para los datos capturados de una red, proporciona estos datos a un sistema que da persistencia a los datos capturados, y toma solicitudes de lectura procedentes de sistemas de usuario externos. La tasa de solicitudes de lectura y de escritura se determina al mantener dos ventanas deslizantes a lo largo de solicitudes de escritura previas, siendo la segunda ventana más larga que la primera. El programador de entrada/salida con estrangulación de lectura está configurado de tal modo que, cuando la actividad de solicitudes de escritura supera un umbral según se determina a lo largo de la primera ventana, el programador de entrada/salida con estrangulación de lectura estrangula el flujo de solicitudes de lectura. Se proporciona un medio de almacenamiento en el cual se reenvían las solicitudes de lectura y de escritura.

Con más detalle, haciendo referencia a la figura 1 se observa un dibujo conceptual que muestra una visión de conjunto de sistema de alto nivel. Se proporciona un aparato de captura de paquetes 101. El aparato de captura de paquetes 101 puede comprender un ordenador que está ejecutando un derivado del sistema operativo LINUX®. El sistema operativo informático LINUX® es un sistema operativo de código abierto que se encuentra disponible bajo una licencia al público general que es administrada por The Linux Foundation, 1796 18th Street, Suite C, San Francisco, California 94107.

Se proporciona al menos una tarjeta de captura de protocolo de comunicación de red 103 para capturar paquetes. La tarjeta de captura de protocolo de comunicación de red 103 puede ser la tarjeta DAG® facilitada por Endace USA Limited, 14425 Penrose Place, Suite 225, Chantilly, Virginia 20151. Un medio de almacenamiento 105, tal como, por ejemplo, una unidad de disco magnético o de tipo flash se proporciona para el almacenamiento, que es posible que esté organizada en una configuración de tipo agrupación redundante de discos independientes (RAID, *redundant array of independent disks*) para aumentar la fiabilidad y el rendimiento de referencia base.

Múltiples fuentes de datos, que se representan mediante la fuente de datos 107, generan y diseminan datos a lo largo de una red 109, tal como, por ejemplo, Internet. Se proporciona una derivación de red pasiva 111 para acceder a los datos que fluyen a través de la red 109. La derivación de red pasiva 111 puede ser una derivación de red Gigamon facilitada por Gigamon, 598 Gibraltar Drive, Milpitas, California 95035. La derivación de red pasiva 111 duplica un flujo de paquetes y encamina el flujo de paquetes hacia la tarjeta de captura de protocolo de comunicación de red 103. La tarjeta de captura de protocolo de comunicación de red 103 entrega los paquetes entrantes a un proceso de escritura 104 en el aparato, el cual los escribe en un medio de almacenamiento 105 por medio del programador de E/S con estrangulación de lectura 113. Estos datos se pueden consultar entonces por medio de un cliente de consulta 115. En paralelo, el tráfico de red es consumido por el destinatario original del flujo de paquetes, el destinatario de datos 117.

La presente invención es pertinente con independencia de cómo se proporcionan los datos de red transmitidos al programador de entrada/salida con estrangulación de lectura. Haciendo referencia a la figura 2, se observa un dibujo conceptual que muestra una visión de conjunto de alto nivel sin una derivación de red pasiva. Se proporciona un aparato de captura de paquetes 120. El aparato de captura de paquetes 120 puede comprender un ordenador que está ejecutando un derivado del sistema operativo LINUX®. El sistema operativo informático LINUX® es un sistema operativo de código abierto que se encuentra disponible bajo una licencia al público general que es administrada por The Linux Foundation, 1796 18th Street, Suite C, San Francisco, California 94107.

Se proporciona al menos una tarjeta de captura de protocolo de comunicación de red 122 para capturar paquetes. La tarjeta de captura de protocolo de comunicación de red 122 puede ser la tarjeta DAG® facilitada por Endace USA Limited, 14425 Penrose Place, Suite 225, Chantilly, Virginia 20151. Un medio de almacenamiento 122, tal como, por ejemplo, una unidad de disco magnético o de tipo flash se proporciona para el almacenamiento, que es posible que esté organizada en una configuración de tipo agrupación redundante de discos independientes (RAID, *redundant array of independent disks*) para aumentar la fiabilidad y el rendimiento de referencia base.

Múltiples fuentes de datos, que se representan mediante la fuente de datos 126, generan y diseminan datos a lo largo de una red 130, tal como, por ejemplo, Internet. La tarjeta de captura de protocolo de comunicación de red 134 entrega los paquetes entrantes a un proceso de escritura 136 en el aparato, el cual los escribe en un medio de almacenamiento 122 por medio del programador de E/S con estrangulación de lectura 142. Estos datos se pueden consultar entonces por medio de un cliente de consulta 146.

Haciendo referencia a la figura 3, se observa un diagrama que ilustra lo que le ocurre a las tasas de lectura y de escritura a medida que sube la tasa de captura con un programador habitual de la técnica anterior. El eje X visualiza la tasa de entrada/salida (E/S) mientras que el eje Y representa el tiempo. Un dispositivo de almacenamiento puede escribir continuamente a su tasa máxima si el dispositivo de almacenamiento está atendiendo solo solicitudes de escritura. El rendimiento máximo comienza a descender tan pronto como se introducen solicitudes de lectura, con la tasa de descenso aumentando a medida que aumenta la relación de las solicitudes de lectura con respecto a las solicitudes de escritura. Esta relación está directamente correlacionada positivamente con la variabilidad de cualquier solicitud de escritura individual y correlacionada negativamente con el rendimiento de escritura.

Cuando el rendimiento de escritura cae por debajo de la tasa de captura de datos, el dispositivo de almacenamiento comienza a generar una contrapresión, dando lugar a que se llene cualquier memoria intermedia existente. Si la diferencia entre la tasa de escritura y la tasa de captura es lo bastante alta, las memorias intermedias se llenarán por completo y, por último, la contrapresión alcanzará la tarjeta de captura de protocolo de comunicación de red, lo que la fuerza a descartar los paquetes. Por conveniencia, se puede hacer referencia, de forma no limitante, a esta diferencia como umbral de estrangulación. Este umbral de estrangulación puede ser un nivel estático y previamente determinado o se puede determinar de una forma dinámica y no previamente determinada.

Haciendo referencia de nuevo a la figura 1, se proporciona un programador de entrada/salida (E/S) con estrangulación de lectura 113. El programador de E/S con estrangulación de lectura 113 está vinculado al umbral de estrangulación. El programador de E/S con estrangulación de lectura 113 puede ser un módulo del núcleo del sistema operativo a través del cual pasan las solicitudes de lectura a partir de un proceso de lectura 115 y las solicitudes de escritura a partir de un proceso de escritura 117 de camino a los medios de almacenamiento 105. El programador de E/S con estrangulación de lectura 113 es libre de dejar pasar las solicitudes de lectura y de escritura en cualquier orden que desee el programador de E/S con estrangulación de lectura 113.

Debido a que la volatilidad de las solicitudes de lectura a partir del proceso de lectura 115 y las solicitudes de escritura a partir del proceso de escritura 117 puede ser bastante alta, en particular si el nivel se encuentra en o cerca del umbral. Con el fin de evitar la histéresis, es aconsejable alguna forma de esquema de cálculo de promedio. La histéresis es una condición en la que el programador oscila, de forma no deseable, entre los modos de estrangulación y de no estrangulación. En una realización, la tasa de captura se supervisa al mantener dos ventanas deslizantes a lo largo de solicitudes de escritura previas, siendo la segunda ventana más larga que la primera. Por conveniencia, se puede hacer referencia, de forma no limitante, a la primera ventana como ventana de ascenso, mientras que se puede hacer referencia, de forma no limitante, a la segunda ventana como ventana de retroceso.

Para entender la ventana de retroceso, imagínese una condición como mercado abierto, en la que la tasa de datos procedente de la bolsa aumenta con mucha rapidez. Esto se observa en la figura 5. A medida que la tasa cruza el umbral, el programador pasa al modo de estrangulación - poniendo en espera las lecturas. Imagínese ahora que la tasa desciende momentáneamente antes de continuar subiendo (esta es una tendencia común en escenarios cerrados y abiertos de mercado). Si ese descenso lleva la tasa por debajo del umbral, el programador saldrá del modo de estrangulación y permitirá el paso de lecturas (potencialmente una gran cantidad de lecturas contenidas); no obstante, debido a que las tasas se encuentran aún en una tendencia al alza general y muy rápida, una gran cantidad de lecturas hubieran pasado en medio de las escrituras de tasa elevada, generando riesgo de contrapresión y de descarte de paquetes. La ventana de retroceso soluciona este problema mediante el uso de una medida de histórico de la tasa para predecir la tendencia actual, que ignora tales descensos en una tendencia por lo demás al alza. El control del tamaño de la ventana permite que el usuario controle cómo de grande ha de ser un descenso para señalizar una inversión real.

Para ilustrar la ventana de ascenso, imagínese un escenario en el que la tasa de captura es bastante plana y se encuentra justo por debajo del umbral, pero presenta algunos picos de tal modo que los valles se encuentran por debajo del umbral y los picos se encuentran justo por encima. Esto se observa en la figura 6. En este escenario, el programador entraría en el modo de estrangulación en cada pico y saldría en cada valle. Si la distancia entre picos y valles es lo bastante pequeña, la tasa de conmutación evitaría que tuviera lugar lectura significativa alguna. El uso de una ventana de ascenso permite que el programador diferencie entre las situaciones en las que la tasa simplemente está oscilando en torno al umbral y en las que la tasa ha roto en la práctica el umbral.

La ventana de ascenso es más pequeña que la ventana de retroceso debido a que es necesario que el valor promedio de la tasa de E/S cambie más rápido cuando se asciende en contraposición al descenso, en donde la tasa promedio de E/S en la ventana debería reducirse más lentamente. Es decir, el ascenso debería estrangular de forma agresiva y el descenso debería ser más conservador en hacer que remita la estrangulación.

Haciendo referencia a la figura 4, se observa un diagrama que ilustra lo que le ocurre a las tasas de lectura y de escritura a medida que sube la tasa de captura utilizando el programador de E/S con estrangulación de lectura 113. Cuando la tasa de captura tal como se determina a lo largo de las ventanas de ascenso supera el umbral de estrangulación, el programador de E/S con estrangulación de lectura 113 deja de atender solicitudes de lectura. Las solicitudes de lectura existentes y entrantes se bloquean durante este periodo. Durante este periodo, los medios de almacenamiento 105 operarán al rendimiento de escritura máximo. Cuando la tasa de captura según se computa a lo largo de la ventana de retroceso cae por debajo del umbral de captura, el programador de E/S con estrangulación de lectura 113 sale del modo de estrangulación y reanuda la atención a solicitudes de lectura.

Hacer corta la ventana de retroceso permite que el programador de E/S con estrangulación de lectura 113 entre en el modo de estrangulación con rapidez si la tasa de captura se dispara con rapidez. De forma similar, hacer más larga la ventana de retroceso permite que el programador de E/S con estrangulación de lectura 113 evite casos de histéresis en los que la tasa de captura puede descender por debajo del umbral antes de continuar ascendiendo. Mediante el aumento o la disminución del umbral de estrangulación, y las longitudes de las ventanas de ascenso y

de retroceso, se puede dar un ajuste fino al comportamiento exacto del programador de E/S con estrangulación de lectura 113.

Las técnicas que se describen en el presente documento no están inherentemente relacionadas con cualquier soporte físico particular u otro aparato. En determinados aspectos, las técnicas que se describen en el presente documento se pueden implementar usando soporte físico o una combinación de soporte lógico y soporte físico, o bien en un servidor dedicado, o bien integrado en otra entidad, o bien distribuido a través de múltiples entidades.

Además de en soporte físico, las técnicas que se describen en el presente documento se pueden implementar usando un código que crea un entorno de ejecución. El código puede constituir un soporte lógico inalterable de procesador, una pila de protocolos, un sistema de gestión de bases de datos, un sistema operativo, o una combinación de uno o más de los mismos que estén almacenados en memoria. La implementación que se describe en el presente documento no se limita a lenguaje de programación particular alguno.

A pesar de que la presente memoria descriptiva contiene muchos detalles específicos, estos no se deberían interpretar como limitaciones al alcance de lo que se puede reivindicar, sino más bien como descripciones de implementaciones particulares de la materia objeto. Determinadas características que se describen en la presente memoria descriptiva en el contexto de implementaciones separadas también se pueden implementar en combinación en una única implementación. A la inversa, diversas características que se describen en el contexto de una única implementación también se pueden implementar en múltiples implementaciones por separado o en cualquier subcombinación adecuada. Además, a pesar de que en lo que antecede se puedan describir características como que actúan en determinadas combinaciones e incluso se puedan reivindicar inicialmente en ese sentido, una o más características procedentes de una combinación reivindicada pueden, en algunos casos, eliminarse de la combinación, y la combinación reivindicada se puede dirigir a una subcombinación o a una variación de una subcombinación.

De forma similar, a pesar de que las operaciones se muestran en los dibujos en un orden particular, esto no se debería entender como que se requiera que tales operaciones se realicen en el orden particular que se muestra, o en orden secuencial, o que se realice la totalidad de las operaciones ilustradas, para lograr unos resultados deseables. En determinadas circunstancias, puede ser ventajoso un procesamiento multitarea y en paralelo. Además, la separación de diversos componentes de sistema en las técnicas que se describen en el presente documento no se debería entender como que se requiera tal separación en todos los aspectos.

A pesar de que la materia objeto se ha descrito con una implementación específica, otras alternativas, modificaciones y variaciones serán evidentes a los expertos en la materia. Por consiguiente, la divulgación tiene por objeto ser ilustrativa, pero no limitante, y la totalidad de tales alternativas, modificaciones y variaciones se encuentran dentro del alcance de las siguientes reivindicaciones.

REIVINDICACIONES

1. Uno o más dispositivos informáticos (101) que están configurados para leer y escribir datos en memoria, comprendiendo los uno o más dispositivos informáticos (101):

un programador de entrada/salida con estrangulación de lectura (113) que está configurado para tomar solicitudes de escritura para los datos capturados de una red (109), para proporcionar estos datos a un sistema (105) que da persistencia a los datos capturados, y para tomar solicitudes de lectura procedentes de sistemas de usuario externos;

el programador de entrada/salida con estrangulación de lectura (113) está configurado adicionalmente para determinar una tasa de captura al mantener dos ventanas deslizantes a lo largo de solicitudes de escritura previas, siendo la segunda ventana más larga que la primera;

el programador de entrada/salida con estrangulación de lectura (113) configurado adicionalmente de tal modo que:

cuando la actividad de solicitudes de escritura supera un umbral según se determina a lo largo de la primera ventana, el programador de entrada/salida con estrangulación de lectura (113) deja de atender solicitudes de lectura; y

cuando la actividad de solicitudes de lectura y de escritura cae por debajo del umbral según se computa a lo largo de la segunda ventana, el programador de entrada/salida con estrangulación de lectura (113) reanuda la atención a solicitudes de lectura

un medio de almacenamiento (105) en el cual se reenvían las solicitudes de lectura y de escritura.

2. Los uno o más dispositivos informáticos (101) según la reivindicación 1, que comprenden adicionalmente una tarjeta de captura de protocolo de comunicación de red (103) que está configurada para capturar datos que están diseminados a lo largo de una red (109) y proporcionar estos datos al programador de entrada/salida con estrangulación de lectura (113).

3. Los uno o más dispositivos informáticos (101) según la reivindicación 2, que comprenden adicionalmente una derivación de red pasiva (111) que está configurada para acceder a datos que fluyen a través de una red (109), y para paquetizar y encaminar los datos hacia la tarjeta de captura de protocolo de comunicación de red (103).

4. Los uno o más dispositivos informáticos (101) según la reivindicación 1, en los que el programador de entrada/salida con estrangulación de lectura (113) está configurado adicionalmente de tal modo que, cuando la actividad de solicitudes de lectura y de escritura supera el umbral, el programador de entrada/salida con estrangulación de lectura (113) bloquea las solicitudes de lectura existentes y entrantes.

5. Los uno o más dispositivos informáticos (101) según la reivindicación 1, en los que el programador de entrada/salida con estrangulación de lectura (113) está configurado adicionalmente de tal modo que, cuando el rendimiento de escritura cae por debajo de la tasa de captura de datos, el programador de entrada/salida con estrangulación de lectura (113) altera el flujo de solicitudes de lectura.

6. Los uno o más dispositivos informáticos (101) según la reivindicación 1, en los que el umbral o bien está previamente establecido o bien se determina de forma dinámica.

7. Los uno o más dispositivos informáticos (101) según la reivindicación 1, en los que el medio de almacenamiento (105) se selecciona de entre el grupo que consiste en unidades de disco magnético, unidades de tipo flash, y combinaciones de las mismas.

8. Los uno o más dispositivos informáticos (101) según la reivindicación 1, en los que el medio de almacenamiento (105) comprende una agrupación redundante de discos independientes.

9. Un método implementado por uno o más dispositivos informáticos (101) que están configurados para programar datos de lectura y de escritura en memoria, comprendiendo el método:

recibir solicitudes de lectura y de escritura;

determinar una tasa de solicitudes de lectura y de escritura al mantener dos ventanas deslizantes a lo largo de solicitudes de escritura previas, siendo la segunda ventana más larga que la primera;

cuando la actividad de solicitudes de escritura supera el umbral según se determina a lo largo de la primera ventana, dejar de atender solicitudes de lectura; y

reenviar la solicitud de lectura y de escritura a un medio de almacenamiento (105); y

cuando la actividad de solicitudes de lectura y de escritura cae por debajo del umbral según se computa a lo largo de la segunda ventana, reanudar la atención a solicitudes de lectura.

10. El método según la reivindicación 9, que comprende adicionalmente al menos una de las siguientes etapas:

(i) cuando la actividad de solicitudes de escritura supera el umbral, bloquear las solicitudes de lectura existentes y entrantes; y

(ii) cuando la actividad de solicitudes de escritura cae por debajo del umbral, reanudar la atención a solicitudes de lectura.

11. El método según la reivindicación 9, que comprende adicionalmente, cuando el rendimiento de escritura cae por debajo de la tasa de captura de datos, alterar el flujo de solicitudes de lectura.

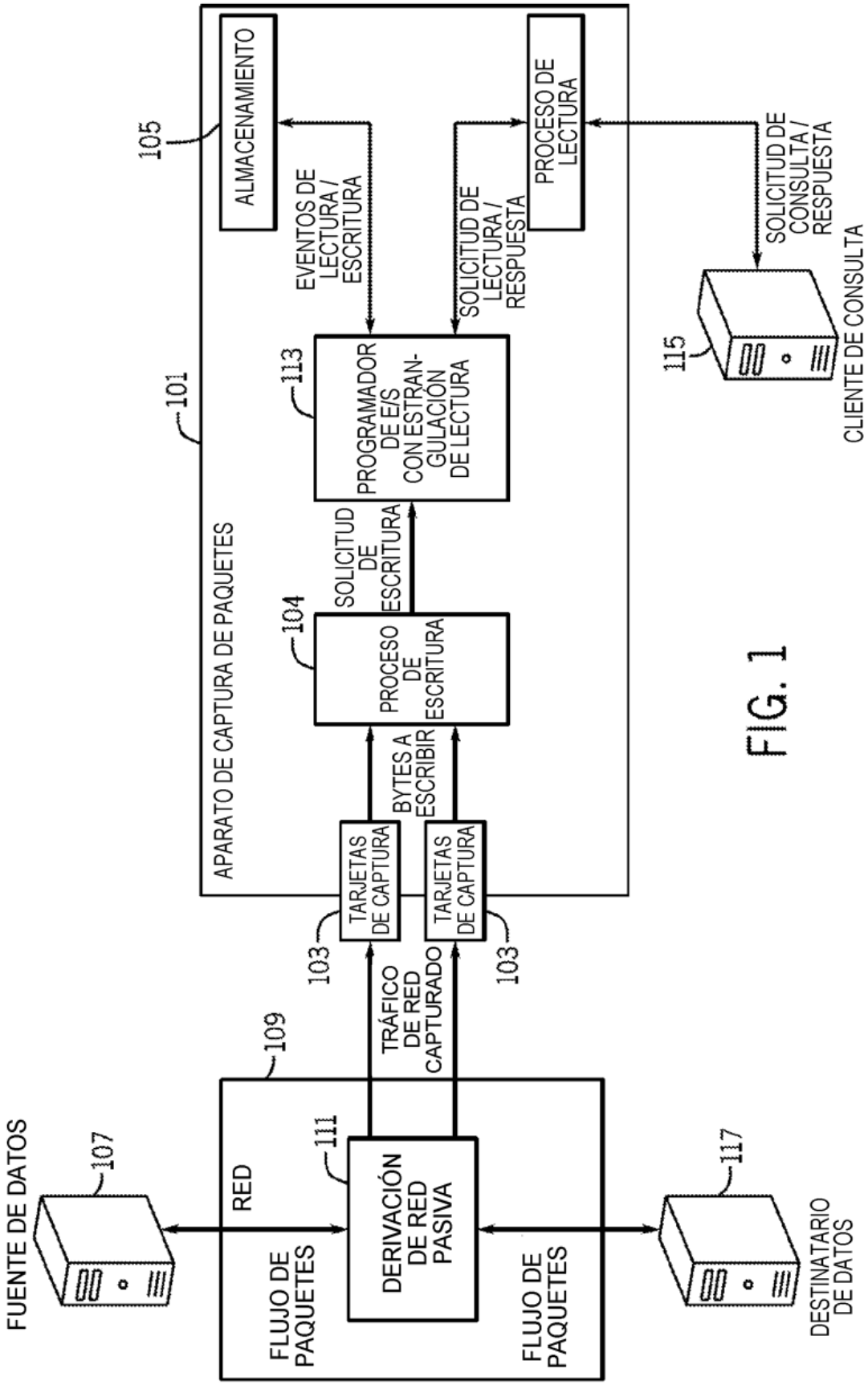
FUENTE DE DATOS
107
RED
111
DERIVACIÓN DE RED PASIVA
117
DESTINATARIO DE DATOS
109
FLUJO DE PAQUETES
FLUJO DE PAQUETES
TRÁFICO DE RED CAPTURADO
103
TARJETAS DE CAPTURA
103
TARJETAS DE CAPTURA
BYTES A ESCRIBIR
104
PROCESO DE ESCRITURA
SOLICITUD DE ESCRITURA
101
APARATO DE CAPTURA DE PAQUETES
113
PROGRAMADOR DE E/S CON ESTRANGULACIÓN DE LECTURA
105
ALMACENAMIENTO
EVENTOS DE LECTURA / ESCRITURA
SOLICITUD DE LECTURA / RESPUESTA
PROCESO DE LECTURA
115
CLIENTE DE CONSULTA
SOLICITUD DE CONSULTA / RESPUESTA

FIG. 1

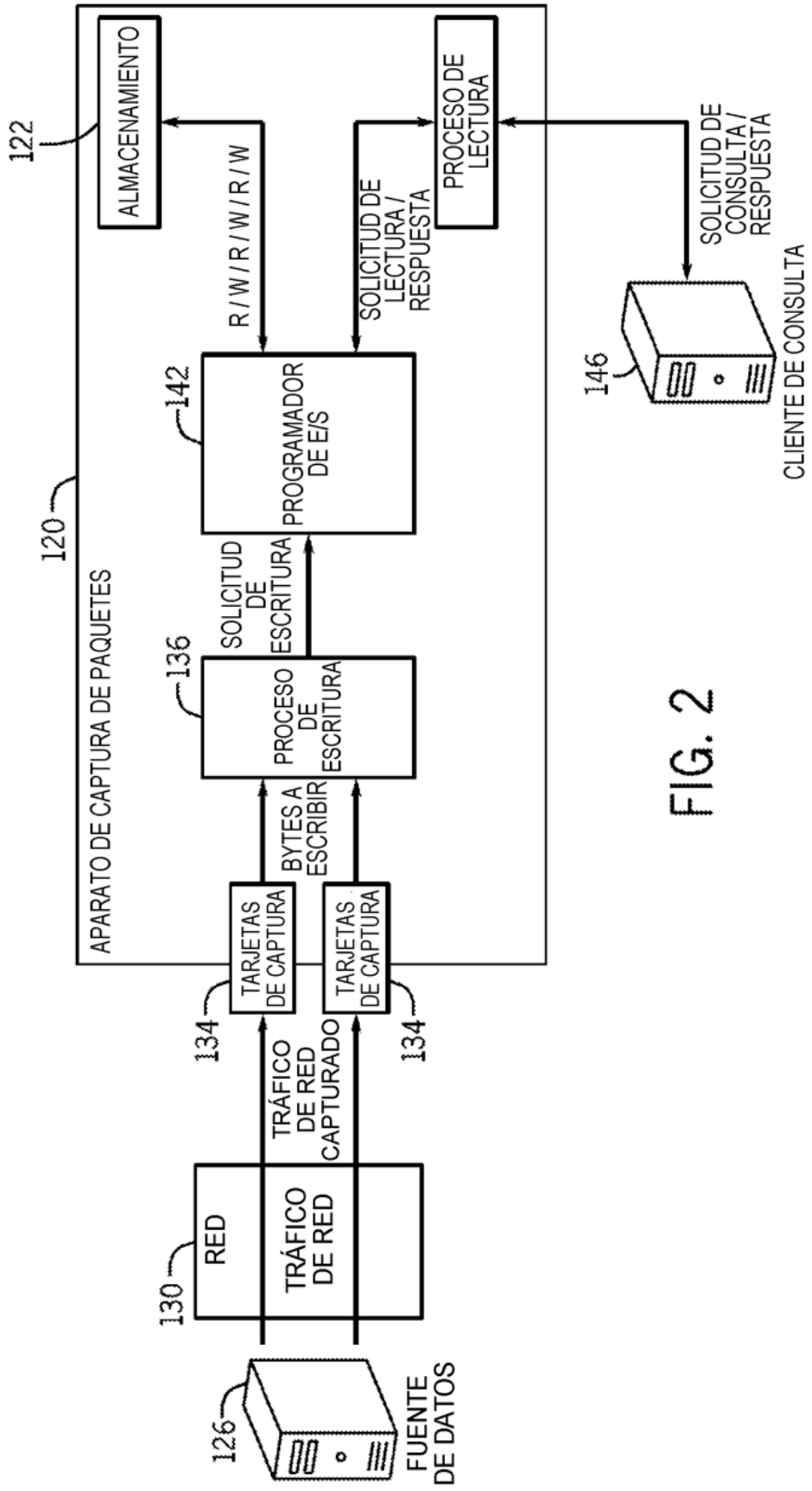
FUENTE DE DATOS
126
RED
130
TRÁFICO DE RED
TRÁFICO DE RED CAPTURADO
TARJETAS DE CAPTURA
134
TARJETAS DE CAPTURA
134
BYTES A ESCRIBIR
PROCESO DE ESCRITURA
136
SOLICITUD DE ESCRITURA
PROGRAMADOR DE E/S
142
APARATO DE CAPTURA DE PAQUETES
120
ALMACENAMIENTO
122
R / W / R / W / R / W
SOLICITUD DE LECTURA / RESPUESTA
PROCESO DE LECTURA
SOLICITUD DE CONSULTA / RESPUESTA
CLIENTE DE CONSULTA
146

FIG. 2

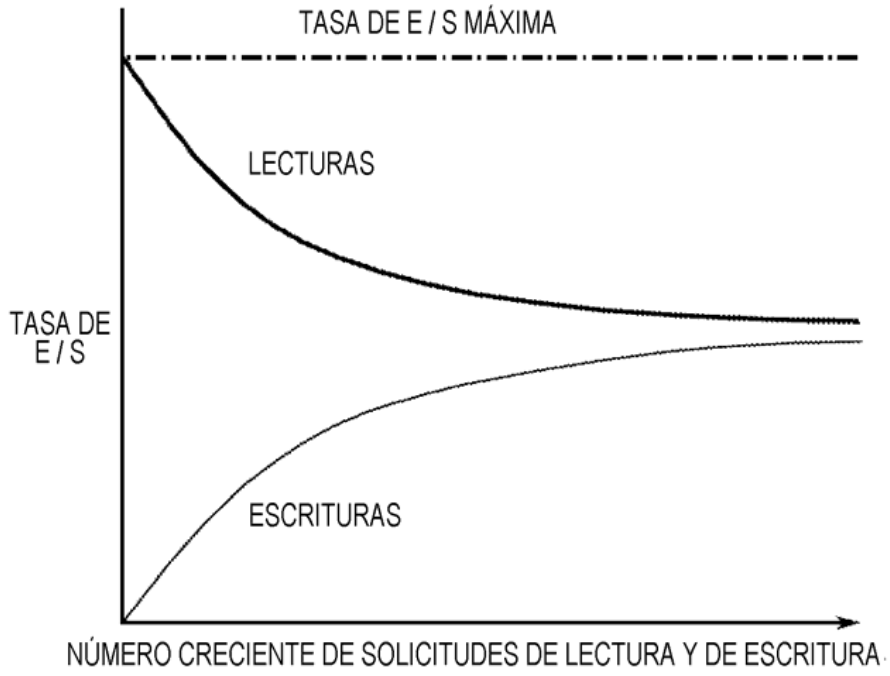
TASA DE E / S MÁXIMA
LECTURAS
TASA DE E / S
ESCRITURAS
NÚMERO CRECIENTE DE SOLICITUDES DE LECTURA Y DE ESCRITURA

FIG. 3

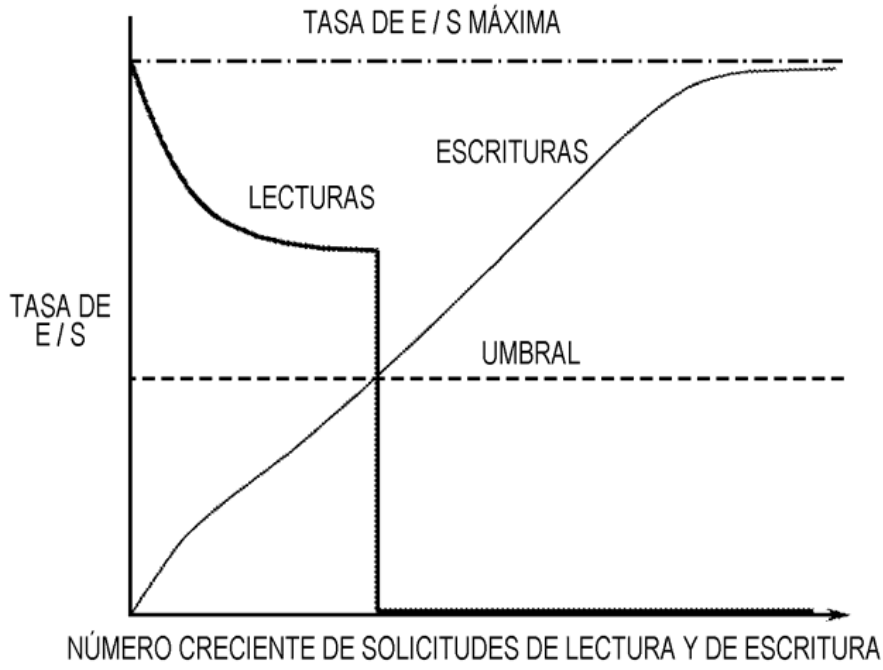
TASA DE E / S MÁXIMA
ESCRITURAS
LECTURAS
TASA DE E / S
UMBRAL
NÚMERO CRECIENTE DE SOLICITUDES DE LECTURA Y DE ESCRITURA

FIG. 4

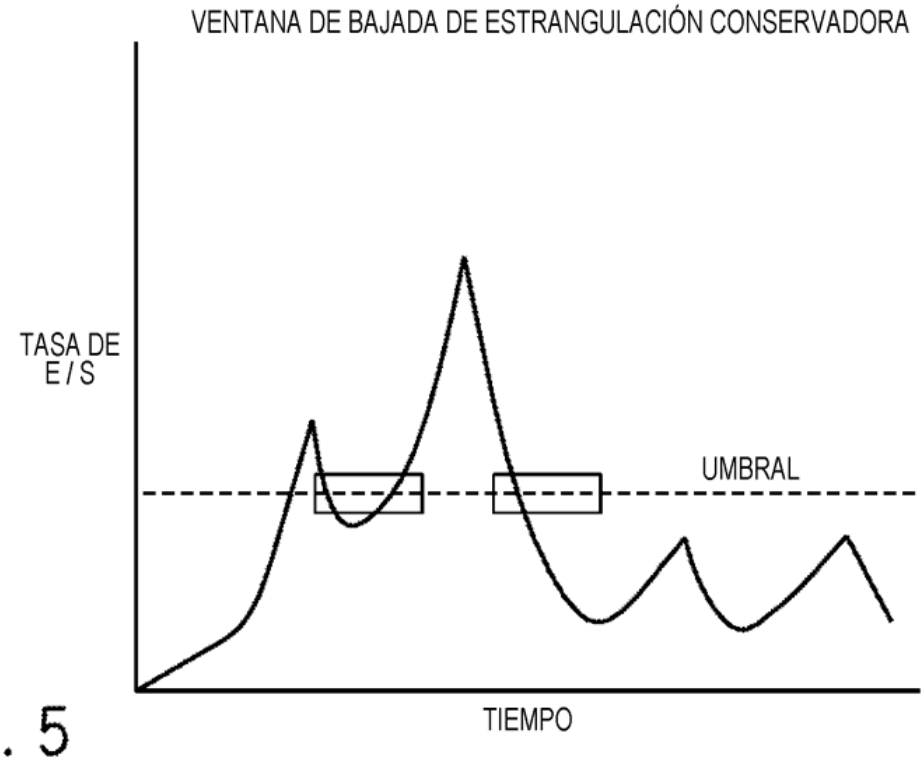
VENTANA DE BAJADA DE ESTRANGULACIÓN CONSERVADORA
TASA DE E / S
UMBRAL
TIEMPO

FIG. 5

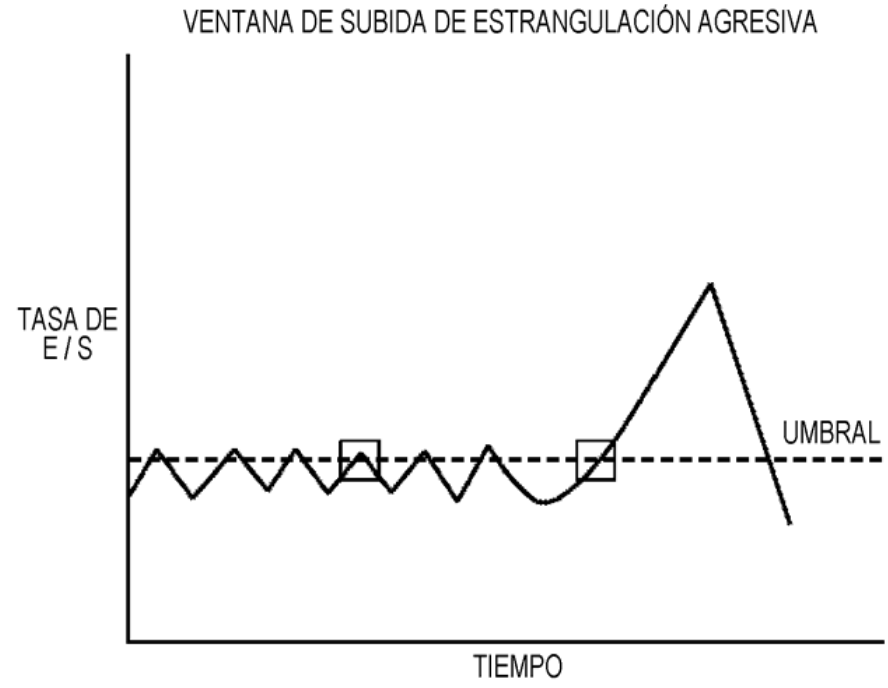
VENTANA DE SUBIDA DE ESTRANGULACIÓN AGRESIVA
TASA DE E / S
UMBRAL
TIEMPO

FIG. 6