(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 614 751**

(51) Int. CI.:

H04L 29/08 (2006.01)
G06F 9/50 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **07.04.2006 PCT/US2006/013677**

(87) Fecha y número de publicación internacional: **12.10.2006 WO06108187**

(96) Fecha de presentación y número de la solicitud europea: **07.04.2006 E 06749901 (2)**

(97) Fecha y número de publicación de la concesión europea: **07.12.2016 EP 1872249**

(54) Título: **Acceso bajo demanda a recursos informáticos**

(30) Prioridad:

07.04.2005 US 669278 P

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

(73) Titular/es:

III HOLDINGS 12, LLC (100.0%)
2711 Centerville Road, Suite 400
Wilmington, DE 19808, US

(72) Inventor/es:

JACKSON, DAVID, BRIAN

(74) Agente/Representante:

SÁEZ MAESO, Ana

DESCRIPCIÓN

Acceso bajo demanda a recursos informáticos.

Aviso sobre derechos de autor (Copyright)

ANTECEDENTES DE LA INVENCIÓN

1. Ámbito de la invención

La presente invención se refiere a un entorno informático bajo demanda y, más específicamente, a un sistema y un método para proporcionar acceso y uso de recursos informáticos bajo demanda desde un entorno informático local.

2. Introducción

Los administradores de clústeres desean un retorno máximo de la inversión, lo que a menudo significa una elevada utilización del sistema y la capacidad de proporcionar diversas calidades de servicio a varios usuarios y grupos. Un clúster se define normalmente como un ordenador paralelo formado por componentes de productos y se ejecuta como su software de productos del software del sistema. Un clúster contiene nodos que contienen uno o más procesadores, memoria compartida por todos los procesadores en el nodo respectivo y dispositivos periféricos adicionales tales como discos de almacenamiento conectados por una red que permite que los datos se muevan entre nodos. Un clúster es un ejemplo de un entorno informático. Otros ejemplos incluyen una cuadrícula, que se define vagamente como un grupo de clústeres, y una granja informática que es otra organización del ordenador para el procesamiento.

A menudo, un conjunto de recursos organizados en un clúster o una cuadrícula puede tener trabajos que se someterán a los recursos que requieren más capacidad que el conjunto de recursos disponible. A este respecto, existe una necesidad en la técnica para poder utilizar de forma fácil, eficaz y bajo petición nuevos recursos o diferentes recursos para gestionar un trabajo. El concepto de recursos informáticos "bajo demanda" se ha desarrollado en la comunidad informática de alto rendimiento recientemente. Un entorno informático bajo demanda permite a las empresas adquirir potencia informática para la demanda media y luego contratar la potencia de procesamiento para ayudar en las cargas máximas o descargar todas sus necesidades informáticas a una instalación remota.

Habilitar la capacidad bajo demanda de una manera fácil de usar es importante para aumentar la omnipresencia del alojamiento en un entorno informático bajo demanda, tal como un entorno informático de alto rendimiento o un entorno de centro de datos. Varias entidades pueden proporcionar una versión de la capacidad bajo demanda que ahí existe multi-hora o de múltiples retrasos en la obtención de acceso al entorno. El retraso se debe a la inflexibilidad de la transferencia de la carga de trabajo porque los centros bajo demanda requieren que las partes participantes se alineen a ciertos hardwares, sistemas operativos o entornos de administrador de recursos. Estos requisitos actúan como inhibidores para amplia adopción del uso de centros bajo demanda y lo hacen demasiado onerosos para que los clientes potenciales prueben el servicio.

Los usuarios deben pagar los cargos y costes no deseados o inesperados para hacer los cambios de infraestructura para su compatibilidad con los centros bajo demanda.

A menudo, un conjunto de recursos organizados en un clúster o una cuadrícula puede tener trabajos para enviarlos a los recursos que requieren más capacidad que el conjunto del recurso tiene disponible. A este respecto, en la técnica existe una necesidad de poder utilizar de forma fácil, eficaz y bajo petición nuevos recursos o diferentes recursos para manejar un trabajo. El concepto de recursos informáticos "bajo demanda" se ha desarrollado en la comunidad informática de alto rendimiento recientemente. Un entorno informático bajo demanda permite a las empresas adquirir potencia informática para la demanda media y luego contratar potencia de procesamiento para ayudar en las cargas máximas o para descargar todas sus necesidades informáticas a una instalación remota. Varios libros de referencia con material relacionado con la informática bajo demanda o la informática de utilidad son Mike Ault, Madhu Tumma, Oracle 10g Grid & Real Application

Clusters, Rampant TechPress, 2004 and Guy Bunker, Darren Thomson, Delivering Utility Computing Business-driven IT Optimization, John Wiley & Sons Ltd, 2006.

En Bunker and Thompson, sección 3.3 en la página 32 titulada "Connectivity: The Great Enabler" donde se analiza cómo la interconexión de ordenadores aumentará drásticamente su uso. Esta divulgación aborda esa cuestión. Sheets et al., Patente de Estados Unidos Nº 6.816.905, describe un enfoque de funcionamiento de un proveedor de servicios alojados para Internet de tal manera que proporcione una gestión dinámica de servicios de alojamiento a través de cuentas de clientes dispares o sitios geográficamente distintos. Su enfoque implica que al menos un primer grupo administrativo del servidor se reasigna automática y dinámicamente a un segundo grupo administrativo en respuesta a la supervisión automática. Sheets et al. no indica nada sobre un centro informático independiente de alto rendimiento bajo demanda. La publicación de PCT WO 01/14987 ofrece un enfoque para establecer una cuadrícula informática que pueda construirse físicamente una vez y luego se divida de forma lógica para varias organizaciones bajo demanda. Sin embargo, muestran un mecanismo de supervisión o un Plano de Control que controla los elementos informáticos, de redes y de almacenamiento de la cuadrícula informática a través de puertos de control o interfaces. Esta referencia tampoco contempla el uso de un entorno informático independiente bajo demanda al requerir que este Plano de Control administre la cuadrícula informática indicada. En la técnica existe una necesidad de soluciones mejoradas que permitan la comunicación y la conectividad con un centro informático de alto rendimiento bajo demanda. La publicación de PCT WO 2006/112981, que es una técnica anterior según el artículo 54(3)EPC, propone un método de gestión de recursos entre un entorno informático local y un entorno informático bajo demanda que comprende detectar un evento en un entorno informático local y, basándose en el evento detectado, establecer automáticamente la comunicación con un entorno informático bajo demanda, aprovisionando recursos dentro del entorno informático bajo demanda y transferir la carga de trabajo desde el entorno informático local de forma transparente al entorno informático bajo demanda.

RESUMEN DE LA INVENCIÓN

En la siguiente descripción, se muestran características adicionales y ventajas de la invención, y en parte serán evidentes a partir de la descripción, o pueden aprenderse mediante la práctica de la invención. Las características y ventajas de la invención pueden realizarse y obtenerse por medio de instrumentos y combinaciones señalados concretamente en las reivindicaciones adjuntas. Estas y otras características de la presente invención se harán más evidentes con la siguiente descripción y las reivindicaciones adjuntas o pueden aprenderse mediante la práctica de la invención tal y como aquí se expone.

Un aspecto de la invención proporciona un método de gestión de recursos en un entorno informático local y un entorno informático bajo demanda. El método incluye la detección de un evento asociado con un entorno informático local que comprende un primer conjunto de nodos informáticos y gestionados por un primer módulo de gestión, en el que el método se caracteriza por: basándose en el evento detectado, identificar información sobre el entorno informático local; establecer comunicación con un entorno informático bajo demanda que comprenda un segundo conjunto de nodos informáticos diferentes del primer conjunto de nodos informáticos y administrado por un segundo módulo de gestión diferente del primer módulo de gestión y transmitir la información sobre el entorno informático local al entorno informático bajo demanda; seleccionar un perfil de una pluralidad de perfiles, el perfil relacionado con la carga de trabajo que pueda procesarse, para producir un perfil seleccionado; transferir el perfil seleccionado al entorno informático bajo demanda; aprovisionar recursos según lo indicado por el segundo módulo de gestión dentro del entorno informático bajo demanda basado en el perfil seleccionado, para proporcionar recursos aprovisionados; y transferir la carga de trabajo desde el entorno informático local) al entorno informático bajo demanda en el que la carga de trabajo transferida consuma los recursos aprovisionados.

El paso de generar al menos un perfil asociado con la carga de trabajo que se puede procesar en un entorno informático puede realizarse antes de recibir peticiones de trabajo en el entorno informático local. También, generar al menos un perfil asociado con la carga de trabajo que se puede procesar en un entorno informático que pueda realizarse de forma dinámica, como solicitudes de trabajo que se reciben en el entorno informático local. Puede haber uno o más perfiles generados. Además, uno o más de pasos del método pueden realizarse después de una operación de un usuario o un administrador, como una operación de un clic. Cualquier perfil de los generados, al menos uno, puede estar relacionado con la configuración de recursos que son diferentes de los recursos disponibles dentro del entorno informático local.

Otro aspecto de la invención proporciona un medio legible por ordenador que almacena instrucciones para controlar un dispositivo informático para administrar flujo de trabajo entre un entorno informático local y un

entorno informático bajo demanda, en el que la mejora se caracteriza por: generar al menos un perfil asociado con una carga de trabajo que se puede procesar en un entorno informático que comprende un primer conjunto de nodos informáticos administrados por un primer módulo de gestión; seleccionar en el entorno informático local un perfil de al menos un perfil para producir un perfil seleccionado, en el que el perfil seleccionado está asociado con la carga de trabajo que se puede procesar; comunicar el perfil seleccionado desde el entorno informático local al entorno informático bajo demanda, el entorno informático bajo demanda incluye un segundo conjunto de nodos informáticos diferentes del primer conjunto de nodos informáticos y gestionados por un segundo módulo de gestión diferente del primer módulo de gestión; aprovisionar recursos dentro del entorno informático bajo demanda de acuerdo con el perfil seleccionado; y transferir la carga de trabajo desde el entorno informático local al entorno informático bajo demanda.

Otro aspecto de la invención proporciona un aparato informático local que comprende un primer conjunto de nodos informáticos administrados por un primer medio de administración, estando los primeros medios de gestión adaptados para administrar la integración de un entorno informático bajo demanda en el aparato informático local, con la característica de que en los primeros medios de administración están adaptados para gestionar la integración por medio de: determinar si existe una condición de carga de trabajo pendiente en el entorno informático local; si es así, analizar la carga de trabajo pendiente para obtener un análisis; comunicar información asociada con el análisis al aparato informático bajo demanda, comprendiendo el aparato informático bajo demanda un segundo conjunto de nodos informáticos diferentes del primer conjunto de nodos informáticos y administrados por un segundo medio de administración independiente; dirigir los segundos medios de administración para aprovisionar al aparato informático bajo demanda de acuerdo con el análisis; y transferir la carga de trabajo pendiente al aparato informático bajo demanda aprovisionado.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

Con el fin de describir la manera en que se pueden obtener las ventajas y características anteriormente descritas y otras de la invención, se hará una descripción más concreta de la invención brevemente descrita anteriormente haciendo referencia a realizaciones específicas de la misma ilustradas en los documentos y dibujos adjuntos. Hay que entender que estos dibujos representan sólo realizaciones típicas de la invención y que, por tanto, no deben considerarse limitativas de su ámbito, la invención se describirá y explicará con especificidad y detalle adicionales mediante el uso de los dibujos adjuntos.

La FIG. 1 ilustra la disposición básica de la presente invención;

La FIG. 2 ilustra los componentes de hardware básicos de acuerdo con una realización de la invención; y

La FIG. 3 ilustra un ejemplo de interfaz gráfica para su uso en la obtención de recursos bajo demanda;

La FIG. 4 ilustra la optimización del almacenamiento inteligente de datos;

La FIG. 5 ilustra varios componentes de la informática basada en la utilidad;

La FIG. 6 ilustra los tipos de cuadrícula;

La FIG. 7 ilustra las combinaciones de relaciones de la cuadrícula;

La FIG. 8 ilustra gráficamente un aspecto servidor web de la descripción; y

La FIG. 9 ilustra un aspecto del método de la descripción.

DESCRIPCIÓN DETALLADA DE LA INVENCIÓN

Varias realizaciones se describen de forma detallada más abajo. Al tiempo que analizan implementaciones específicas, debe entenderse que esto se hace sólo con fines ilustrativos.

Para que los centros de alojamiento obtengan la máxima ventaja, deben simplificar la experiencia de los clientes potenciales y permitir un control detallado sobre el intercambio de recursos y también ajustar de forma dinámica lo que se proporciona en función de las necesidades de cada cliente. El control de inteligencia adicional optimiza la entrega de recursos para que los centros de alojamiento puedan reducir los costes y proporcionar ofertas competitivas que sean más fáciles de adoptar y utilizar.

Esta descripción se refiere al acceso y administración de recursos informáticos bajo demanda o de utilidad en un centro de alojamiento. La FIG. 1 ilustra la disposición básica y la interacción entre un entorno informático local 104 y un centro de alojamiento bajo demanda 102. El entorno informático local puede comprender un clúster, una cuadrícula o cualquier otra variación en estos tipos de nodos múltiples y entornos gestionados comúnmente. El centro de alojamiento bajo demanda o el entorno informático bajo demanda 102 comprende una pluralidad de nodos que están disponibles para el aprovisionamiento y, preferiblemente, tienen un nodo dedicado que contiene un maestro de alojamiento 128 que puede comprender un módulo de gestión esclavo 106 y/o al menos otro módulo como el administrador de entidades 129 y el dispositivo de provisión de nodos 118.

A lo largo de la descripción, los términos software, administrador de carga de trabajo (WM), módulo de gestión, sistema y así sucesivamente pueden usarse para referirse de manera general al software que ejecuta funciones similares a uno o más de los productos Moab™ de Cluster Resources, Inc., pero ciertamente no se limitan a la implementación exacta de Moab™ (por ejemplo, Moab Workload Manager®, Moab Grid Monitor®, etc.). Generalmente, el término "WM" puede usarse para referirse al software que realiza los pasos que se están analizando. Dicho software proporciona un servicio para la optimización de un entorno informático local y de acuerdo con los principios de la invención también puede usarse para controlar el acceso a recursos bajo demanda. En términos de control de entorno local, el software proporciona un análisis en recursos locales sobre cómo y cuándo, como dispositivos de hardware y software, se están utilizando para fines de anulación, planificación, auditoría, solución de problemas e informes internos o externos. Dicha optimización permite ajustar el entorno local para aprovechar al máximo los recursos en el entorno informático local. Sin embargo, hay momentos en los que se necesitan más recursos que los que están disponibles en el entorno local. Aquí es donde el centro bajo demanda o de alojamiento puede proporcionar recursos adicionales.

Normalmente, un centro de alojamiento 102 tendrá los siguientes atributos. Permite a una organización proporcionar recursos o servicios a clientes donde los recursos o servicios se adaptan a las necesidades del cliente. Respaldar la verdadera informática de utilidad requiere generalmente crear un centro de alojamiento 102 con una o más capacidades de la manera siguiente: acceso remoto seguro, disponibilidad de recursos garantizada en un tiempo fijo o una serie de veces, servicios integrados de auditoría/contabilidad/facturación, nivel de servicio escalonado (QOS/SLA), aprovisionamiento dinámico de nodos de ejecución, administración completa del entorno sobre recursos informáticos, de red, almacenamiento y basados en aplicaciones/servicios, optimización inteligente de la carga de trabajo, alta disponibilidad, recuperación de fallos y reasignación automatizada.

Un módulo de gestión 108 habilita la informática de utilidad permitiendo que los recursos informáticos se reserven, asignen y aprovisionen dinámicamente para satisfacer las necesidades de carga de trabajo interna o externa. Por lo tanto, en los tiempos de carga máxima de trabajo o basados en algunos otros criterios, el entorno informático local no necesita ser construido con el uso máximo en mente. A medida que se requieren recursos máximos periódicos, los desencadenantes pueden provocar un desbordamiento en el entorno bajo demanda y así ahorrar dinero al cliente. El módulo 108 es capaz de responder a solicitudes manuales o generadas automáticamente y puede garantizar la disponibilidad de recursos sujeta a los acuerdos existentes de acuerdo de nivel de servicio (SLA) o de calidad de servicio (QOS). Como ejemplo, la FIG. 1 muestra un usuario 110 que envía un trabajo o una consulta al clúster o entorno local 104. El entorno local suele ser un clúster o una cuadrícula con carga de trabajo local. Pueden enviarse trabajos que tengan requisitos explícitos de recursos. La carga de trabajo puede tener requisitos explícitos. El entorno local 104 tendrá diversos atributos tales como sistemas operativos, arquitectura, tipos de red, aplicaciones, software, capacidades de ancho de banda, etc., que el trabajo espera implícitamente. En otras palabras, los trabajos normalmente esperan que el entorno local tenga ciertos atributos que le permitan consumir los recursos de una manera esperada. Estos atributos

esperados pueden duplicarse en un entorno bajo demanda o sustituir recursos (que pueden ser una mejora o menos óptima) pueden darse en el entorno bajo demanda.

Otro software se muestra a modo de ejemplo en un administrador de recursos distribuidos como Torque 128 y varios nodos 130, 132 y 134. Los módulos de administración (maestro y/o esclavo) pueden interactuar y funcionar con cualquier gestor de recursos, como Torque, LSF, SGE, PBS y LoadLeveler y son independientes en este sentido. Los expertos en la técnica reconocerán estos diferentes paquetes de software de administrador de recursos distribuidos.

Un módulo de gestión de alojamiento o maestro de alojamiento 106 puede ser también una instancia de un producto de software Moab™ con capacidades de centro de alojamiento para permitir a una organización controlar dinámicamente recursos de red, informáticos, de aplicación y de almacenamiento y proporcionar dinámicamente sistemas operativos, de seguridad, credenciales y otros aspectos de un entorno completo informático de extremo a extremo. El Módulo 106 es responsable de conocer todas las políticas, garantías, promesas y también de administrar el aprovisionamiento de recursos dentro del espacio informático de utilidad 102- En un sentido, el módulo 106 puede ser denominado el módulo "maestro" porque se acopla y necesita conocer toda la información asociada tanto con el entorno de la utilidad y el entorno local. Sin embargo, en otro sentido, puede denominarse el módulo esclavo o intermediario de aprovisionamiento, que toma las instrucciones del módulo de gestión de clientes 108 para aprovisionar recursos y construir cualquier entorno que se solicite en el centro bajo demanda 102. Un módulo esclavo no tendría ninguna de sus propias políticas locales, sino que sigue todas las solicitudes de otro módulo de gestión. Por ejemplo, cuando el módulo 106 es el módulo esclavo, entonces un módulo maestro 108 presentaría solicitudes automatizadas o manuales (a través de un administrador o usuario) que el módulo esclavo 106 simplemente sigue para administrar la construcción del entorno solicitado. Así, tanto para los TI como para los usuarios finales, una sola interfaz fácilmente utilizable puede aumentar la eficiencia, reducir los costes, incluidos los costes de administración y mejorar las inversiones en el entorno del cliente local. La interfaz con el entorno local que también tiene acceso al entorno puede ser una interfaz web o un portal de acceso también. Sólo pueden existir restricciones de factibilidad. El módulo de cliente 108 tendría derechos y propiedad de todos los recursos. Los recursos asignados no serían compartidos sino dedicados al solicitante. Dado que el módulo esclavo 106 sigue todas las órdenes desde el módulo maestro 108, cualquier restricción de política ocurrirá preferiblemente en el módulo maestro 108 en el entorno local.

Los módulos también proporcionan servicios de administración de datos que simplifican la adición de recursos desde un entorno local. Por ejemplo, si el entorno local comprende una red de área amplia, el módulo de gestión 108 proporciona un modelo de seguridad que asegura, cuando el entorno lo exige, que los administradores puedan confiar en el sistema incluso cuando los recursos no confiables a cierto nivel se hayan añadido al entorno local o al entorno bajo demanda. Además, los módulos de administración cumplen con las arquitecturas basadas en servicios web de n-niveles y, por lo tanto, la escalabilidad y la generación de informes son partes inherentes del sistema. Un sistema operativo según los principios expuestos en la presente memoria también tiene la capacidad de rastrear, registrar y archivar información sobre trabajos u otros procesos que se han ejecutado en el sistema.

Un centro de alojamiento 102 proporciona recursos dedicados programados a clientes para diversos propósitos y tiene normalmente una serie de atributos clave: acceso remoto seguro, disponibilidad garantizada de recursos a un momento fijo o series de veces, servicios de auditoría/contabilidad estrechamente integrados, niveles variables de calidad de servicio que proporcionan acceso privilegiado a un conjunto de usuarios; administración de imágenes por nodos que permiten al centro de alojamiento restaurar una imagen exacta del cliente específico antes de permitir el acceso. Los recursos disponibles para un módulo 106, que también puede denominarse intermediario de recursos de proveedores, tendrán atributos rígidos (arquitectura, RAM, espacio en disco local, etc.) y flexibles (OS, colas, aplicaciones instaladas, etc.). El proveedor o intermediario de recursos bajo demanda 106 puede proporcionar normalmente (modificar dinámicamente) atributos flexibles pero no rígidos. El intermediario proveedor 106 puede poseer recursos múltiples cada uno con tipos diferentes con atributos rígidos (es decir, nodos de un solo procesador y procesador dual, nodos Intel, nodos AMD, nodos con 512 MB de RAM, nodos con 1 GB de RAM, etc.).

Esta combinación de atributos presenta restricciones únicas en un sistema de gestión. Se describen aquí cómo los módulos de gestión 108 y 106 son capaces de gestionar, modificar y proporcionar recursos de manera eficaz en este entorno y proporcionar una gama completa de servicios además de estos recursos. Las herramientas avanzadas de reserva y gestión de políticas de los módulos de gestión proporcionan soporte para el establecimiento de acuerdos de nivel de servicio amplios, facturación automatizada y creación de gráficos e informes instantáneos.

La tecnología informática basada en utilidades permite a un centro de alojamiento 102 aprovechar rápidamente los recursos informáticos existentes, co-asignar dinámicamente los recursos y proporcionarlos automáticamente en un clúster virtual sin interrupciones. La Solicitud N° 11/276.852 de Estados Unidos describe un método para formar un clúster privado virtual. El proceso implica la agregación de recursos informáticos y el establecimiento de particiones de los recursos informáticos agregados. A continuación, el sistema presenta sólo los recursos compartidos accesibles por una organización para su uso dentro de la organización. Por lo tanto, en el centro bajo demanda, a medida que se necesitan recursos, el control y establecimiento de un entorno de carga de trabajo desde un entorno local puede producirse por medio de la creación de un clúster privado virtual (VPC) para el usuario local dentro del centro bajo demanda. Hay que tener en cuenta que en la solicitud 852 se encuentran más detalles sobre la creación y el uso de VPC. En cada uno de los casos descritos en el presente documento en el que los recursos informáticos bajo demanda se identifican, aprovisionan y consumen por la carga de trabajo del entorno local, los medios por los cuales esto se logra pueden ser mediante la creación de un VPC dentro del centro bajo demanda.

En la FIG. 1 también se muestran otros componentes tales como un gestor de identidad 112 y un dispositivo de provisión de nodo 118 como parte del centro de alojamiento 102. El maestro de alojamiento 128 puede incluir una interfaz de gestor de identidad 112 que puede coordinar información global y local con respecto a usuarios, grupos, cuentas y clases asociadas con recursos informáticos. La interfaz de gestor de identidad 112 también puede permitir que el módulo de gestión 106 cree y modifique automática y dinámicamente cuentas de usuario y atributos de credenciales de acuerdo con las necesidades de carga de trabajo actuales. El maestro de alojamiento 128 permite a los sitios una gran flexibilidad cuando se trata de definir el acceso de credenciales, los atributos y las relaciones. En la mayoría de los casos, el uso de los parámetros USERCFG, GROUPCFG, ACCOUNTCFG, CLASSCFG y QOSCFG es adecuado para especificar la configuración necesaria. Sin embargo, en ciertos casos, como los siguientes, este enfoque puede no ser ideal o ni siquiera adecuado: entornos con conjuntos de usuarios muy grandes; entornos con configuraciones de credenciales muy dinámicas en términos de objetivos de proporcionalidad equilibrada, prioridades, restricciones de acceso a servicios y relaciones de credenciales; entornos de cuadrícula con servicios externas de información de asignación de credenciales; entornos empresariales con políticas de imparcialidad basadas en el uso de varios clústeres.

Los módulos abordan estas y otras cuestiones similares a través del uso del gestor de identidad 112. El gestor de identidad 112 permite al módulo intercambiar información con un servicio de gestor de identidad externo. Al igual que con las interfaces del administrador de recursos del módulo, este servicio puede ser un paquete comercial completo diseñado para este propósito, o algo mucho más simple por el cual el módulo obtiene la información necesaria para un servicio web, un archivo de texto o una base de datos.

La atención se dirige ahora al dispositivo de provisión de nodos 118 y como ejemplo de su funcionamiento, el dispositivo de provisión de nodos 118 puede permitir la asignación de recursos en el centro de alojamiento 102 para la carga de trabajo desde un entorno informático local 104. Como se mencionó anteriormente, un aspecto de este proceso puede ser crear un VPC dentro del centro de alojamiento según lo dirigido por el módulo 108. El módulo de gestión de clientes 108 se comunicará con el módulo de gestión de alojamiento 106 para comenzar el proceso de aprovisionamiento. En un aspecto, el módulo de provisión 118 puede generar otra instancia del software de gestión necesario 120 y 122 que se creará en el entorno del centro de alojamiento así como los nodos informáticos 124 y 126 para ser consumido por un trabajo enviado. El nuevo módulo de gestión 120 se crea sobre la marcha, puede estar asociado con una petición específica y preferiblemente estará operativo en un nodo dedicado si el nuevo módulo de gestión 120 está asociado con una petición o trabajo específico, ya que el trabajo consume los recursos asociados con los nodos informáticos aprovisionados 124, 126 y el trabajo se completa, entonces el sistema quitaría el módulo de gestión 120 puesto que sólo se creó para la solicitud específica de la matriz. El nuevo módulo de gestión 120 puede conectarse a otros módulos tales como el módulo 108. El módulo 120 no tiene necesariamente que ser creado sino que puede ser generado sobre la marcha según sea necesario para ayudar en la comunicación y aprovisionamiento y uso de los recursos en el entorno de utilidad 102. Por ejemplo, el módulo 106 puede continuar y asignar nodos dentro del entorno informático de utilidad 102 y conectar estos nodos directamente al módulo 108, pero en ese caso puede perder alguna capacidad de lote como una compensación. El maestro de alojamiento 128 con el módulo de gestión 106, el gestor de identidad 112 y el dispositivo de provisión de nodo 118 preferiblemente está situado conjuntamente con el entorno informático de utilidad pero puede ser distribuido. El módulo de gestión en el entorno local 108 puede comunicarse directamente con el módulo de gestión 120 creado en el centro de alojamiento para administrar la transferencia de la carga de trabajo y el consumo de los recursos del centro bajo demanda. El módulo de gestión creado 120 puede o no ser parte de un VPC.

Como se muestra en la FIG. 2, un sistema ejemplar para implementar la invención incluye un dispositivo informático de uso general 200, que incluye una unidad de procesamiento (CPU) 220, una memoria de sistema 230 y un bus de sistema 210 que acopla varios componentes de sistema incluyendo la memoria de sistema 230 a la unidad de procesamiento 220. El bus de sistema 210 puede ser cualquiera de varios tipos de estructuras de bus incluyendo un bus de memoria o un controlador de memoria, un bus periférico y un bus local usando cualquiera de una variedad de arquitecturas de bus. El sistema también puede incluir otra memoria tal como memoria de sólo lectura (ROM) 240. En el ROM 240 normalmente se almacena un BIOS básico (entrada/salida) que contiene la rutina básica que ayuda a transferir información entre elementos dentro del dispositivo informático 200, tal como durante la puesta en marcha. El dispositivo informático 200 incluye además medios de almacenamiento tales como una unidad de disco duro 250, una unidad de disco magnético, una unidad de disco óptico, unidad de cinta o similar. El dispositivo de almacenamiento 260 está conectado al bus de sistema 210 por una interfaz de accionamiento. Los accionamientos y los medios legibles por ordenador asociados proporcionan almacenamiento no volátil de instrucciones legibles por ordenador, estructuras de datos, módulos de programa y otros datos para el dispositivo informático 200. Los componentes básicos son conocidos por los expertos en la técnica y se contemplan variaciones adecuadas según tipo de dispositivo, tanto si el dispositivo es un dispositivo informático portátil pequeño, como un ordenador de escritorio o un servidor informático.

Aunque el entorno del ejemplo aquí descrito utiliza el disco duro, debe ser valorado por los expertos en la técnica que otros tipos de medios legibles por ordenador que pueden almacenar datos accesibles por un ordenador, tales como casetes magnéticos, tarjetas de memoria flash, discos de vídeo digitales, cartuchos de memoria, memorias de acceso aleatorio (RAM), memoria de sólo lectura (ROM), y similares, también pueden usarse en el entorno operativo de ejemplo. El sistema anterior proporciona un servidor o dispositivo informático de ejemplo que puede utilizarse y conectarse en red con un clúster, clústeres o una cuadrícula para gestionar los recursos de acuerdo con los principios establecidos aquí. También se reconoce que en el futuro se pueden desarrollar otras configuraciones de hardware sobre las que el método puede ser operativo.

Tal como se ha mencionado, un concepto útil pero no necesario para habilitar la tecnología incluye una función bajo demanda de capacidad fácil de usar y VPCs dinámicos. La solicitud de Patente de Estados Unidos número 11/276.852 presentada el 16 de marzo de 2006, mencionada anteriormente, proporciona detalles adicionales con respecto a los VPC y la capacidad está habilitada en el código fuente incorporado en la primera solicitud provisional. Respecto a la capacidad fácil de usar bajo demanda, la FIG. 3 ilustra una interfaz de ejemplo 300 que un usuario puede utilizar para conectarse a un centro bajo demanda mediante una configuración simple de varios parámetros en cada sitio. Estos parámetros pueden ser preconfigurados y activados de una manera tan simple como usar un botón "enable now " (habilitar ahora) 302. Preferentemente, los términos y acuerdos de licencia pueden ser predeterminados o aceptados con otras licencias del software durante un proceso de instalación o pueden ser revisados a través de un formulario web como respuesta a la activación del servicio. El administrador puede configurar los requisitos de recursos 308 en el centro bajo demanda fácilmente para controlar cuántos procesadores simultáneos, nodos y demás pueden ser utilizados en el centro bajo demanda. Se pueden establecer otros parámetros tales como el tamaño de pasos incrementales, duración mínima y horas de procesador por mes. La interfaz 300 también incluye capacidades de ejemplo tales como personalizar límites de capacidad 304, personalizar las políticas de nivel de servicio 306 y otros permisos de externalización. Por ejemplo, el usuario puede variar los permisos de usuarios, grupos, clases y cuentas con quién puede tener el nivel de permisos de externalización.

Como puede verse en la interfaz 300, hay otros parámetros mostrados, tales como límites de capacidad máxima y de nivel de servicio, límites de tiempo real y niveles de calidad de servicio. De este modo, un usuario puede proporcionar un enfoque personalizado para utilizar el centro bajo demanda. El usuario puede habilitar políticas de cumplimiento de nivel de servicio y aplicar las políticas a diversas gradaciones de la carga de trabajo, como a toda la carga de trabajo con tiempos de espera excesivos, sólo carga de trabajo de alta prioridad con excesivo tiempo de espera que haya aplicado la etiqueta de externalización. También se contemplan otras gradaciones, tales como permitir al usuario definir con más detalle el tiempo de espera "excesivo" o cuán alta es la carga de trabajo de alta prioridad.

El VPC dinámico permite predeterminar, asegurar, optimizar y garantizar la entrega correcta de recursos en entornos de clúster, cuadrícula y centros de alojamiento. El VPC se utiliza para la partición virtual de varios tipos de recursos (como diferentes recursos de hardware, licencias de software, VLAN, almacenamiento, etc.) en unidades que pueden tratarse como clústers independientes. Estos clústeres virtuales independientes pueden tener sus propios controles de política, seguridad, garantías de recursos, optimización, facturación y generación de informes. El VPC utiliza la programación de software de gestión y los controles de política para cambiar automáticamente los límites virtuales para que coincidan con los recursos necesarios con la carga de trabajo

asociada. Por ejemplo, si un cliente necesitaba por primera vez los recursos de una granja informática tradicional de Linux, pero luego con el tiempo tenía una carga de trabajo que necesitaba cada vez más recursos SMP, el VPC dinámico podría adaptar de forma óptima los recursos correctos para que coincidan con los requisitos de la carga de trabajo. El VPC dinámico proporciona flexibilidad para administrar y modificar los recursos en el centro bajo demanda. De lo contrario, los servicios de alojamiento son demasiado rígidos, haciendo que los clientes pasen por las tareas de redefinir y renegociar qué recursos se proporcionan o hacer que paguen por recursos que no concuerdan con sus necesidades cambiantes.

Otros diferenciadores habilitados en el software de gestión incluyen el conocimiento detallado y el control definido de la carga de trabajo que incluye la asignación de la carga de trabajo (CPU frente carga de trabajo intensiva de datos), almacenamiento optimizado de datos, afinidad de recursos, co-asignación de recursos altamente optimizada, integración de aprovisionamiento, integración gestión de seguridad: los controles de cumplimiento del nivel de servicio se relacionan con los tiempos de respuesta garantizados y el tiempo de actividad garantizado. Existen amplias capacidades de gestión, como el soporte de gestores de recursos múltiples y flexibilidad en módulos de gestión, como imágenes de sistema único. A continuación se muestran más detalles sobre estas características.

En cuanto a la asignación de la carga de trabajo, una de las capacidades de inteligencia habilitadas por el conocimiento detallado y el control sobre la carga de trabajo es su capacidad de diferenciar entre la carga de trabajo intensiva de la CPU y la carga de trabajo intensiva de datos. Cuando el software programa la carga de trabajo de HPC para un centro de alojamiento, puede enviar automáticamente carga de trabajo más intensiva de la CPU al sitio de alojamiento, al tiempo que centra la carga de trabajo intensiva de datos localmente. Esto significa que los trabajos con archivos de datos grandes no necesitan unir redes y reduce el tiempo total de respuesta de la carga de trabajo de los clientes. Los clientes estarían más satisfechos porque su trabajo se hace antes y el centro de alojamiento estaría más satisfecho porque puede centrarse en la carga de trabajo que es más rentable para el modelo de facturación "Hora CPU".

El almacenamiento optimizado de datos es otro aspecto del conocimiento detallado del software y control de la carga de trabajo. Esta tecnología aumenta el rendimiento de la carga de trabajo intensiva de datos al romper la reserva de un trabajo en los dos, tres (o más) elementos de los datos de almacenamiento, el procesamiento de la carga de trabajo y los resultados del almacenamiento. Otras tecnologías de programación reservan el procesador y otros recursos en un nodo mientras duran los tres, dejando la CPU inactiva durante el almacenamiento de datos y la capacidad de E/S prácticamente inactiva durante el período de procesamiento. El software de gestión de la presente invención tiene servicio de consulta de información que analiza tanto los servicios de información de archivo como de red y, a continuación, programa inteligentemente los tres procesos de una manera optimizada. La capacidad de E/S está programada para evitar conflictos entre períodos de almacenamiento de datos y la programación de la CPU está optimizada para permitir el uso más completo del procesador subyacente. Una vez más, esto ayuda al cliente final a conseguir más en un período de tiempo más corto y optimiza los recursos de los proveedores de alojamiento para evitar el tiempo de CPU inactivo. La FIG. 4 ilustra cómo funciona el almacenamiento inteligente de datos. La parte superior 402 de esta figura muestra el método tradicional de reservar un nodo completo, incluyendo la CPU, para todo el almacenamiento de datos y tiempo de cómputo. La mitad inferior 404 muestra cómo el software programa el almacenamiento y procesamiento de datos para superponerse y optimizar la carga de trabajo. De este modo, los "eventos" utilizarán la CPU durante los periodos de almacenamiento en lugar de dejar la CPU inactiva durante esos tiempos.

En cuanto a la afinidad de recursos, el módulo de gestión aprovecha su conocimiento detallado de solicitudes de carga de trabajo solicitando trabajos al tipo de recurso capaz de proporcionar el tiempo de respuesta más rápido. Por ejemplo, si es probable que un trabajo se ejecute más rápido en AIX sobre Linux, en un sistema SMP en lugar de una granja de CPU tradicional, o funciona mejor en un tipo de red específico, tales afinidades pueden configurarse manual o automáticamente para que se produzcan de manera que la carga de trabajo se optimice. El software también tiene la capacidad de rastrear estas variables y aplicar tasas más elevadas a los que utilizan los sistemas más costosos.

El software asocia peticiones de carga de trabajo con controles de cumplimiento de nivel de servicio, tales como garantizar el tiempo de respuesta y garantizar el tiempo de actividad. Es importante que los centros informáticos de alto rendimiento bajo demanda sean capaces de administrar el cumplimiento del nivel de servicio, de lo contrario su clientela nunca repetirá con ellos. Una aplicación de esta capacidad es que puede establecer reglas que empujen automáticamente toda la carga de trabajo pendiente de un sitio a un centro de alojamiento. Esta capacidad se puede denominar protección de sobrecarga de carga de trabajo. Los algoritmos avanzados de

programación y las capacidades de gestión de políticas se pueden configurar para satisfacer estas necesidades. A continuación se muestran las industrias que tienen necesidades específicas para tales garantías: Homeland Security (garantía de tiempos de respuesta, así como garantía de tiempo de actividad, protección contra sobrecargas de carga de trabajo); El National Institute of Health deseaba recursos de garantía de software en caso de una crisis nacional, hasta el punto de anticipar todos los trabajos en toda la cuadrícula. Esta característica llamada "Run Now" proporciona el tiempo de respuesta inmediata garantizada requerida. Para ello, realiza una serie de consultas complejas para proporcionar el tiempo de respuesta al menor costo posible para los sitios participantes. El software puede lograr esto ejecutando a través de más de 8 niveles (cualquier número puede servir) de políticas cada vez más agresivas para proporcionar los recursos - comenzando con los niveles menos impactantes y agotando completamente sus opciones antes de aumentar al siguiente nivel más agresivo. De forma similar, la inteligencia del software permite que los sitios de alojamiento ofrezcan niveles de SLA prometidos que mantengan al cliente plenamente satisfecho, al tiempo que proporcionan el mayor retorno posible al proveedor de alojamiento; películas multi-media, juegos, simulación y otras áreas de representación intensa (tiempo de respuesta de garantía); petróleo y gas (tiempo de respuesta de garantía, protección de sobrecarga de carga de trabajo); aeroespacial (tiempo de respuesta de garantía); financiero (garantía de disponibilidad y tiempo de respuesta de garantía, protección contra sobrecarga de carga de trabajo); Fabricantes: farmacéutica, automoción, chips y otras industrias intensivas "First to Market" (garantía de tiempo de respuesta, protección contra sobrecarga de carga de trabajo). Como se puede ver, el software ofrece características aplicables en muchos mercados.

Otra característica se refiere a la arquitectura del software que permite el monitoreo, programación y administración simultáneas de múltiples tipos de recursos, y se puede desplegar en diferentes entornos o utilizarse como un punto central de conexión para entornos distintos. En cuanto a la amplia compatibilidad, los elementos del lado del servidor del software funcionan al menos en entornos Linux, Unix y Mac OS X (puede administrar Linux, Unix, Mac OS X, Windows y entornos de sistema central, según lo que soporte el gestor de recursos local). El software del lado del cliente funciona en entornos Linux, Unix, Mac OS X y Windows, así como en otros entornos.

El soporte de gestor de múltiples recursos permite que el software funcione prácticamente en todos los gestores de recursos informáticos convencionales. Estos gestores de recursos informáticos incluyen, pero no están limitados a LoadLeveler, LSF, PBSPro, TORQUE, OpenPBS y otros. Esto no sólo incrementa el número de entornos en los que puede utilizarse para proporcionar capacidad bajo demanda, sino que ofrece al cliente un conjunto más amplio de opciones que avanzan porque no las bloquea en la solución de un proveedor concreto. Además, con el soporte de administrador de múltiples recursos, el software puede interoperar con múltiples gestores de recursos informáticos al mismo tiempo, lo que permite capacidades de cuadrícula incluso en entornos mixtos.

Más allá del gestor de recursos informáticos tradicional que administra la sumisión de trabajos a nodos informáticos, el software puede integrarse con gestores de recursos de almacenamiento, gestores de recursos de red, gestores de recursos de licencias de software, etc. Utiliza esta multiplicidad de fuentes de información para hacer que sus decisiones de política sean más efectivas. El software también puede conectarse para monitorear hardware como Ganglia, scripts personalizados, ejecutables y bases de datos para obtener información adicional que la mayoría de los administradores de recursos locales informáticos no tendrían disponibles. Esta información adicional puede ser consultada y evaluada por el software o un administrador para aplicarse a las decisiones de colocación de la carga de trabajo y a otras políticas del sistema.

La FIG. 5 ilustra gráficamente 500 cómo el WM se integra con otras tecnologías. Los elementos de la parte inferior son tipos de recursos como almacenamiento, licencias y redes. Los elementos de la izquierda son mecanismos de interfaz para usuarios finales y administradores. Los elementos del lado derecho de la figura son servicios con los que el software puede integrarse para proporcionar capacidades ampliadas adicionales, tales como aprovisionamiento, generación de informes centrada en la base de datos y gestión de la asignación. Los ejemplos de paquetes de software mostrados en la FIG. 5 son principalmente productos de IBM, pero por supuesto puede integrarse otro software.

Con respecto a la flexibilidad de los modelos de gestión, el software permite proporcionar la capacidad bajo demanda de cualquier entorno de clúster soportado o entorno de cuadrícula. El software puede configurarse para permitir múltiples tipos de cuadrícula y modelos de gestión. Los dos tipos de cuadrículas preferibles habilitadas por el software son las cuadrículas de área local y cuadrículas de área amplia, aunque también están habilitadas otras. La FIG. 6 ilustra 600 ejemplos de varios tipos de cuadrícula así como varios escenarios de gestión de cuadrícula. Una "Cuadrícula de Área Local" (LAG) utiliza una instancia de un administrador de carga

de trabajo WM, como Moab, dentro de un entorno que comparte un espacio de usuario y de datos a través de varios clústeres, que pueden o no tener varios tipos de hardware, sistemas operativos y administradores de recursos informáticos (por ejemplo, LoadLeveler, TORQUE, LSF, PBSPro, etc). Los beneficios de una LAG son que es muy fácil de configurar y aún más fácil de manejar. Básicamente, todos los clústeres se combinan en una LAG utilizando una instancia de WM, eliminando la administración de políticas redundantes y los informes. Los clústeres parecen ser un conjunto mixto de recursos en un solo clúster grande. Una "Cuadrícula de Área Amplia" (WAG) utiliza múltiples instancias WM que trabajan juntas dentro de un entorno que puede uno o más usuarios y espacios de datos a través de varios clústeres, que pueden o no tener tipos de hardware mixtos, sistemas operativos y gestores de recursos informáticos (por ejemplo, LoadLeveler, TORQUE, LSF, PBSPro, etc). Las reglas de gestión de WAG pueden ser centralizadas, controladas localmente o mezcladas. El beneficio de una WAG es que una organización puede mantener la gestión soberana de su propio clúster local, al tiempo que establece políticas estrictas o relajadas de intercambio de sus recursos a la cuadrícula externa. La colaboración puede ser facilitada con un conjunto muy flexible de políticas opcionales en las áreas de propiedad, control, intercambio de información y privacidad. Los sitios son capaces de elegir la cantidad de recursos e información de su clúster que comparten con la cuadrícula externa.

Las cuadrículas son intrínsecamente de naturaleza política y flexibilidad para gestionar qué información se comparte y qué información no es fundamental para establecer tales cuadrículas. Usando el software, los administradores pueden crear políticas para administrar el intercambio de información en entornos políticos difíciles.

Las organizaciones pueden controlar el intercambio de información y la privacidad de al menos tres maneras diferentes: (1) Permitir que todos los recursos (por ejemplo, nodos, almacenamiento, etc.), carga de trabajo (por ejemplo, trabajos, reservas, etc.) y de política de información (por ejemplo, reglas de intercambio y priorización) se compartan para dar cuentas e informes completos; (2) Permitir que otros sitios solo vean recursos, cargas de trabajo y de información de políticas que les concierna para que los detalles completos del recurso puedan mantenerse privados y más simplificados; (3) Permitir que otros sitios sólo vean un único bloque de recursos, revelando nada más que el volumen agregado de recursos disponibles para el otro sitio. Esto permite que los recursos, la carga de trabajo y la información de política se mantengan en privado, al mismo tiempo que permite que se lleven a cabo relaciones compartidas. Por ejemplo, un sitio que tiene 1.024 procesadores puede mostrar públicamente sólo 64 procesadores a otros sitios de la cuadrícula.

Los tipos de cuadrícula y los escenarios de administración mencionados anteriormente se pueden combinar junto con las reglas de intercambio de información y privacidad para crear relaciones personalizadas que coincidan con las necesidades de las organizaciones subyacentes. La FIG. 7 ilustra un ejemplo de cómo se pueden combinar cuadrículas. Muchas combinaciones son posibles.

El software es capaz de facilitar virtualmente cualquier relación de cuadrícula tal como uniendo cuadrículas de área local en cuadrículas de área amplia; uniendo cuadrículas de área amplia a otras cuadrículas de área amplia (ya sean administradas centralmente, localmente - "punto a punto" o mixtas); compartir recursos en una dirección (por ejemplo, para usar con centros de alojamiento o alquilar los propios recursos); permitir múltiples niveles de relaciones de cuadrícula (por ejemplo, conglomerados dentro de conglomerados). Como se puede apreciar, el entorno local puede ser una de muchas configuraciones como se discutió en el ejemplo anterior.

A continuación se describen varios aspectos de la descripción sobre al acceso a un centro bajo demanda desde un entorno local. Un aspecto se refiere a permitir la detección automática de un evento tal como umbrales de recursos o umbrales de servicio dentro del entorno informático 104. Por ejemplo, si se cumple un umbral del 95% del consumo del procesador, ya que se utilizan 951 procesadores de los 1000 procesadores en el entorno, entonces el WM 108 puede establecer automáticamente una conexión con el entorno bajo demanda 102. Un umbral de servicio, un umbral basado en políticas, un umbral basado en hardware o cualquier otro tipo de umbral puede activar la comunicación al centro de alojamiento 102. Otros eventos también pueden desencadenar la comunicación con el centro de alojamiento, como una carga de trabajo con una cierta configuración. El WM 108 entonces puede comunicarse con WM 106 para proporcionar o personalizar los recursos bajo demanda 102. La creación de un VPC en el centro bajo demanda puede ocurrir. Los dos entornos intercambian la información necesaria para crear reservas de recursos, proveer los recursos, administrar licencias, etc., necesarios para permitir la transferencia automática de trabajos u otra carga de trabajo desde el entorno local 104 al entorno de trabajo bajo demanda 102. Nada cambia sobre un trabajo de usuario 110 enviado a un WM 108. El entorno físico del entorno informático local 104 también se puede replicar en el centro bajo demanda. El entorno bajo demanda 102 inmediatamente comienza a ejecutar el trabajo sin ningún cambio en el trabajo o tal vez incluso cualquier conocimiento del emisor.

En otro aspecto, los eventos predichos también pueden ser desencadenantes. Por ejemplo, un fallo previsto de nodos dentro del entorno local, eventos predichos internos o externos al entorno o una unión prevista de umbrales puede desencadenar la comunicación con el centro bajo demanda. Todos estos son configurables y pueden activar automáticamente la migración de trabajos o carga de trabajo o pueden desencadenar una notificación al usuario o al administrador para tomar una decisión sobre si migrar la carga de trabajo o acceder al centro bajo demanda.

En el caso del análisis y transferencia de la carga de trabajo pendiente, la realización del método proporciona determinar si existe una condición de carga de trabajo pendiente en el entorno informático local. Si la condición de carga de trabajo pendiente existe, entonces el sistema analiza la carga de trabajo pendiente, comunica la información asociada con el análisis al entorno informático bajo demanda, proporciona el entorno informático bajo demanda de acuerdo con la carga de trabajo pendiente analizada y transfiere la carga de trabajo pendiente al entorno informático bajo demanda. Es preferible que el aprovisionamiento del entorno informático bajo demanda comprenda también crear un clúster privado virtual en el entorno informático bajo demanda. El análisis de la carga de trabajo puede comprender determinar al menos un tipo de recurso asociado con la carga de trabajo de reserva para el aprovisionamiento en el entorno informático bajo demanda.

En otro aspecto, analizar la carga de trabajo pendiente, comunicar la información asociada con el análisis al entorno informático bajo demanda, el aprovisionamiento del entorno informático bajo demanda de acuerdo con la carga de trabajo pendiente y la transferencia de la carga de trabajo pendiente al entorno informático aprovisionado se produce en respuesta a una operación de un solo clic de un administrador. Sin embargo, el proceso de aprovisionamiento y transferencia de carga de trabajo pendiente al centro bajo demanda puede comenzar en función de cualquier número de eventos. Por ejemplo, un usuario puede interactuar con una interfaz de usuario para iniciar la transferencia de carga de trabajo pendiente. Un evento interno tal como un umbral, por ejemplo, un tiempo de espera que alcanza un máximo, puede ser un evento que podría desencadenar los análisis y transferir. Un evento externo también puede desencadenar la transferencia de la carga de trabajo pendiente, tal como un ataque terrorista, condiciones climáticas, apagones, etc.

Hay varios aspectos de esta invención que se muestran en el código fuente adjunto. Una es la capacidad de intercambiar información. Por ejemplo, para la transferencia automática de carga de trabajo al centro bajo demanda, el sistema importará clases remotas, información de política de configuración, información de hardware físico, sistemas operativos y otra información desde el entorno WM 108 al WM 106 esclavo para uso del entorno bajo demanda 102. La información sobre el entorno informático bajo demanda, los recursos, las políticas, etc., también se comunican desde el WM 106 esclavo a la WM 108 local.

Por tanto, un método puede proporcionar un método para gestionar recursos entre el entorno informático local y el entorno bajo demanda. Un método a modo de ejemplo comprende detectar un evento asociado con un entorno informático local. Como se ha mencionado, el evento puede ser cualquier tipo de desencadenante o umbral. El software entonces identifica información sobre el entorno local, establece la comunicación con un entorno informático bajo demanda y transmite la información sobre el entorno local al entorno de informático bajo demanda. Con esa información, el software proporciona recursos dentro del entorno informático bajo demanda para duplicar sustancialmente el entorno local y transferir la carga de trabajo desde el entorno local al entorno informático bajo demanda. En otro aspecto, el aprovisionamiento no necesariamente duplica el entorno local, sino que especialmente aprovisiona el entorno bajo demanda para la carga de trabajo migrada al centro bajo demanda. Como ejemplo, la información comunicada sobre el entorno local puede estar relacionada con al menos hardware y/o un sistema operativo. El establecimiento de la comunicación con el entorno informático bajo demanda y la transmisión de la información sobre el entorno local al entorno informático bajo demanda se pueden realizar de forma automática o manual a través de una interfaz de usuario. El uso de una interfaz de este tipo puede permitir al usuario proporcionar una petición de un solo clic o una acción para establecer la comunicación y migrar la carga de trabajo al centro bajo demanda.

En algunos casos, cuando el software busca proporcionar recursos, un recurso particular puede no ser capaz de duplicarse en el entorno informático bajo demanda. En este escenario, el software identificará y seleccionará un recurso sustituto. Este proceso de identificación y selección de un recurso sustituto puede realizarse ya sea en el entorno bajo demanda o mediante negociación entre un gestor de carga de trabajo esclavo en el entorno bajo demanda y un gestor maestro de carga de trabajo en el entorno informático local. El método comprende además identificar un tipo de carga de trabajo para transferir al entorno bajo demanda, y en el que la transferencia de carga de trabajo desde el entorno local al entorno informático bajo demanda comprende además transferir solamente el tipo de carga de trabajo identificado al centro bajo demanda. En otro aspecto, la transferencia del

tipo de carga de trabajo identificado al centro bajo demanda se basa en diferentes capacidades de hardware y/o software entre el entorno bajo demanda y el entorno informático local.

Otro aspecto de la descripción es la capacidad para automatizar la gestión de datos entre dos sitios. Esto implica la manipulación transparente de la gestión de datos entre el entorno bajo demanda 102 y el entorno local 104 que es transparente para el usuario. En otras palabras, puede realizarse sin una acción explícita o una configuración por parte del usuario. También puede ser desconocida para el usuario. Otro aspecto más se refiere a un mecanismo sencillo y fácil para permitir la integración de centros bajo demanda. Este aspecto de la invención implica la capacidad del usuario o de un administrador para, en una sola acción como el clic de un botón, tocar una pantalla sensible al tacto, detección de movimiento, u otra acción simple, para poder ordenar la integración de una información y capacidad del centro bajo demanda en la WM 108 local. A este respecto, el sistema de la invención será capaz de intercambiar e integrar automáticamente toda la información y conocimiento de recursos necesarios en un solo clic para ampliar el conjunto de recursos que pueden estar disponibles para los usuarios que tienen acceso inicialmente sólo al entorno informático local 104. La información puede incluir los diversos aspectos de los recursos disponibles en el centro bajo demanda, tales como el marco temporal, el coste de los recursos, el tipo de recursos, etc.

Uno de los aspectos de la integración de un entorno bajo demanda 102 y un entorno informático local 104 es que los datos generales aparecen localmente. En otras palabras, el WM 108 tendrá acceso a los recursos y conocimientos del entorno bajo demanda 102, pero la visión de esos recursos, con el adecuado cumplimiento de los requisitos de las políticas locales, se maneja localmente y aparece localmente a los usuarios y administradores del entorno local 104.

Otro aspecto está habilitado con el código fuente adjunto: es la capacidad de especificar la información de configuración asociada con el entorno local 104 y alimentarlo al centro de alojamiento 102. Por ejemplo, la interacción entre los entornos informáticos soporta reservas estáticas. Una reserva estática es una reserva que un usuario o administrador no puede cambiar, eliminar o destruir. Es una reserva que está asociada con el WM 108 en sí. Una reserva estática bloquea los marcos de tiempo cuando los recursos no están disponibles para otros usos. Por ejemplo, si para habilitar un entorno informático para ejecutar (consumir) recursos, un trabajo tarda una hora en aprovisionar recursos, entonces el WM 108 puede hacer una reserva estática de recursos para el proceso de aprovisionamiento. El WM 108 creará localmente una reserva estática para el componente de aprovisionamiento de ejecutar el trabajo. El WM 108 informará sobre estas restricciones asociadas con la reserva estática creada.

A continuación, el WM 108 se comunicará con el WM 106 esclavo si se necesitan recursos bajo demanda para ejecutar un trabajo. El WM 108 se comunica con el WM 106 esclavo e identifica qué recursos se necesitan (20 procesadores y 512 MB de memoria, por ejemplo) y pregunta cuándo pueden estar disponibles esos recursos. Supongamos que WM 106 responde que los procesadores y la memoria estarán disponibles en una hora y que el WM 108 puede tener esos recursos durante 36 horas. Una vez que toda la información adecuada ha sido comunicada entre el WM 106 y el WM 108, entonces WM 108 crea una reserva estática para bloquear la primera parte de los recursos que requiere la hora de aprovisionamiento. El WM 108 también puede bloquear los recursos con una reserva estática desde la hora 36 hasta el infinito, hasta que los recursos desaparezcan. Por lo tanto, de cero a una hora se bloquea por una reserva estática y desde el final de las 36 horas hasta el infinito se bloquea. De esta manera, el programador 108 puede optimizar los recursos bajo demanda y asegurar que estén disponibles para los servicios locales. La comunicación entre los WM 106 y 108 se realiza preferiblemente mediante tunelado.

Otro aspecto más es la capacidad de tener un solo agente como el WM 108 u otro agente de software que detecte un parámetro, evento o configuración en el entorno local 104. El entorno, en este sentido, incluye hardware y software y otros aspectos del entorno. Por ejemplo, un entorno de clúster 104 puede tener, además de las políticas y restricciones en usuarios y grupos como se ha explicado anteriormente, cierta configuración de hardware/software tal como cierto número de nodos, cierta cantidad de memoria y espacio en disco, sistemas operativos y software cargado en los nodos y así sucesivamente. El agente (que puede ser WM 108 u otro módulo de software) determina los aspectos físicos del entorno informático 104 y se comunica con el centro de alojamiento bajo demanda para proporcionar un aprovisionamiento automático de recursos dentro del centro 102 de tal manera que el entorno local se duplica. La duplicación puede coincidir con la misma configuración de hardware/software o puede sustituir de forma dinámica o manual los componentes alternativos. La comunicación y transferencia de carga de trabajo a un entorno replicado dentro del centro de alojamiento 102 puede producirse automáticamente (por ejemplo, en la detección de un valor de umbral) o al presionar un botón de un

administrador. Por lo tanto, se examina la información relativa al entorno local y el WM 108 u otro agente de software transfiere dicha información al centro 102 de alojamiento para su replicación.

La replicación, por tanto, implica proporcionar el mismo -o quizás similar- número de nodos, sistemas operativos de aprovisionamiento, arquitectura del sistema de archivos y memoria y cualquier otro aspecto de hardware o software del centro de alojamiento 102 usando WM 106 para replicar el entorno informático 104. Los expertos en la técnica entenderán que otros elementos podrán necesitarse para ser aprovisionados para duplicar el entorno. Cuando el entorno exacto no puede replicarse en el centro de alojamiento 102, el WM 106 puede tomar decisiones o mediante negociación entre WM 106 y WM 108 para determinar un aprovisionamiento alternativo.

En otro aspecto, un usuario del entorno informático 104 tal como un administrador puede configurar en el sitio de cliente 104 un entorno informático y cuando la carga de trabajo se transfiere al centro de alojamiento 102, se puede proporcionar el entorno informático deseado. En otras palabras, el administrador podría configurar un entorno mejor o más adecuado que el entorno informático 104 existente. Como ejemplo, una empresa puede querer construir un entorno informático 104 que será utilizado por trabajos intensivos de procesador y trabajos intensivos de memoria. Puede ser más barato para el administrador del entorno 104 crear un entorno que se adapte mejor a los trabajos intensivos del procesador. El administrador puede configurar un entorno intensivo de procesador en el clúster local 104 y cuando se envía un trabajo intensivo de memoria 110, el entorno intensivo de memoria se puede aprovisionar en el centro de alojamiento 102 para descargarlo.

A este respecto, el administrador puede generar perfiles de varias configuraciones para varias provisiones de "un solo clic" en el centro de alojamiento 102. Por ejemplo, el administrador puede tener perfiles para trabajos informáticos intensivos, trabajos intensivos en memoria, tipos de sistema operativo, tipos de software, cualquier combinación de requisitos de software y hardware y otros tipos de entornos. Los expertos en la técnica comprenderán los diversos tipos de perfiles que se pueden crear. El clúster local 104 tiene una relación con el centro de alojamiento 102 en el que el administrador puede transferir la carga de trabajo basada en uno de los múltiples perfiles creados. Esto puede hacerse automáticamente si el WM 108 identifica un trabajo de usuario 110 que coincide con un perfil o puede realizarse manualmente por el administrador a través de una interfaz de usuario que puede o no ser gráfica. El administrador, en la opción "un clic" puede seleccionar una opción para transferir el componente intensivo de memoria de esta carga de trabajo al centro de alojamiento para aprovisionar y procesar según el perfil de memoria intensiva.

La relación entre el centro de alojamiento 102 y el clúster local 104 por medio de la organización de la gestión de la carga de trabajo puede establecerse de antemano o dinámicamente. El ejemplo anterior ilustra el escenario donde la disposición se crea de antemano donde existen perfiles para su selección por un sistema o un administrador. El escenario dinámico puede ocurrir cuando el administrador local para el entorno 104 tiene un nuevo usuario con un perfil deseado diferente al de los perfiles ya creados. El nuevo usuario quiere utilizar los recursos 104. Los perfiles configurados para nuevos usuarios o grupos pueden agregarse y/o negociarse manualmente entre el centro de alojamiento 102 y el clúster local 104 o pueden ser automáticos. Pueden existir disposiciones para la identificación automática de un tipo diferente de perfil y WM 108 (u otro módulo) puede comunicarse con WM 106 (u otro módulo) para organizar la disponibilidad/capacidad del centro bajo demanda para manejar la carga de trabajo de acuerdo con el nuevo perfil y para organizar el costo, etc. Si no se puede crear un nuevo perfil, entonces se seleccionará un perfil por defecto o genérico, o el más cercano que exista previamente, para cubrir las necesidades del trabajo del nuevo usuario. De esta manera, el sistema puede administrar de forma fácil y dinámica la adición de nuevos usuarios o grupos al clúster local 104.

A este respecto, cuando WM 108 somete una consulta al WM 106 indicando que necesita cierto conjunto de recursos, pasa el perfil(s) y también. WM 106 identifica cuándo están disponibles los recursos en dimensiones estáticas (tal como identifica que cierta cantidad de memoria, nodos y/o otros tipos de arquitectura están disponibles). Este paso identificará si el solicitante obtiene los recursos sin procesar para satisfacer esas necesidades. A continuación, el WM 106 gestionará la instalación del cliente y el aprovisionamiento del software, los sistemas operativos y demás según el perfil recibido. De esta manera, se puede satisfacer toda la especificación de las necesidades de acuerdo con el perfil.

Otro aspecto de la invención se refiere a observar la carga de trabajo que desborda al centro de alojamiento. El sistema puede personalizar el entorno para la carga de trabajo de desbordamiento en particular. Esto se mencionó anteriormente. El agente 108 puede examinar la carga de trabajo en el clúster local 104 y determinar qué parte de esa carga de trabajo o si toda esa carga de trabajo puede transferirse al centro de alojamiento 102. El agente identifica si el entorno local está sobrecargado de trabajo y qué tipo de trabajo está causando la sobrecarga. El agente puede identificar de forma preventiva la carga de trabajo que sobrecargaría el entorno

local o puede identificar dinámicamente el trabajo de sobrecarga que se está procesando. Por ejemplo, si se envía un trabajo 110 que es intensivo en procesador e intensivo en memoria, el WM 108 reconocerá esto y se comunicará inteligentemente con el WM 106 para transferir la parte intensiva del procesador de la carga de trabajo al centro de alojamiento 102. Esto puede ser preferible por varias razones. Quizás sea más barato utilizar el tiempo de procesamiento del centro de alojamiento 102 para un tiempo de procesamiento intensivo. Tal vez el entorno local 104 sea más adecuado para el componente intensivo de memoria de la carga de trabajo. Además, tal vez las restricciones tales como el ancho de banda, las políticas de usuario, las reservas actuales en el entorno local o el entorno de alojamiento, etc., puedan regir cuándo se procesa la carga de trabajo. Por ejemplo, la decisión de dónde procesar la carga de trabajo puede responder al conocimiento de que el entorno 104 no es tan adecuado para el componente intensivo del procesador de la carga de trabajo o debido a otros trabajos que se ejecutan o están programados para ejecutarse en el entorno 104. Como se mencionó anteriormente, el WM 106 gestiona el aprovisionamiento adecuado del entorno del centro de alojamiento para la carga de trabajo de desbordamiento.

Cuando el agente ha identificado cierto tipo de carga de trabajo que está causando la sobrecarga, el sistema puede proporcionar automáticamente en el centro de alojamiento los tipos apropiados de recursos para que coincidan con la carga de trabajo de sobrecarga y luego transferir esa carga de trabajo.

Otro ejemplo de cómo esto funciona, se puede conseguir un umbral para el trabajo que se procesa en el clúster local 104. El umbral puede ser satisfecho por la cantidad de potencia de procesamiento que se está utilizando, la cantidad de memoria disponible, si el usuario ha alcanzado una restricción de permisos, una calidad de servicio puede no ser cumplida ni ningún otro parámetro. Una vez que se alcanza ese umbral, ya sea automáticamente o a través de un administrador, puede presionarse un botón y WM 108 analiza la carga de trabajo en el entorno 104. Puede identificar que hay un trabajo pendiente y determina que se necesitan más nodos (o más de cualquier tipo específico de recurso). El WM 108 se comunicará con WM 106 y recursos de autoprovisión dentro del centro de alojamiento para satisfacer las necesidades de los trabajos pendientes. Los recursos, el hardware, el software, los permisos y las políticas apropiados pueden duplicarse exactamente o de una manera aceptable para resolver el retraso. Además, el auto-aprovisionamiento se puede realizar con referencia a las necesidades de carga de trabajo pendiente en lugar de la configuración de entorno local. A este respecto, se identifica y analiza la carga de trabajo de desbordamiento y se adapta el aprovisionamiento en el centro de alojamiento a la carga de trabajo en sí (en contraste con la adecuación del entorno local) para el procesamiento cuando se transfiere la carga de trabajo de reserva. Por lo tanto, el aprovisionamiento puede estar basado en un tipo de recurso específico que resolverá más eficientemente la carga de trabajo pendiente.

Un aspecto de esta descripción se refiere a la aplicación de los conceptos anteriores para proporcionar un servidor de sitio web con potencia de computación de seguridad a través de un centro de alojamiento 102. Este aspecto de la invención se muestra por el sistema 800 en la Fig. 8. El centro de alojamiento 102 y WM 106 se configuran como se ha explicado anteriormente y se hace el ajuste necesario para comunicarse con un servidor web 802. Una versión de sitio web del gestor de carga de trabajo (WM) 804 operaría en el servidor web 302. Los ajustes conocidos se hacen para permitir que el Servicio de Nombres de Dominio (DNS) proporcione la configuración del desbordamiento del tráfico de red para dirigirse al servidor web 802 o al centro de alojamiento 102. En otro aspecto, el servidor web preferentemente manejaría todo el desvío del tráfico al centro bajo demanda una vez que fuera aprovisionado para el tráfico del Web del desbordamiento. En otro aspecto, un servicio de red separado puede proporcionar el control del control del tráfico web dirigido al servidor web o al centro bajo demanda. Un experto en la técnica entenderá la información básica acerca de cómo los paquetes de información de protocolo de Internet (IP) son enrutados entre un navegador web en un dispositivo de computación de cliente y un servidor web 802.

A este respecto, el WM 804 supervisaría el tráfico web 306 y los recursos en el servidor web 802. Naturalmente, el servidor web 802 puede ser un grupo o grupo de servidores configurados para proporcionar un sitio web. El WM 804 está configurado para tratar el tráfico web 806 y todo lo relacionado con cómo el tráfico web consume recursos dentro del servidor web 802 como un trabajo o un grupo de trabajos. Un evento como un umbral es detectado por WM 804. Si se pasa el umbral o se produce el evento, el WM 804 se comunica con el WM 106 del centro de alojamiento 102, el WM 106 autoprovisiona los recursos y permite que el tráfico web fluya al centro de alojamiento 102 donde se recibirán las solicitudes y páginas web y se devuelve el contenido web. El aprovisionamiento de recursos también puede realizarse manualmente, por ejemplo, en la preparación para un tráfico web aumentado por alguna razón. Como ejemplo, si una compañía de seguros sabe que un huracán está llegando puede proporcionar y prepararse para el aumento del tráfico del sitio web.

La gestión del tráfico web 806 al servidor web 802 y al centro de alojamiento 102 también puede coordinarse de tal manera que una parte de las solicitudes vaya directamente al centro de alojamiento 102 o sea encaminada desde el servidor web 802 al centro de alojamiento 102 para la respuesta. Por ejemplo, una vez completada la provisión en el centro de alojamiento 102, un agente (que puede comunicarse con el WM 804) puede entonces interceptar el tráfico web dirigido al servidor web 302 y dirigirlo al centro de alojamiento 102, que puede proporcionar contenido del sitio web directamente al navegador del cliente (no mostrado) solicitando la información. Los expertos en la técnica reconocerán que hay varias maneras en que el tráfico web 806 pueda ser interceptado y enrutado a los recursos aprovisionados en el centro de alojamiento 102, de tal manera que sea transparente para el navegador web del cliente que un centro de alojamiento 102 -en lugar del servidor web 802- esté dando servicio a la sesión web.

La identificación del umbral puede basarse en un aumento del tráfico actual o puede identificarse desde otra fuente. Por ejemplo, si el New York Times o algún otro medio de comunicación importante menciona un sitio web, ese evento puede causar un aumento previsible del tráfico. A este respecto, un aspecto de la invención es un control de posibles desencadenantes para aumentar la actividad de la banda. El seguimiento puede ser a través de una búsqueda automática Google (o de cualquier tipo) del nombre del sitio web en medios de comunicación como www.nytimes.com. www.washingtonpostcom o www.powerlineblog.com. Si el sitio web se identifica en estos medios, entonces un administrador o automáticamente el aprovisionamiento pueden producirse en un momento predecible de cuándo se produciría el aumento de tráfico.

Otro aspecto de la invención se ilustra en un ejemplo. En un caso, un sitio web pequeño (podemos denominarlo www.smallsite.com) se mencionó en la página del motor de búsqueda de Google™. Debido al gran número de ususarios de Google, www.smallsite.com se vino abajo. Para evitar que esto suceda, cuando una fuente de tráfico elevado como www.google.com o www.nytimes.com enlazan a o se remiten a un sitio web de poco tráfico, entonces el aprovisionamiento puede realizarse automáticamente, por ejemplo, si se creó el enlace de Google a www.smallsite.com, y el sistema (ya sea Google o una función especial disponible en cualquier sitio web) identificó que se estableció un vínculo de este tipo que probablemente causaría una mayor cantidad de tráfico, entonces podría producirse el aprovisionamiento necesario, reflejando el contenido, etc, podrían producirse entre el servidor web 802 y el centro de alojamiento 102 y las modificaciones DNS necesarias para permitir la descarga de parte o de la totalidad del tráfico web al centro de alojamiento.

Si parte del tráfico se dirige al centro de alojamiento 102, entonces se hacen las provisiones para enviar ese tráfico, directa o indirectamente, al centro de alojamiento 102. En un aspecto, los datos se reflejan en el centro de alojamiento 102 y el centro de alojamiento puede manejar exclusivamente el tráfico hasta que se alcance cierto umbral y el tráfico web se pueda transferir automáticamente al servidor web 802.

La descarga del tráfico de la web puede aparecer como un cargo añadido disponible para sitios web, así como cargos o tarifas por los servicios que se pueden utilizar para identificar cuándo puede aumentar el tráfico. Las fuerzas externas (como mencionar un sitio web en las noticias) pueden desencadenar el aumento, así como las fuerzas internas. Por ejemplo, si una oferta especial se publica en un sitio web a un precio reducido para un producto, entonces el sitio web puede esperar un aumento en el tráfico. A este respecto, puede haber una opción de "un clic" para identificar un período de tiempo (1 día de descarga) y un tiempo de inicio (2 horas después de que se publique la oferta) para que se produzca la descarga.

Como se puede apreciar, los principios de la presente invención permiten al usuario medio "navegar" en la red para disfrutar el acceso y la experiencia a sitios web que de otro modo podrían no estar disponibles debido al gran tráfico de Internet. El beneficio ciertamente acostumbra a los propietarios de sitios web y operadores que evitarán el tiempo de inactividad no deseado y el impacto negativo que puede tener en su negocio.

La FIG. 9 ilustra un aspecto del método de la realización de servidor web de la invención. En este caso, se describe un método de gestión de recursos entre un servidor web y un entorno informático bajo demanda, el método incluye determinar si el tráfico web dirigido al servidor web debe ser al menos parcialmente servido a través del entorno informático bajo demanda (902), aprovisionando los recursos dentro del entorno informático bajo demanda para permitirle que responda al tráfico web por el servidor web (904), establecer una ruta de al menos parte del tráfico web desde el servidor web al entorno informático bajo demanda (906) y comunicar datos entre un navegador de cliente y el entorno informático bajo demanda, de manera que el uso del entorno informático bajo demanda para el tráfico web sea transparente (908).

Aunque las reivindicaciones siguientes son reivindicaciones de método, se entiende que las etapas pueden ser practicadas por módulos informáticos en una realización del sistema de la invención, así como relacionadas con

instrucciones para controlar un dispositivo informático almacenado en un medio legible por ordenador. La invención también puede comprender un entorno informático local 104 y/o un centro bajo demanda 102 configurado para operar como se ha descrito anteriormente. Un servidor/servidores web (802) y/o el centro bajo demanda (102) con cualquier otro nodo de red configurado para permitir la descarga del tráfico web (806) también puede ser una realización de la invención. Esto también puede implicar una alteración de software adicional en un navegador web para permitir la descarga del tráfico web. Además, cualquier sistema o red de hardware también se puede realizarse en la invención.

Las realizaciones dentro del ámbito de la presente invención también pueden incluir medios legibles por ordenador para llevar o tener instrucciones ejecutables por ordenador o datos estructurados de ahí almacenados. Tales medios legibles por ordenador pueden ser cualquier medio disponible al que pueda acceder un ordenador con propósito general o de propósito especial. A modo de ejemplo, y sin limitación, tales medios legibles por ordenador pueden comprender RAM, ROM, EEPROM, CD-ROM u otro almacenamiento en disco óptico, almacenamiento en disco magnético u otros dispositivos de almacenamiento magnético, o cualquier otro medio que pueda ser utilizado para transportar o almacenar los medios de código de programa deseados en forma de instrucciones ejecutables por ordenador o estructuras de datos. Cuando la información se transfiere o se proporciona a través de una red u otra conexión de comunicaciones (ya sea cableada, inalámbrica o combinación de los mismos) a un ordenador, el ordenador considera correctamente la conexión como un medio legible por ordenador. Por lo tanto, cualquier conexión de este tipo se denomina correctamente medio legible por ordenador. Las combinaciones de lo anterior también deben incluirse dentro del ámbito de los medios legibles por ordenador.

Las instrucciones ejecutables por ordenador incluyen, por ejemplo, instrucciones y datos que hacen que un ordenador de propósito general, un ordenador de propósito especial o un dispositivo de procesamiento de propósito especial realicen una determinada función o grupo de funciones. Las instrucciones ejecutables por ordenador también incluyen módulos de programa que son ejecutados por ordenadores en entornos autónomos o en red. Generalmente, los módulos de programa incluyen rutinas, programas, objetos, componentes y estructuras de datos, etc. que realizan tareas particulares o implementan tipos de datos abstractos concretos. Las instrucciones ejecutables por ordenador, las estructuras de datos asociadas y los módulos de programa representan ejemplos de los medios de código de programa para ejecutar pasos de los métodos aquí descritos. La secuencia particular de tales instrucciones ejecutables o estructuras de datos asociadas representa ejemplos de actos correspondientes para implementar las funciones descritas en dichos pasos.

Los expertos en la técnica apreciarán que se pueden practicar otras realizaciones de la invención en entornos informáticos en red con muchos tipos de configuraciones del sistema informático, incluyendo ordenadores personales, dispositivos portátiles, sistemas multiprocesador, sistemas basados en microprocesador o electrónica de consumo programable, PCs de red, miniordenadores, ordenadores de marco principal y similares. También se pueden llevar a cabo realizaciones en entornos informáticos distribuidos en los que las tareas sean realizadas por dispositivos de procesamiento locales y remotos que estén enlazados (ya sea por enlaces cableados, enlaces inalámbricos o por una combinación de los mismos) a través de una red de comunicaciones. En un entorno informático distribuido, los módulos de programa pueden estar ubicados en dispositivos de almacenamiento de memoria local y remota.

Aunque la descripción anterior puede contener datos específicos, no deben interpretarse como limitantes de las reivindicaciones de ninguna manera. Otras configuraciones de las realizaciones descritas de la invención son parte del ámbito de esta invención como se define en las reivindicaciones adjuntas.

REIVINDICACIONES

1. Un método para gestionar recursos en un entorno de informático local y un entorno informático bajo demanda, el método incluye la detección de un evento asociado con un entorno informático local (104) que comprende un primer conjunto de nodos informáticos (130, 132, 134) y gestionado por un primer módulo de gestión (108) para producir un evento detectado, en el que el método se caracteriza por:

basándose en el evento detectado, identificar información sobre el entorno informático local (104);

establecer una comunicación con un entorno informático bajo demanda (102) que comprende un segundo conjunto de nodos informáticos (124, 126) diferentes del primer conjunto de nodos informáticos y gestionado por un segundo módulo de gestión (106) diferente del primer módulo de gestión y que transmita la información sobre el entorno informático local (104) al entorno informático bajo demanda (102);

en una disposición creada previamente donde existe una pluralidad de perfiles para la selección, seleccionar un perfil de la pluralidad de perfiles, el perfil relacionado con la carga de trabajo que se puede procesar, para producir un perfil seleccionado;

transferir el perfil seleccionado al entorno informático bajo demanda;

aprovisionar recursos (118) según lo indicado por el segundo módulo de gestión (106) dentro del entorno informático bajo demanda (102) basado en el perfil seleccionado para producir recursos aprovisionados; y

transferir la carga de trabajo desde el entorno informático local (104) al entorno informático bajo demanda (102) en el que la carga de trabajo transferida consume los recursos aprovisionados (124, 126).

2. El método de la reivindicación 1, en el que el evento es al menos uno de los pasos de un umbral o un evento desencadenante.

3. El método de la reivindicación 1, en el que la información sobre el entorno informático local está relacionada con al menos uno de: hardware y un sistema operativo.

4. El método de la reivindicación 1, en donde el establecimiento de la comunicación con el entorno informático bajo demanda y la transmisión de la información sobre el entorno informático local al entorno informático bajo demanda se realiza automáticamente.

5. El método de la reivindicación 1, en donde el evento es un paso de un solo clic realizado por un usuario.

6. El método de la reivindicación 1, en donde el aprovisionamiento de recursos dentro del entorno informático bajo demanda comprende además:

si un recurso en particular no puede duplicarse en el entorno informático bajo demanda, entonces, identificar y seleccionar un recurso sustituto.

7. El método de la reivindicación 6, en el que la identificación y selección de un recurso sustituto se lleva a cabo ya sea en el entorno informático bajo demanda o mediante la negociación entre el segundo módulo de gestión en el entorno informático bajo demanda y el primer módulo de gestión en el entorno informático local.

8. El método de la reivindicación 1, incluye además:

identificar un tipo de carga de trabajo para transferir al entorno informático bajo demanda y en el que la transferencia de carga de trabajo desde el entorno informático local al entorno informático bajo demanda comprende además transferir el tipo de carga de trabajo identificado al entorno informático bajo demanda.

9. El método de la reivindicación 8, en donde la transferencia del tipo de carga de trabajo identificado al entorno informático bajo demanda se basa en diferentes capacidades de hardware y/o software entre el entorno informático bajo demanda y el entorno informático local.

10. El método de la reivindicación 1, en donde el aprovisionamiento de recursos dentro del entorno informático bajo demanda comprende además crear un clúster privado virtual en el entorno informático bajo demanda.

11. El método de la reivindicación 10, en el que el clúster privado virtual comprende además un agregado de recursos informáticos en el entorno informático bajo demanda que tiene particiones y en donde sólo los recursos particionados son accesibles por el entorno informático local.

12. Un medio legible por ordenador (240, 250, 260) que almacena instrucciones informáticas para implementar las etapas del método de una cualquiera de las reivindicaciones 1 a 11.

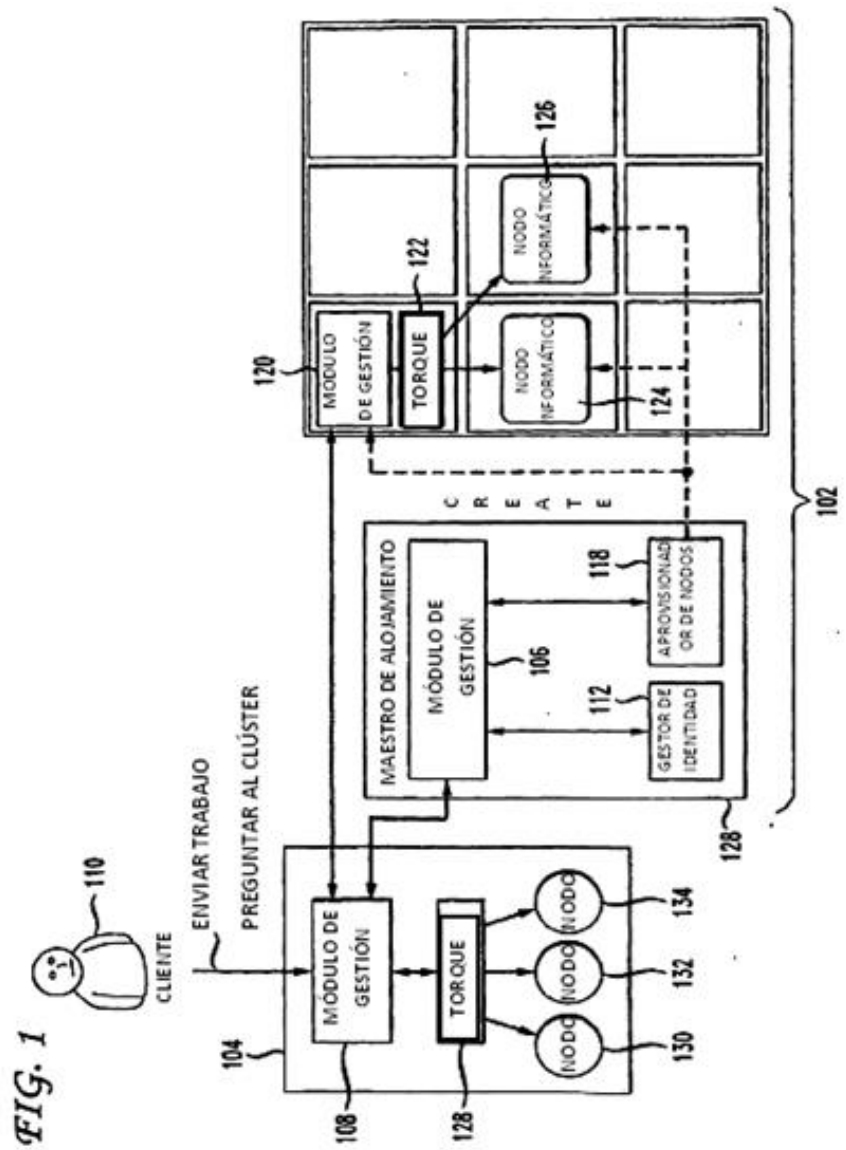

110
CLIENTE
ENVIAR TRABAJO
PREGUNTAR AL CLÚSTER
104
MÓDULO DE GESTIÓN
TORQUE
NODO
NODO
NODO
108
128
134
132
130
MAESTRO DE ALOJAMIENTO
MÓDULO DE GESTIÓN
106
APROVISIONADOR DE NODOS
118
GESTOR DE IDENTIDAD
112
CREATE
128
102
120
MÓDULO DE GESTIÓN
TORQUE
122
NODO INFORMÁTICO
NODO INFORMÁTICO
126
124

FIG. 1

FIG. 2

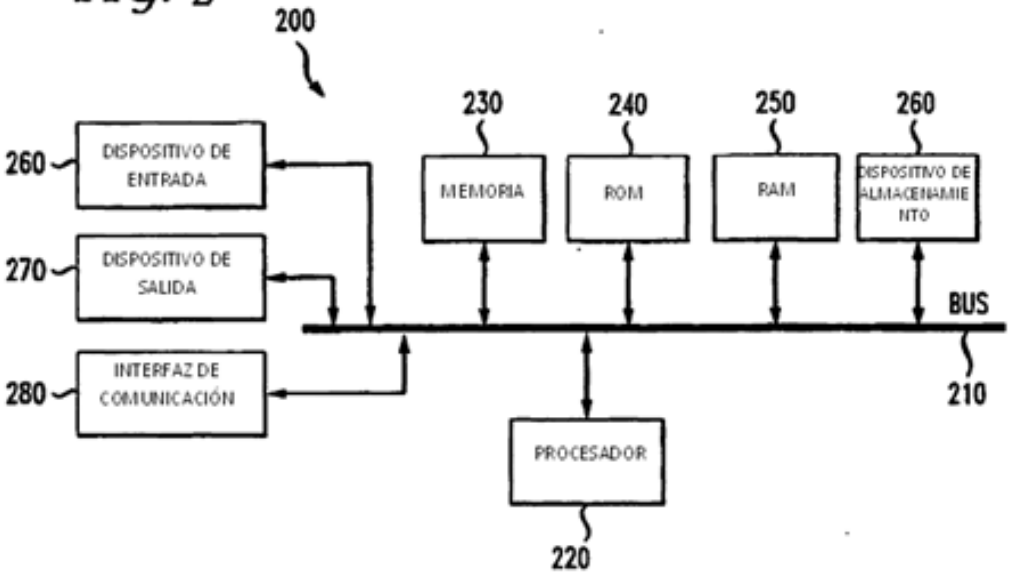

200
230
240
250
260
260
DISPOSITIVO DE ENTRADA
MEMORIA
ROM
RAM
DISPOSITIVO DE ALMACENAMIENTO
270
DISPOSITIVO DE SALIDA
BUS
280
INTERFAZ DE COMUNICACIÓN
210
PROCESADOR
220

FIG. 3

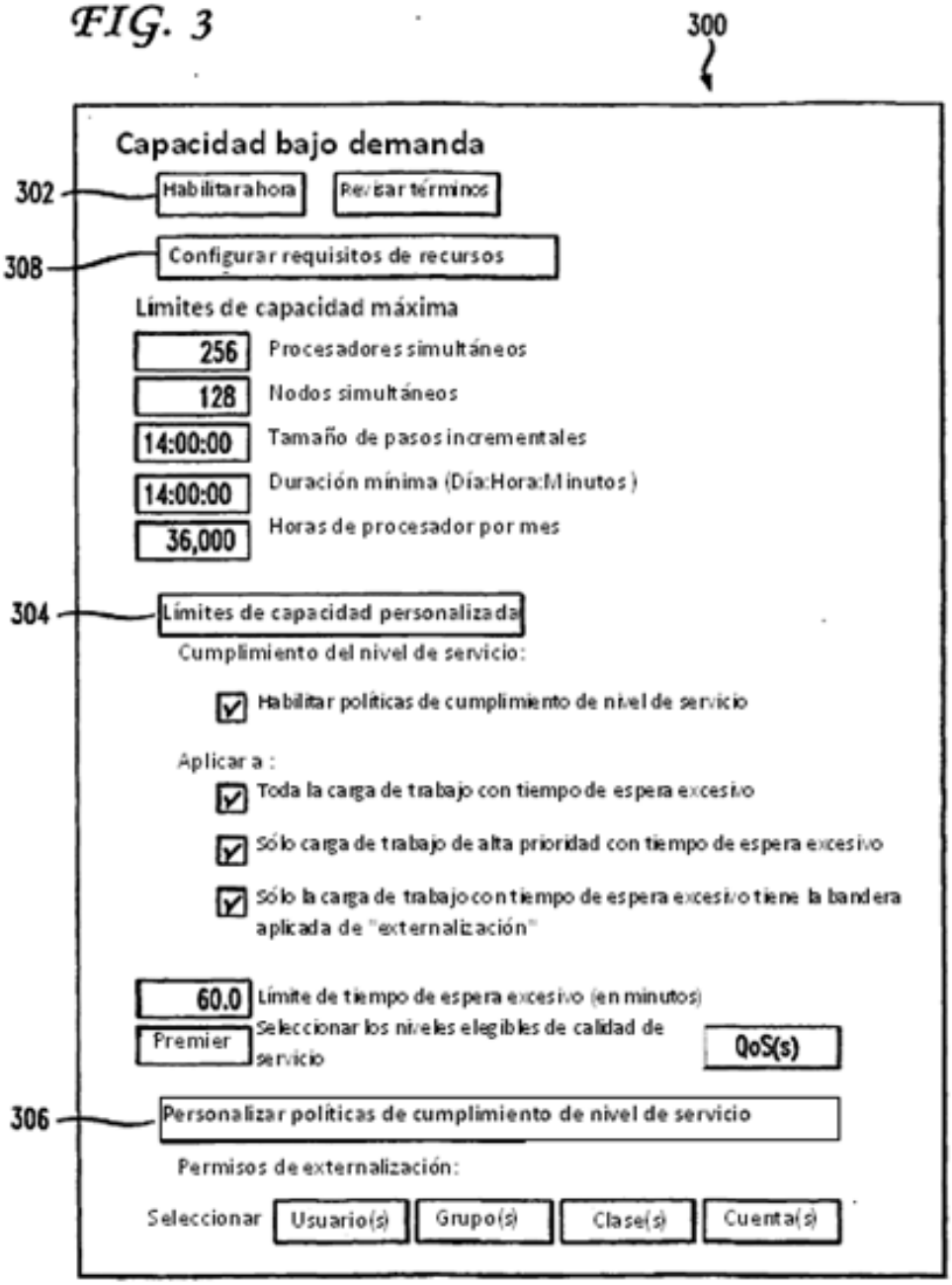

300
Capacidad bajo demanda
302
Habilitar ahora
Revisar términos
308
Configurar requisitos de recursos
Límites de capacidad máxima
256
Procesadores simultáneos
128
Nodos simultáneos
14:00:00
Tamaño de pasos incrementales
14:00:00
Duración mínima (Día:Hora:Minutos)
36,000
Horas de procesador por mes
304
Límites de capacidad personalizada
Cumplimiento del nivel de servicio:
Habilitar políticas de cumplimiento de nivel de servicio
Aplicar a:
Toda la carga de trabajo con tiempo de espera excesivo
Sólo carga de trabajo de alta prioridad con tiempo de espera excesivo
Sólo la carga de trabajo con tiempo de espera excesivo tiene la bandera aplicada de "externalización"
60.0
Límite de tiempo de espera excesivo (en minutos)
Premier
Seleccionar los niveles elegibles de calidad de servicio
QoS(s)
306
Personalizar políticas de cumplimiento de nivel de servicio
Permisos de externalización:
Seleccionar
Usuario(s)
Grupo(s)
Clase(s)
Cuenta(s)

FIG. 4

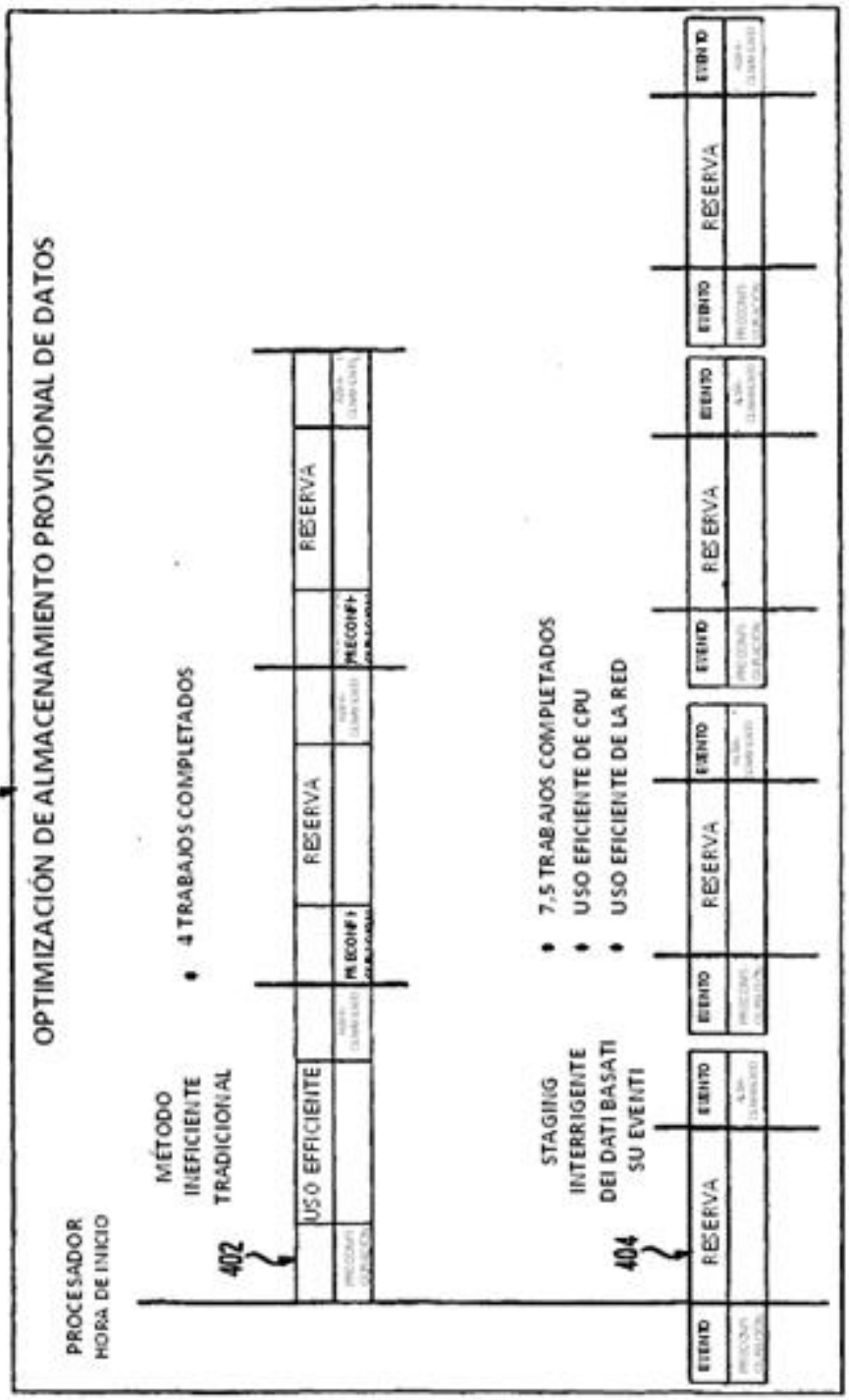

400
OPTIMIZACIÓN DE ALMACENAMIENTO PROVISIONAL DE DATOS
PROCESADOR
HORA DE INICIO
MÉTODO INEFICIENTE TRADICIONAL
• 4 TRABAJOS COMPLETADOS
402
USO EFICIENTE
RESERVA
RESERVA
STAGING INTERRIGENTE DEI DATI BASATI SU EVENTI
• 7,5 TRABAJOS COMPLETADOS
• USO EFICIENTE DE CPU
• USO EFICIENTE DE LA RED
404
EVENTO
RESERVA
EVENTO
RESERVA
EVENTO
RESERVA
EVENTO
RESERVA
EVENTO

FIG. 5

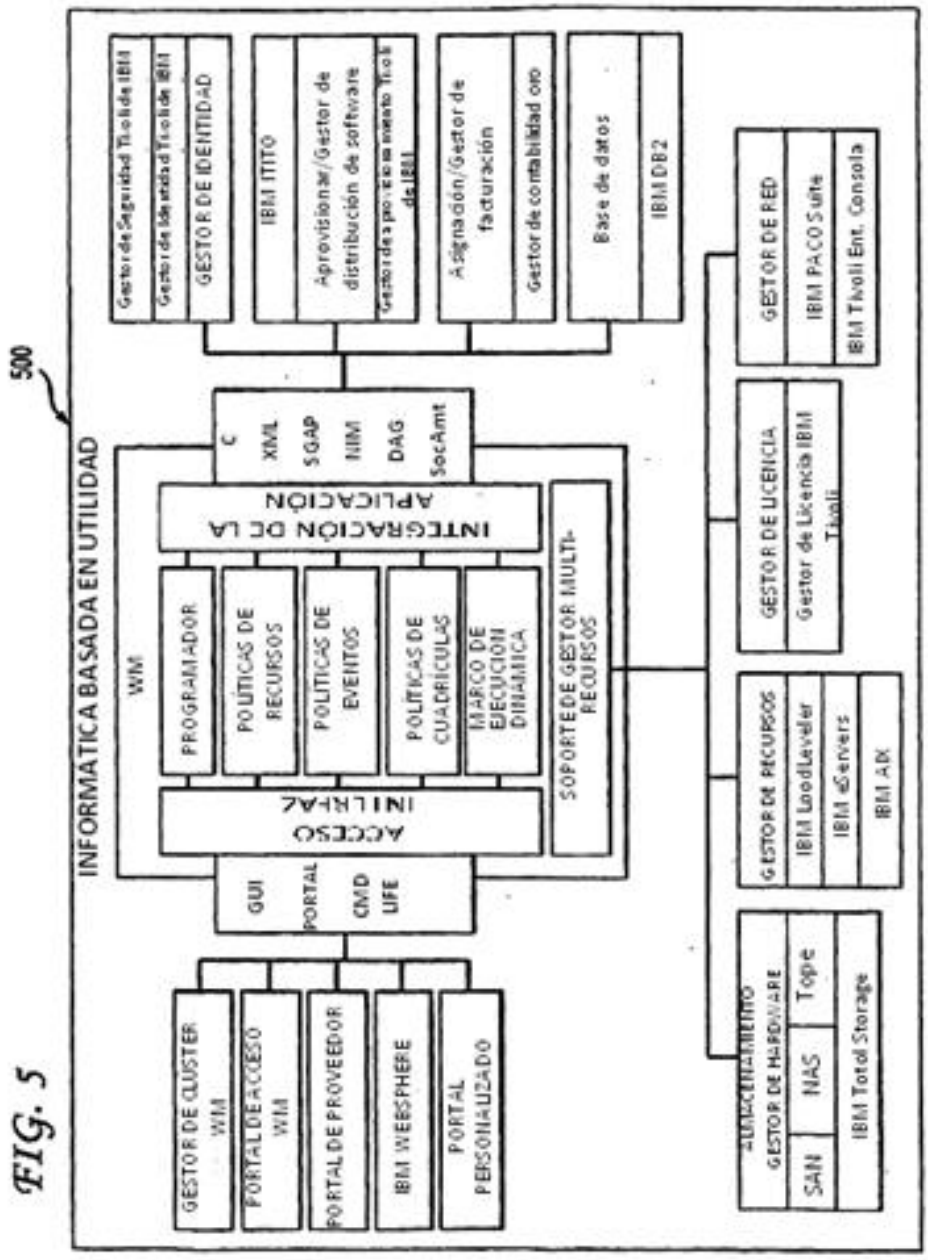
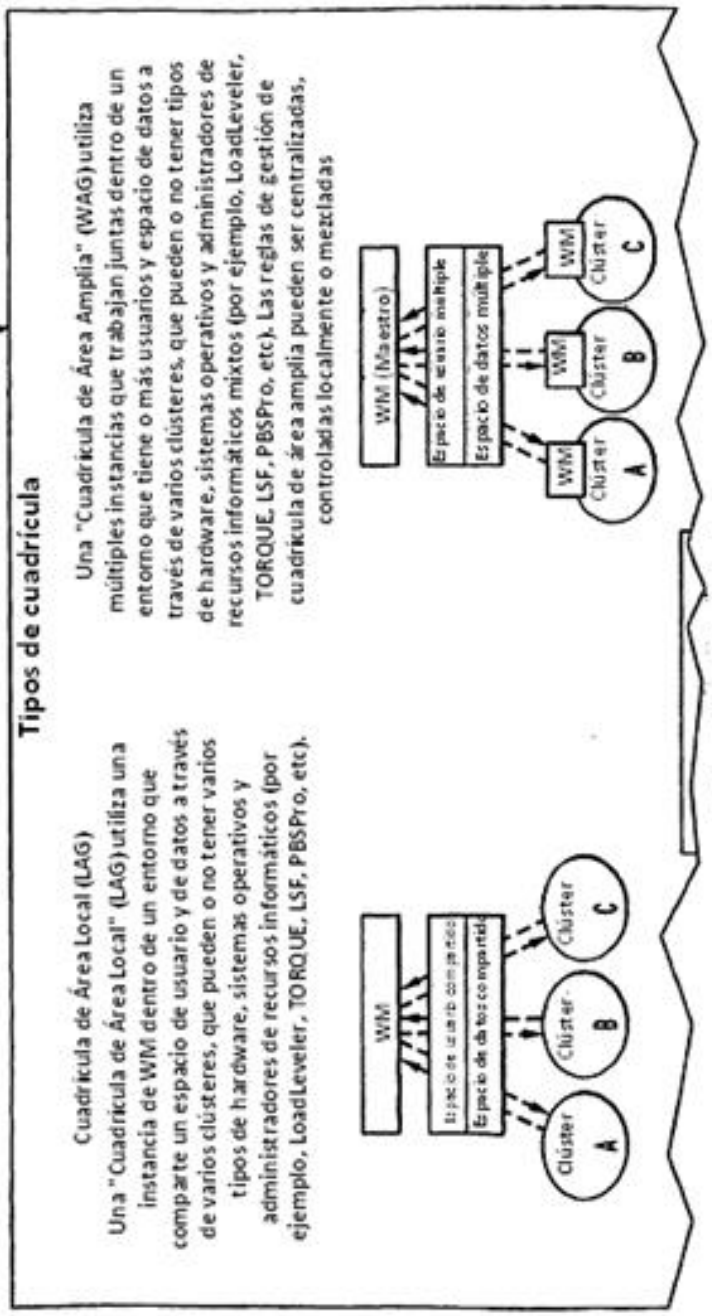

500
INFORMATICA BASADA EN UTILIDAD
Gestor de Seguridad Tivoli de IBM
Gestor de Identidad Tivoli de IBM
GESTOR DE IDENTIDAD
IBM ITITO
Aprovisionar/Gestor de distribución de software
Gestor de gestión de sistemas Tivoli de IBM
Asignación/Gestor de facturación
Gestor de contabilidad oro
Base de datos
IBM DB2
GESTOR DE RED
IBM PACO Suite
IBM Tivoli Ent. Consola
GESTOR DE LICENCIA
Gestor de Licencia IBM Tivoli
GESTOR DE RECURSOS
IBM LoadLeveler
IBM eServers
IBM AIX
ALMACENAMIENTO GESTOR DE HARDWARE
SAN
NAS
Tape
IBM Total Storage
C
XML
SOAP
IIM
DAG
SecAmt
INTEGRACIÓN DE LA APLICACIÓN
WM
PROGRAMADOR
POLÍTICAS DE RECURSOS
POLÍTICAS DE EVENTOS
POLÍTICAS DE CUADRÍCULAS
MARCO DE EJECUCIÓN DINÁMICA
SOPORTE DE GESTOR MULTI-RECURSOS
ACCESO INTERFAZ
GUI
PORTAL
CMD LIFE
GESTOR DE CLUSTER WM
PORTAL DE ACCESO WM
PORTAL DE PROVEEDOR
IBM WEBSPHERE
PORTAL PERSONALIZADO

FIG. 6



A Fig. 6 (continuada)

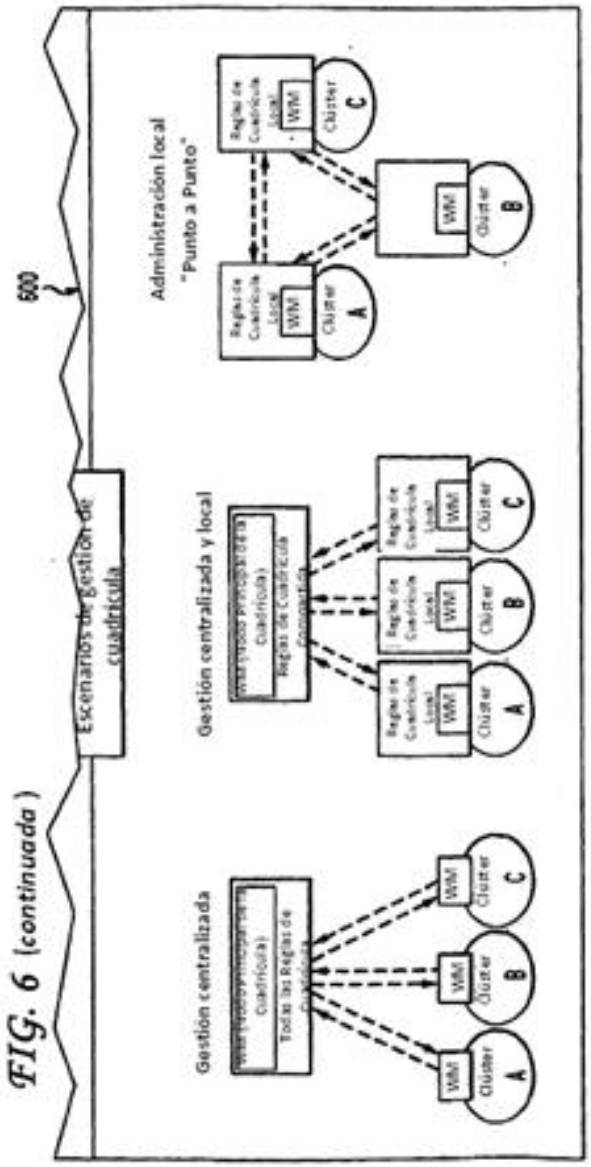


FIG. 6 (continuada)

FIG. 7

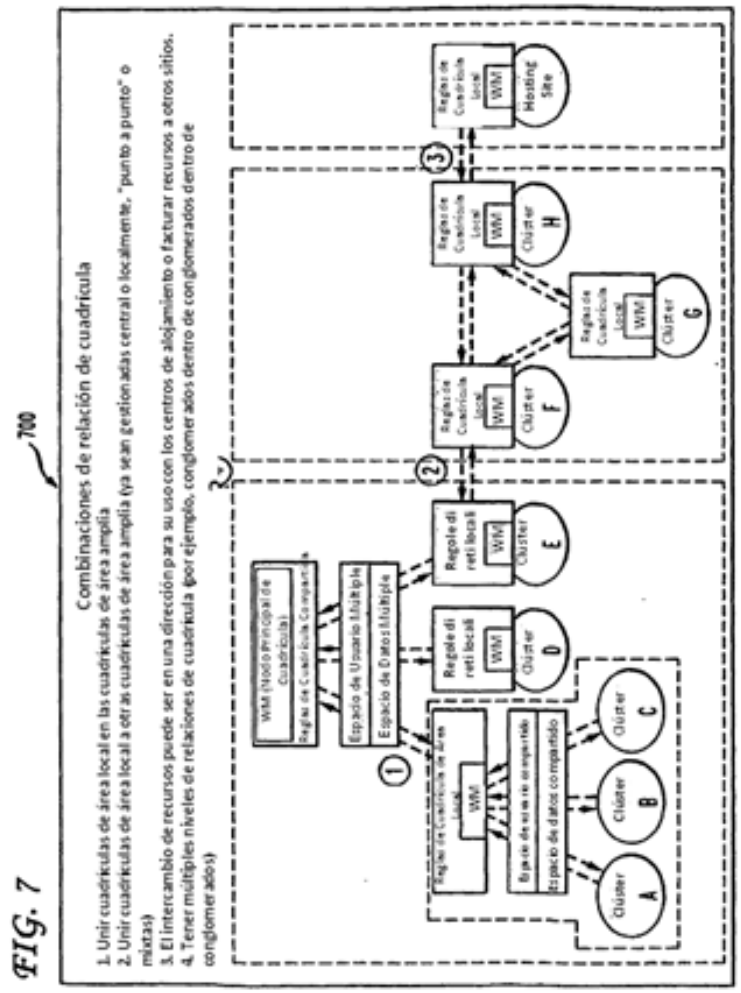

100
Combinaciones de relación de cuadrícula
1. Unir cuadrículas de área local en las cuadrículas de área amplia
2. Unir cuadrículas de área local a otras cuadrículas de área amplia (ya sean gestionadas central o localmente, "punto a punto" o mixtas)
3. El intercambio de recursos puede ser en una dirección para su uso con los centros de alojamiento o facturar recursos a otros sitios.
4. Tener múltiples niveles de relaciones de cuadrícula (por ejemplo, conglomerados dentro de conglomerados)
WM (Nodo Principal de Cuadrícula)
Regla de Cuadrícula Compartida
Espacio de Usuario Múltiple
Espacio de Datos Múltiple
Regla de Cuadrícula Local
WM
Espacio de usuario compartido
Espacio de datos compartido
Regla de reti local
Clúster A
Clúster B
Clúster C
Clúster D
Clúster E
Clúster F
Clúster G
Clúster H
Hosting Site
1
2
3

FIG. 8

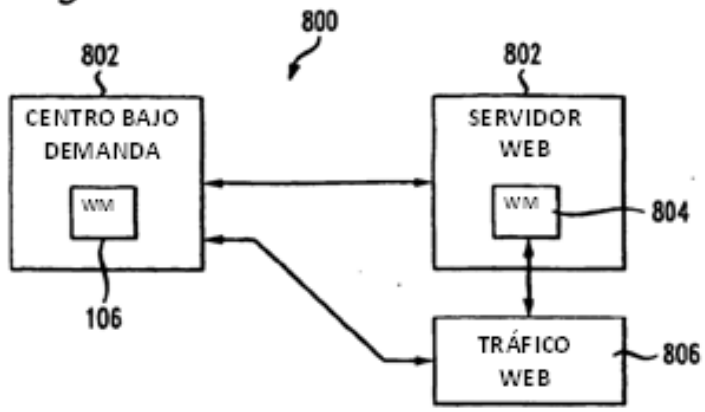
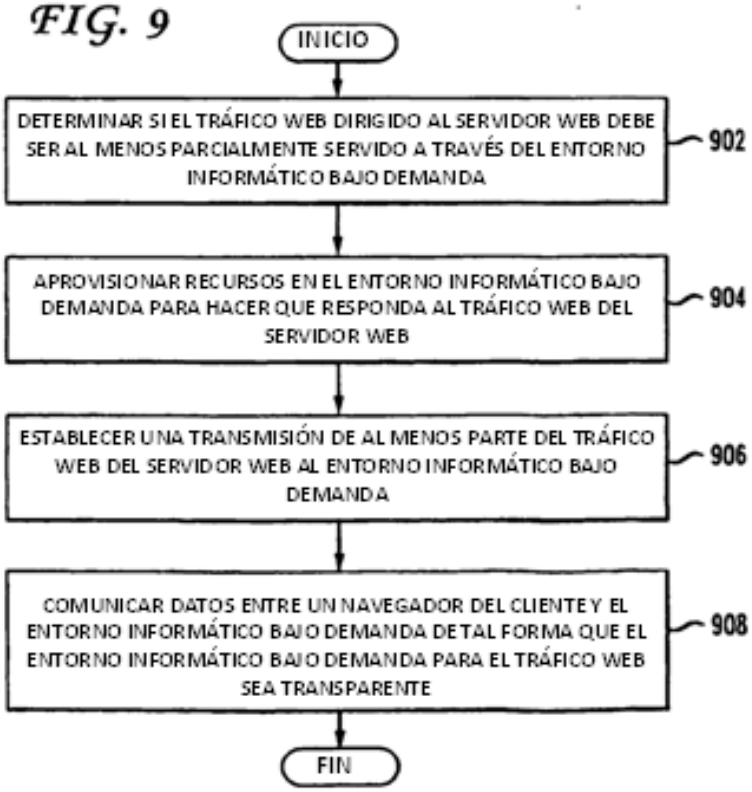

800
802
802
CENTRO BAJO DEMANDA
WM
106
SERVIDOR WEB
WM
804
TRÁFICO WEB
806

FIG. 9


INICIO
DETERMINAR SI EL TRÁFICO WEB DIRIGIDO AL SERVIDOR WEB DEBE SER AL MENOS PARCIALMENTE SERVIDO A TRAVÉS DEL ENTORNO INFORMÁTICO BAJO DEMANDA
902
APROVISIONAR RECURSOS EN EL ENTORNO INFORMÁTICO BAJO DEMANDA PARA HACER QUE RESPONDA AL TRÁFICO WEB DEL SERVIDOR WEB
904
ESTABLECER UNA TRANSMISIÓN DE AL MENOS PARTE DEL TRÁFICO WEB DEL SERVIDOR WEB AL ENTORNO INFORMÁTICO BAJO DEMANDA
906
COMUNICAR DATOS ENTRE UN NAVEGADOR DEL CLIENTE Y EL ENTORNO INFORMÁTICO BAJO DEMANDA DE TAL FORMA QUE EL ENTORNO INFORMÁTICO BAJO DEMANDA PARA EL TRÁFICO WEB SEA TRANSPARENTE
908
FIN