(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 585 178**

(51) Int. CI.:

G06F 9/50 (2006.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(96) Fecha de presentación y número de la solicitud europea: **18.12.2003 E 03780885 (4)**

(97) Fecha y número de publicación de la concesión europea: **06.04.2016 EP 1696324**

(54) Título: **Sistema servidor/cliente, dispositivo de distribución de carga, procedimiento de distribución de carga y programa de distribución de carga**

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
04.10.2016

(73) Titular/es:

G-CLUSTER GLOBAL CORPORATION (100.0%)
4-14, Akasaka 8-chome, Minato-ku
Tokyo 107-0052, JP

(72) Inventor/es:

HASHIMOTO, TARO y
ONODA, TETSUYA

(74) Agente/Representante:

PONTI SALES, Adelaida

DESCRIPCIÓN

Sistema servidor/cliente, dispositivo de distribución de carga, procedimiento de distribución de carga y programa de distribución de carga

CAMPO TÉCNICO

[0001] La presente invención se refiere a un sistema servidor/cliente según la parte de preámbulo de la reivindicación 1 tal como se describe en el documento EP0384339A2, un procedimiento de distribución de carga según la reivindicación 3 y un programa de distribución de carga que distribuyen accesos a servidores, y más específicamente a un sistema servidor/cliente, un procedimiento de distribución de carga y un programa de distribución de carga que seleccionan el servidor óptimo de entre una pluralidad de servidores monitorizando y evaluando un estado de funcionamiento de los servidores.

TÉCNICA ANTECEDENTE

[0002] Convencionalmente, se utiliza un sistema de tipo servidor-cliente en el cual se efectúa una petición de ejecución de un programa servidor, a un servidor, mediante un programa de aplicación en un cliente a través de una red. En este sistema, en respuesta a la petición de ejecutar un programa procedente de un cliente, el servidor que proporciona un servicio como resultado de la ejecución del programa da inicio a un proceso que es una unidad de procesamiento del programa para proporcionar el servicio. Cuando un proceso de un programa de distribución de contenido web, un software de operaciones comerciales, un software de juego, etc. es iniciado en el servidor, se proporciona un servicio en forma de, por ejemplo, datos de texto, datos de imagen, etc. al cliente que ha efectuado la petición al servicio.

[0003] Cuando un procesamiento en tal sistema servidor/cliente es ejecutado por una pluralidad de servidores, se utiliza una técnica que distribuye cargas de un procesamiento solicitado desde el cliente. Por ejemplo, en un caso de servidores web que distribuyen contenido web, los servidores que han de ser asignados para ejecutar un procesamiento se seleccionan utilizando esquemas tales como un esquema de asignación cíclica que comparte el procesamiento de accesos desde clientes secuencialmente a servidores preparados y un esquema de conexión mínima que selecciona de los servidores el servidor que tiene el mínimo número de sesiones.

[0004] Cuando la carga es distribuida en el esquema de asignación cíclica o el esquema de conexión mínima, los servidores deben ser asignados independientemente de la cantidad de recursos (recursos guardados por un ordenador personal tal como una CPU, una memoria, etc.) que han de ser consumidos por la ejecución del proceso o la cantidad de recursos guardada por los servidores.

[0005] El documento EP0384339A2 describe un mecanismo intermediario en una red informática para asignar una pluralidad de servidores, que tienen cada uno una capacidad de recursos disponible, a una pluralidad de clientes para suministrar uno de varios servicios al cliente. El intermediario recibe peticiones de clientes de los servicios y sugiere uno de los servidores basándose en una política de red y la capacidad de recursos disponible.

[0006] La fig. 19 es una vista explicativa que muestra un ejemplo de una relación entre una cantidad de recursos guardados por un servidor y una cantidad de recursos consumidos por un proceso. En la fig. 19, la coordenada X representa una cantidad de CPU del servidor y la coordenada Y representa una cantidad de memoria del servidor que son consumidas para ejecutar un programa del servidor. “Xmax” en la coordenada X mostrada en la fig. 19 representa el “100 %” que es el valor máximo del consumo de la CPU, e “Ymax” en la coordenada Y representa el “100 %” que es el valor máximo del consumo de la memoria. Una posición ideal en la cual la CPU y la memoria son consumidas respectivamente al 100 % está representada por las coordenadas 1902. Cuando un proceso 1901 que consume la CPU y la memoria es ejecutado en este servidor, el consumo de la memoria se vuelve el 100 % aun cuando el consumo de la CPU sea inferior al 100 %.

[0007] La fig. 20 es una vista explicativa que muestra otro ejemplo de la relación entre la cantidad de recursos guardados por un servidor y la cantidad de recursos consumidos por un proceso. En la fig. 20, al igual que la fig. 19, la coordenada X representa una cantidad de CPU del servidor y la coordenada Y representa una cantidad de memoria del servidor que son consumidas para ejecutar un programa servidor. “Xmax” en la coordenada X representa el “100 %” que es el valor máximo del consumo de la CPU, e “Ymax” en la coordenada Y representa el “100 %” que es el valor máximo del consumo de la memoria. Una posición ideal en la cual la CPU y la memoria son consumidas respectivamente al 100 % está representada por las coordenadas 2002. Cuando un proceso 2001 que

consume la CPU y la memoria es ejecutado en este servidor, al contrario que el caso mostrado en la fig. 19, el consumo de la CPU se vuelve el 100 % aun cuando el consumo de la memoria sea inferior al 100 %.

[0008] Existe un sistema tal que un dispositivo de distribución de carga toma el número de sesiones en cada uno de los servidores, y selecciona un servidor adecuado de entre los servidores para asignar una sesión basándose en los números de sesiones y los pesos determinados respectivamente basándose en un rendimiento de la máquina de cada uno de los servidores (por ejemplo, la publicación de solicitud de patente japonesa abierta a consulta por el público Nº 2002-269061).

[0009] Sin embargo, en la distribución de carga según el esquema de asignación cíclica o el esquema de conexión mínima descritos anteriormente, la carga es distribuida basándose en el número de sesiones con clientes de la misma independientemente de los recursos guardados por los servidores y los recursos que han de ser consumidos por el proceso. Así, la suficiencia del recurso se juzga cuando el proceso que se ha solicitado que sea ejecutado es iniciado en el servidor. Como resultado, se produce retardo al responder a la petición, y se produce un tiempo de latencia en el cliente.

[0010] Por otra parte, como la cantidad de recursos que son guardados por cada servidor difiere según un rendimiento de cada dispositivo que ha de utilizarse como servidor, cantidades de recursos utilizados después de distribuir la carga no se consideran cuando se utiliza un esquema de distribución de carga basado simplemente en los números de sesiones tal como el esquema de asignación cíclica, el esquema de conexión mínima, etc. Por lo tanto, un proceso puede ser asignado a un servidor al que le queda una pequeña cantidad de recursos o a un servidor que tiene un consumo desequilibrado de los recursos.

[0011] Además, cuando se hace que un servidor en el que el consumo está concentrado o bien en la CPU o bien en la memoria ejecute un proceso, no puede ser ejecutado un nuevo proceso aun cuando un recurso que tenga un menor consumo tenga una región vacía porque sólo una región de recursos que tenga un mayor consumo no es suficiente. Por lo tanto, tal servidor que tiene un consumo desequilibrado de recursos puede ejecutar menos procesos comparado con un servidor en el que los recursos son consumidos por igual.

[0012] Un objeto de la presente invención es solucionar los problemas de las técnicas convencionales descritas anteriormente, y proporcionar un sistema servidor/cliente, un procedimiento de distribución de carga y un programa de distribución de carga que pueda seleccionar un servidor óptimo de entre una pluralidad de servidores instalados en uno o más lugares evaluando numéricamente los recursos y los estados operacionales de los servidores que han de ser asignados a un proceso, y pueda hacer que cada uno de los servidores ejecute el proceso eficientemente.

DESCRIPCIÓN DE LA INVENCIÓN

[0013] Para solucionar los problemas anteriores, un sistema servidor/cliente según la presente invención incluye las características expuestas en la reivindicación 1.

[0014] Además, un procedimiento de distribución de carga según la presente invención es un procedimiento de distribución de cargas de servidores utilizados en un sistema servidor/cliente en el cual una pluralidad de servidores y una pluralidad de clientes están conectados a través de una red, y los servidores ejecutan un proceso basándose en una petición de proceso procedente del cliente y transmiten un resultado de proceso al cliente. El procedimiento de distribución de carga incluye las etapas enumeradas en la reivindicación 3.

[0015] Por otra parte, un programa de distribución de carga según la presente invención incluye las etapas de programa definidas en la reivindicación 5.

[0016] La unidad (etapa) de determinación incluye una primera unidad (etapa) de cálculo configurada para calcular, para cada uno de los servidores, una primera distancia desde un punto de estimación que indica un consumo estimado hasta una línea de consumo ideal, el consumo estimado obtenido añadiendo una cantidad de recursos que han de ser consumidos por la ejecución del proceso a un punto que indica una cantidad de recursos que han sido consumidos por cada uno de los servidores, siendo la línea de consumo ideal una línea recta que conecta un origen y un punto que indica una capacidad de recursos máxima de cada uno de los servidores expresada en un espacio que tiene parámetros de recursos como ejes; y una segunda unidad (etapa) de cálculo de distancia configurada para calcular, para cada uno de los servidores, una segunda distancia desde el punto estimado hasta el origen en el espacio. La unidad (etapa) de determinación está configurada (incluye) para

determinar el servidor basándose en al menos una de la primera distancia y la segunda distancia.

[0017] Además, los parámetros pueden incluir al menos una de una cantidad de carga de una unidad de procesamiento central, una cantidad de carga de una memoria de sistema, una cantidad de carga de una unidad de procesamiento gráfico, una cantidad de carga de una memoria de vídeo, y una cantidad de carga de una tarjeta de interfaz de red.

BREVE DESCRIPCIÓN DE LOS DIBUJOS

[0018] La fig. 1 es un diagrama esquemático de una configuración de sistema de un sistema servidor/cliente según una realización de la presente invención; la fig. 2 es un diagrama de bloques de una configuración de hardware de un dispositivo de distribución de carga, un dispositivo terminal, y un servidor del sistema servidor/cliente según la realización de la presente invención; la fig. 3 es un diagrama de bloques de una configuración de hardware de un teléfono móvil del sistema servidor/cliente según la realización de la presente invención; la fig. 4 es un diagrama de bloques de una configuración funcional del sistema servidor/cliente según la realización de la presente invención; la fig. 5 es una vista explicativa que muestra una estructura de un procedimiento de un procedimiento de distribución de carga según la realización de la presente invención; la fig. 6 es una vista explicativa que muestra un ejemplo de una asignación de procesos ideal según la realización de la presente invención; la fig. 7 es una vista explicativa que muestra otro ejemplo de la asignación de procesos ideal según la realización de la presente invención; la fig. 8 es una vista explicativa que muestra un ejemplo de un consumo de recursos óptimo en el servidor del sistema servidor/cliente según la realización de la presente invención; la fig. 9 es una vista explicativa que muestra otro ejemplo del consumo de recursos óptimo en el servidor del sistema servidor/cliente según la realización de la presente invención; la fig. 10 es una vista explicativa que muestra otro ejemplo del consumo de recursos óptimo en el servidor del sistema servidor/cliente según la realización de la presente invención; la fig. 11 es una vista explicativa que muestra un ejemplo de un consumo de recursos según la realización de la presente invención; la fig. 12 es una vista explicativa que muestra una estructura de un procedimiento de evaluación de los servidores según la realización de la presente invención; la fig. 13 es una vista explicativa que muestra un ejemplo de una función para un procedimiento de evaluación de un servidor que ha de ser asignado según la realización de la presente invención; la fig. 14 es un diagrama de flujo que muestra un ejemplo de un procedimiento de evaluación y determinación de un servidor según la realización de la presente invención; la fig. 15 es un esquema que muestra otro ejemplo del procedimiento de evaluación y determinación del servidor según la realización de la presente invención; la fig. 16 es una vista explicativa que muestra un ejemplo de una tabla que contiene información sobre un proceso según la realización de la presente invención; la fig. 17 es una vista explicativa que muestra un ejemplo de una tabla que contiene información sobre los servidores del sistema servidor/cliente según la realización de la presente invención; la fig. 18 es una vista explicativa que muestra un ejemplo de una distribución regional de los servidores del sistema servidor/cliente según la presente invención; la fig. 19 es una vista explicativa que muestra un ejemplo de una relación entre una cantidad de recursos guardados por un servidor y una cantidad de recursos consumidos por un proceso; y la fig. 20 es una vista explicativa que muestra otro ejemplo de la relación entre una cantidad de recursos guardados por un servidor y una cantidad de recursos consumidos por un proceso.

MEJOR(ES) MODO(S) DE LLEVAR A CABO LA INVENCIÓN

[0019] Más adelante se describirán en detalle realizaciones ejemplares de un sistema servidor/cliente, un procedimiento de distribución de carga y un programa de distribución de carga según la presente invención con referencia a los dibujos adjuntos.

(Configuración de sistema)

[0020] Se describirá la configuración de sistema de un sistema servidor/cliente según una realización de la presente invención que incluye un dispositivo de distribución de carga. La fig. 1 es un diagrama esquemático de una configuración de sistema de un sistema servidor/cliente según una realización de la presente invención. Tal como se muestra en la fig. 1, los servidores 101a a 101n están configurados para estar conectados a través de una red 100 tal como Internet, respectivamente con un dispositivo terminal (cliente) 102 o una oficina telefónica 105. La oficina telefónica 105 está conectada a una estación de base radioeléctrica 104 y un teléfono móvil 103 está conectado a través de esta estación de base radioeléctrica 104 con los servidores 101a a 101n.

[0021] Los servidores 101a a 101n son, por ejemplo, ordenadores personales utilizados para un uso como servidores denominados servidores PC, y son administrados y operados por una entidad comercial de servicios que realiza la distribución de contenido web y la distribución de juegos en red. Un servidor PC es alterado como un

dispositivo de distribución de carga que tiene una función de distribución de carga, instalando software de aplicación dedicada en el servidor PC. Uno o más servidores que tienen una función de distribución de carga pueden estar provistos en un sistema entero. El servidor 101a se describirá más adelante como un dispositivo de distribución de carga.

[0022] Por lo tanto, los servidores (servidores de procesamiento de procesos) 101b a 101n distintos del servidor 101a también pueden ser alterados como dispositivos de distribución de carga que tienen respectivamente una función de distribución de carga instalando software de aplicación dedicada en esos servidores 101b a 101n y, cuando el servidor 101a que ha sido puesto en marcha como un dispositivo de distribución de carga 101a detiene el funcionamiento del mismo debido a una avería, se puede cambiar a uno cualquiera de los servidores 101b a 101n y puede sustituir al dispositivo de distribución de carga 101a.

[0023] Cuando una petición de ejecución de un proceso es enviada desde el dispositivo terminal 102 o el teléfono móvil 103 descrito más adelante que es utilizado por un usuario, el dispositivo de distribución de carga 101a asigna la ejecución del proceso a los servidores 101b a 101n. Los servidores 101b a 101n ejecutan el proceso solicitado y proporcionan servicios a través de la red 100, como resultado de la ejecución, al dispositivo terminal 102 o el teléfono móvil 103 descrito más adelante.

[0024] El dispositivo terminal 102 es, por ejemplo, un dispositivo terminal de información, tal como un ordenador personal, etc., y es utilizado por un individuo o en una compañía. El teléfono móvil 103 es, por ejemplo, un teléfono móvil que está equipado con una función de comunicación de información, tal como “i-mode (R)” de NTT DoCoMo Inc. o “EZweb (R)” de au, que puede acceder al menos a Internet y puede acceder a la red 100 utilizando un navegador o una aplicación Java (R). La estación de base radioeléctrica 104 convierte los datos de comunicación recibidos desde la oficina telefónica 105 en una onda radioeléctrica y transmite la onda radioeléctrica al teléfono móvil 103, y transmite una onda radioeléctrica recibida desde el teléfono móvil 103 a la oficina telefónica 105. La oficina telefónica 105 conmuta las líneas cuando la comunicación de datos es ejecutada entre el teléfono móvil 103, y el dispositivo de distribución de carga 101a y los servidores 101b a 101n.

[0025] El sistema para el teléfono móvil 103 puede considerarse como un sistema para el denominado “servicio de punto de acceso inalámbrico” cuando se supone que el teléfono móvil es sustituido por un ordenador portátil o un PDA provisto de/que contiene un adaptador LAN radioeléctrico, y la estación de base radioeléctrica 104 del teléfono móvil es sustituida por una estación de base LAN radioeléctrica.

[0026] En la comunicación ejecutada entre el dispositivo de distribución de carga 101a, los servidores 101b a 101n, el dispositivo terminal 102, o el teléfono móvil 103 a través de la red 100, es preferible que la protección de confidencialidad se asegure utilizando una función de seguridad que emplee un esquema SSL, etc., o una técnica de cifrado, etc.

(Configuración de hardware)

[0027] Se describirán las configuraciones de hardware del dispositivo terminal y los servidores según la realización de la presente invención. La fig. 2 es un diagrama de bloques de una configuración de hardware del dispositivo de distribución de carga, el dispositivo terminal y los servidores del sistema servidor/cliente según la realización de la presente invención.

[0028] Tal como se muestra en la fig. 2, el dispositivo terminal 102, el dispositivo de distribución de carga 101a y los servidores 101b a 101n incluyen respectivamente una CPU 201, una ROM 202, una RAM 203, una HDD (unidad de disco duro) 204, un HD (disco duro) 205, una FDD (unidad de disco flexible) 206, un FD (disco flexible) 207 como ejemplo de un medio de grabación extraíble, una pantalla 208, una I/F (interfaz) de red 209, un teclado 210, un ratón 211, una impresora 212, una unidad de CD-ROM 213, un CD-ROM 214 como ejemplo de un medio de grabación extraíble 214, y un altavoz 215. Cada componente está conectado unos con otros por un bus 200.

[0029] La CPU 201 administra el control de todo el dispositivo terminal 102, todo el dispositivo de distribución de carga 101a y todos los servidores 101b a 101n. La ROM 202 almacena programas tales como los programas básicos de entrada/salida, un programa de arranque, etc. La RAM 203 se utiliza como área de trabajo de la CPU 201. La HDD 204 controla la lectura/escritura de datos del/en el HD 205 según un control de la CPU 201.El HD 205 almacena datos escritos según un control de la HDD 204.

[0030] La FDD 206 controla la lectura/escritura de datos del/en el FD 207 según el control de la CPU 201. El

FD 207 almacena los datos escritos por el control de la FDD 206. Como medio de grabación extraíble, puede utilizarse un CD-RW, un MO, un DVD (disco versátil digital), etc. además del FD 207. La pantalla 208 muestra un cursor, menús, o ventanas (navegadores) relacionadas con datos tales como textos, imágenes, información funcional, etc., y es un CRT, una pantalla de cristal líquido TFT, una pantalla de plasma, etc.

[0031] La red I/F 209 está conectada a la red 100 y está conectada al dispositivo terminal 102 a través de esta red 100. La red I/F 209 administra la interfaz entre la red 100 y la parte interna del sistema, y controla la entrada/salida de datos del dispositivo de distribución de carga 101a, los servidores 101b a 101n y el dispositivo terminal 102. Por ejemplo, puede emplearse un módem, un adaptador LAN, etc. como la red I/F 209.

[0032] El teclado 210 incluye teclas para introducir letras, valores numéricos, instrucciones varias, etc., y ejecuta la entrada de datos. El ratón 211 ejecuta el movimiento del cursor, la selección de una zona, o el desplazamiento y cambio de tamaño de ventanas. El ratón 211 puede ser una bola de control de cursor, una palanca de control, un mando para juegos, etc., que incluye igualmente la función como dispositivo apuntador. La impresora 212 imprime datos de texto. Por ejemplo, puede emplearse una impresora láser, una impresora de inyección de tinta, etc. como la impresora 212. La unidad de CD-ROM 213 controla la lectura de datos del CD-ROM 214 según el control de la CPU 201. El CD-ROM 214 es un medio de grabación extraíble. El altavoz 215 (incluyendo unos cascos o unos auriculares) produce sonido, música, etc.

[0033] Se describirá una configuración de hardware del teléfono móvil según la realización de la presente invención. La fig. 3 es un diagrama de bloques de la configuración de hardware del teléfono móvil del sistema servidor/cliente según la realización de la presente invención. Tal como se muestra en la fig. 3, el teléfono móvil 103 incluye una CPU 301, una ROM 302, una RAM 303, una pantalla 304, teclas operativas 305, un micrófono 306, un altavoz 307, una unidad de control de comunicación 308, una antena 309, un terminal de conexión externa 310, y un dispositivo de almacenamiento externo 311. La antena 309 está conectada comunicativamente a la estación de base radioeléctrica 104. Cada componente está conectado respectivamente con los otros por un bus 300.

[0034] La CPU 301 administra el control de todo el teléfono móvil 103. La ROM 302 almacena programas tales como los programas básicos de entrada/salida, un programa de arranque, etc. La RAM 303 se utiliza como área de trabajo de la CPU 301. La pantalla 304 es una pantalla de cristal líquido y muestra ventanas (navegadores) relacionados con datos tales como textos, imágenes, información funcional, etc. Las teclas operativas 305 introducen letras, dígitos, instrucciones varias, etc. El micrófono 306 convierte voces de entrada en una señal eléctrica. El altavoz 307 convierte la señal eléctrica de entrada en las voces y produce las voces. La unidad de control de comunicación 308 transmite/recibe ondas electromagnéticas a/desde la estación de base radioeléctrica 104 a través de la antena 309 y realiza un control de la misma. El terminal de conexión externa 310 es un punto de conexión con el dispositivo de almacenamiento externo 311 tal como una memoria flash, etc.

(Configuración funcional)

[0035] Se describirá una configuración funcional del sistema servidor/cliente según la realización de la invención. La fig. 4 es un diagrama de bloques de la configuración funcional del sistema servidor/cliente según la realización de la presente invención. Tal como se muestra en la fig. 4, el dispositivo de distribución de carga 101a que constituye el sistema servidor/cliente incluye una unidad de recepción de información de proceso 401, una unidad de determinación 402 y una unidad de transmisión de información de servidor 403. La unidad de determinación 402 incluye una unidad de cálculo de distancia 404. El cliente 102 que constituye el sistema servidor/cliente incluye una unidad de transmisión de información de proceso 411, una unidad de recepción de información de servidor 412 y una unidad de transmisión de petición de proceso 413.

[0036] La unidad de recepción de información de proceso 401 recibe información sobre el proceso procedente del cliente 102 a través de la red 100. La unidad de determinación 402 determina los servidores que se ha de hacer que procesen el proceso de entre la pluralidad de servidores de procesamiento de procesos 101b a 101n basándose en la información sobre el proceso recibida por la unidad de recepción de información de proceso 401. La unidad de cálculo de distancia 404 calcula distancias entre el consumo de recursos del proceso y una línea recta que conecta el origen y las capacidades disponibles máximas de los parámetros en un espacio que tiene parámetros de recursos como los ejes del mismo. Los servidores que se ha de hacer que ejecuten el proceso se determinan basándose en las distancias calculadas por la unidad de cálculo de distancia 404. La unidad de transmisión de información de servidor 403 transmite la información sobre los servidores determinada por la unidad de determinación 402 al cliente 102.

[0037] Por otra parte, la unidad de transmisión de información de proceso 411 transmite la información sobre el proceso al dispositivo de distribución de carga 101a a través de la red 100 antes de solicitar el proceso. La unidad de recepción de información de servidor 412 recibe a través de la red 100 la información sobre los servidores transmitida por la unidad de transmisión de información de servidor 403 del dispositivo de distribución de carga 101a. La unidad de transmisión de solicitud de proceso 413 transmite información sobre una petición para procesar al proceso a los servidores relacionada con la información recibida por la unidad de recepción de información de servidor 412, es decir, los servidores determinados por la unidad de determinación 402 del dispositivo de distribución de carga 101a (cualquiera de los servidores de procesamiento de procesos 101b a 101n). Los servidores anteriores procesan basándose en esta petición de proceso y transmiten los resultados del procesamiento al cliente 102.

(Procedimiento de procesamiento de distribución de carga)

[0038] Se describirá una estructura de un procedimiento de un procedimiento de distribución de carga según la realización de la presente invención utilizando el dispositivo de distribución de carga 101a. La fig. 5 es una vista explicativa que muestra la estructura del procedimiento de un procedimiento de distribución de carga según la realización de la presente invención utilizando el dispositivo de distribución de carga. Tal como se muestra en la fig. 5, el dispositivo de distribución de carga 101a distribuye la carga para ejecutar un proceso (por ejemplo, un proceso de juego tal como un juego online, etc.) solicitado desde los dispositivos terminales 102a a 102n, y asigna las cargas distribuidas respectivamente a los servidores 101b a 101n.

[0039] Cuando se juega a un juego online utilizando el dispositivo terminal 102a, la ejecución del proceso de juego es solicitada tal como se indica por una flecha 501 desde el dispositivo terminal 102a al dispositivo de distribución de carga 101a. El dispositivo de distribución de carga 101a al que se la solicitado ejecutar el proceso, evalúa los servidores de juego a los que se ha de hacer ejecutar el proceso de juego de los servidores 101b a 101n. Los servidores de juego son evaluados basándose en los recursos guardados por cada uno de los servidores 101b a 101n. Como resultado de la evaluación por el dispositivo de distribución de carga 101a, el servidor 101b es determinado como el destino de distribución de carga tal como se indica por una flecha 502. Cuando se determina el servidor de juego, la ejecución del proceso de juego es solicitada desde el dispositivo de distribución de carga 101a y el proceso de juego es ejecutado en el servidor 101b. Así, el juego puede empezar a utilizarse por el dispositivo terminal 102.

[0040] Tal como se describió anteriormente, el dispositivo de distribución de carga 101a es activado, por el hecho de que la ejecución del proceso de juego es solicitada desde los dispositivos terminales 102a a 102n, para seleccionar el servidor óptimo para ejecutar este proceso de juego desde los servidores 101b a 101n. Sin limitarse a tal juego online tal como se describió anteriormente, cuando un proceso es ejecutado por otro software de aplicación, la carga puede ser distribuida utilizando el dispositivo de distribución de carga 101a.

[0041] Se describirá la relación entre las cantidades de recursos del servidor y las cantidades de recursos consumidos por el proceso según la realización de la presente invención. La fig. 6 y la fig. 7 son vistas explicativas que muestran ejemplos de una asignación de procesos ideal según la realización de la presente invención. En la fig. 6, el consumo de CPU de los servidores 101b a 101n está representado por una coordenada X y un consumo de memoria está representado por una coordenada Y. Este gráfico muestra una tendencia en el consumo de los recursos utilizados por el proceso ejecutado por los servidores 101b a 101n.

[0042] Las características del gráfico mostrado en la fig. 6 se describirán más adelante suponiendo que el servidor 101b tiene estas características. En este gráfico, una cantidad que corresponde al 100 % que es la cantidad máxima de un consumo de CPU está indicada por un valor numérico, “20”. Este es el valor numérico que representa la proporción con respecto al consumo de las CPU guardada por el servidor 101c en la fig. 7 descrita después (suponiendo que el consumo de las CPU del servidor 101c es “10”), y muestra que el servidor 101b tiene un rendimiento capaz de consumir una cantidad de CPU el doble de grande que la del servidor 101c.

[0043] Una cantidad que corresponde al 100 % que es el valor máximo de un consumo de memoria está indicada por un valor numérico, “10”. Análogamente al caso de las CPU, este es el valor numérico que representa la proporción con respecto al consumo de memorias guardada por el servidor 101c en la fig. 7 descrita después (suponiendo que el consumo de la memoria en la fig. 7 es “20”), y muestra que el servidor 101b tiene un rendimiento capaz de consumir una cuarta parte de las memorias del servidor 101c.

[0044] Un proceso 601 mostrado en el gráfico de la fig. 6 representa un consumo de CPU y un consumo de memoria del proceso ejecutado en el servidor 101b. Los consumos de las CPU y las memorias ocupadas por la

ejecución de este proceso 601 son respectivamente “10” para las CPU y “5” para las memorias y ocupan respectivamente la mitad de los recursos guardados por el servidor 101b.

[0045] Un estado en el que el primer proceso 601 ha sido ejecutado está representado como un rectángulo formado por el origen, la posición de “10” del consumo de CPU y la posición de “5” del consumo de memoria. Los valores máximos de las CPU y las memorias que han de ser consumidas están representados respectivamente como las coordenadas 602 (en lo sucesivo, las coordenadas constituidas por los valores máximos de las CPU y las memorias que han de ser consumidas por un proceso se denominarán “coordenadas de recursos consumidos”. Cuando es ejecutado un segundo proceso 601, este segundo proceso puede estar representado en forma escalonada con las coordenadas 602 del primer proceso como el punto de partida. Debido a este segundo proceso 601 que ha sido asignado, el consumo de CPU y el consumo de memoria se convierten respectivamente en los valores máximos de los recursos respectivos del servidor 101b, y las coordenadas que indican estos valores máximos están representadas como las coordenadas 603.

[0046] Tal como se describió anteriormente, cuando son ejecutados dos procesos 601, tanto el consumo de CPU como el consumo de memoria en el servidor 101b se convierten en los valores máximos, y no quedan recursos restantes que asignar a otro proceso tanto para la CPU como para la memoria. Por lo tanto, se produce un estado en el que los recursos son consumidos idealmente.

[0047] En la fig. 7, el consumo de CPU de los servidores 101b a 101n está representado como la coordenada X y el consumo de memoria de los servidores está representado como la coordenada Y. Este gráfico muestra la tendencia de los consumos de los recursos utilizados por procesos ejecutados por los servidores 101b a 101n.

[0048] A continuación se describirán las características del gráfico mostrado en la fig. 7 suponiendo que el servidor 101c tiene las características. En este gráfico, una cantidad que corresponde al 100 % que es la cantidad máxima del consumo de CPU está indicada por un valor numérico, “10”. Esto representa la mitad de la cantidad consumible de la CPU guardada por el servidor 101b de la fig. 6 descrita anteriormente, y representa que el servidor 101c guarda una función que puede consumir la mitad del consumo de CPU del servidor 101b.

[0049] Una cantidad que corresponde al 100 % que es el valor máximo del consumo de memoria está indicada por un valor numérico, “20”. Análogamente al caso de la CPU, esto representa una cantidad el doble de grande que la cantidad consumible de la memoria guardada por el servidor 101b de la fig. 6 descrita anteriormente, y representa que el servidor 101c guarda una función que puede consumir el doble que el consumo de memoria del servidor 101b.

[0050] Un proceso 701 mostrado en el gráfico de la fig. 7 representa un consumo de CPU y un consumo de memoria para el proceso ejecutado en el servidor 101c. Cada uno de los consumos de la CPU y la memoria ocupadas por la ejecución de este proceso 701 es “5” para la CPU y “10” para la memoria, y esos consumos ocupan respectivamente la mitad de los recursos guardados por el servidor 101c.

[0051] Un estado en el que un primer proceso 701 ha sido ejecutado está representado como un rectángulo formado por el origen, la posición de “5” de un consumo de CPU, y la posición de “10” de un consumo de memoria, y los valores máximos de la CPU y la memoria que han de ser consumidas pueden estar representados respectivamente como las coordenadas 702. Cuando es ejecutado un segundo proceso 701, este proceso puede estar representado en forma escalonada con las coordenadas 702 del primer proceso 701 como punto de partida. Debido a este segundo proceso 701 que ha sido asignado, el consumo de CPU y el consumo de memoria se convierten respectivamente en los valores máximos de los recursos respectivos del servidor 101c, y las coordenadas que indican estos valores máximos están representadas como las coordenadas 703.

[0052] Tal como se describió anteriormente, cuando son ejecutados dos procesos 701, tanto el consumo de CPU como el consumo de memoria en el servidor 101c se convierten en los valores máximos, y no quedan recursos restantes que asignar a otro proceso tanto para la CPU como para la memoria. Por lo tanto, se produce un estado en el que los recursos son consumidos idealmente.

[0053] Se describirá el consumo de recursos óptimo en los servidores según la realización de la presente invención. La fig. 8 a la fi. 10 son vistas explicativas que muestran un ejemplo del consumo de recursos óptimo en los servidores del sistema servidor/cliente según la realización. En la fig. 8, el consumo de la CPU está representado por la coordenada X y el consumo de la memoria está representado por la coordenada Y. En la fig. 8, y la fig. 9 y la fig. 10 descritas después, se representa que, cuando existen los servidores 101d a 101f, las cantidades de recursos

que pueden ser consumidos respectivamente por los servidores difieren entre sí.

[0054] Las características del gráfico de la fig. 8 se describirán más adelante suponiendo que el servidor 101d tiene estas características. En este gráfico, una cantidad que corresponde al 100 % que es la cantidad máxima del consumo de CPU está representada por un valor máximo 801. Este representa el 80 % del valor máximo del consumo de CPU de la fig. 9 descrita después, y representa el 50 % del valor máximo del consumo de CPU de la fig. 10. Una cantidad que corresponde al 100 % que es la cantidad máxima de un consumo de memoria está representada por un valor máximo 802. Este representa la misma cantidad que el valor máximo de un consumo de memoria de la fig. 9 descrita después, y representa el 200 % que es el doble de grande que el valor máximo del consumo de memoria de la fig. 10.

[0055] Tal como se muestra en la fig. 8, un proceso 803 iniciado en primer lugar está representado desde la posición del origen; los procesos iniciados después del proceso 803 están representados en una forma escalonada en la secuencia temporal desde los procesos 804 hasta 808; y, por último, se alcanzan las coordenadas 809 que es la intersección del valor máximo 801 del consumo de CPU y el valor máximo 802 del consumo de memoria. Así, el servidor 101d está en un estado en el que las CPU y las memorias de las mismas son consumidas idealmente.

[0056] En la fig. 9, el consumo de CPU está representado por la coordenada X y el consumo de memoria está representado por la coordenada Y en el gráfico de la misma. En la fig. 9, análogamente a la fig. 8, se representa que, cuando existen los servidores 101d a 101f, las cantidades de recursos que pueden ser consumidos respectivamente por los servidores difieren entre sí.

[0057] Las características del gráfico mostrado en la fig. 9 se describirán más adelante suponiendo que el servidor 101e tiene estas características. En este gráfico, una cantidad que corresponde al 100 % que es la cantidad máxima del consumo de CPU está representada por un valor máximo 901. Este representa el 125 % del valor máximo del consumo de CPU de la fig. 8 descrita anteriormente, y representa el 62,5 % del valor máximo del consumo de CPU de la fig. 10 descrita después.

[0058] Una cantidad que corresponde al 100 % que es la cantidad máxima del consumo de memoria está representada por un valor máximo 902. Este representa la misma cantidad que el valor máximo del consumo de memoria de la fig. 8 descrita anteriormente, y es el doble de grande que el valor máximo del consumo de memoria de la fig. 10 descrita después.

[0059] Tal como se muestra en la fig. 9, un proceso 903 iniciado en primer lugar está representado desde la posición del origen; los procesos iniciados después del proceso 903 están representados en una forma escalonada en la secuencia temporal desde los procesos 904 hasta 908; y, por último, se alcanzan las coordenadas 909 que es la intersección del valor máximo 901 del consumo de CPU y el valor máximo 902 del consumo de memoria. Así, el servidor 101e está en un estado en el que la CPU y la memoria de la misma son consumidas idealmente.

[0060] En la fig. 10, el consumo de CPU está representado por la coordenada X y el consumo de memoria está representado por la coordenada Y. En la fig. 10, análogamente a la fig. 8 y la fig. 9, se representa que, cuando existen los servidores 101d a 101f, las cantidades de recursos que pueden ser consumidos respectivamente por los servidores difieren entre sí.

[0061] Las características del gráfico mostrado en la fig. 10 se describirán más adelante suponiendo que el servidor 101f tiene las características. En este gráfico, una cantidad que corresponde al 100 % que es la cantidad máxima del consumo de CPU está representada por un valor máximo 1001. Este representa el 200 % del valor máximo del consumo de CPU de la fig. 8 descrita anteriormente, y representa el 160 % del valor máximo del consumo de CPu de la fig. 9. Una cantidad que corresponde al 100 % que es la cantidad máxima del consumo de memoria está representada por un valor máximo 1002. Este valor máximo 1002 es el 50 % del valor máximo del consumo de memoria de la fig. 8 y la fig. 9 descritas anteriormente.

[0062] Tal como se muestra en la fig. 10, un proceso 1003 iniciado en primer lugar está representado desde la posición del origen; los procesos iniciados después del proceso 1003 están representados en una forma escalonada en la secuencia temporal desde los procesos 1004 hasta 1008; y, por último, se alcanzan las coordenadas 1009 que es la intersección del valor máximo 1001 del consumo de CPU y el valor máximo 1002 del consumo de memoria. Así, el servidor 101f está en un estado en el que la CPU y la memoria de la misma son consumidas idealmente.

[0063] Se describirá un ejemplo de consumos de recursos en los servidores según la realización de la presente invención. La fig. 11 es una vista explicativa que muestra un ejemplo de consumo de recursos de los servidores según la realización de la presente invención. En la fig. 11, Xmax en la coordenada X indica el 100 % que es el valor máximo del consumo de CPU, e Ymax en la coordenada Y indica el 100 % que es el valor máximo del consumo de memoria.

[0064] Una línea 1101 es un ejemplo en el que el recurso de CPU y el recurso de memoria son asignados a cada proceso cada vez que ocurre el proceso, y por último se alcanzan las coordenadas 1102 que son una intersección de la Xmax que es el valor máximo del consumo de CPU y la Ymax que es el valor máximo del consumo de memoria. Así, se muestra que la línea 1101 es un gráfico del caso en el que el proceso está distribuido sin ningún desperdicio del consumo de CPU y el consumo de memoria, y se representa el resultado obtenido consumiendo idealmente los recursos.

[0065] Una línea 1103 es un ejemplo en el que, como resultado de asignar el recurso de CPU y el recurso de memoria a cada proceso cada vez que ocurre, el consumo de memoria ha alcanzado el valor máximo Ymax del mismo antes de que el consumo de CPU alcance el valor máximo Xmax del mismo. Esto muestra que la línea 1103 es un gráfico del caso en el que el consumo de CPU está parcialmente desperdiciado a diferencia del consumo de memoria, y muestra que a la CPU le queda una cantidad de área utilizable indicada por una sección 1105.

[0066] Una línea 1106 muestra un ejemplo en el que, como resultado de asignar el recurso de CPU y el recurso de memoria a cada proceso cada vez que ocurre, el consumo de CPU ha alcanzado el valor máximo Xmax del mismo antes de que el consumo de memoria alcance el valor máximo Ymax del mismo. Esto muestra que la línea discontinua 1106 es un gráfico del caso en el que el consumo de memoria está parcialmente desperdiciado a diferencia del consumo de CPU, y muestra que a la memoria le queda una cantidad de área utilizable indicada por una sección 1108.

[0067] Según el procedimiento de asignar el proceso por el dispositivo de distribución de carga convencional, la carga es distribuida sin juzgar los consumos de recursos de los servidores, y por lo tanto, es normal que se obtengan los estados en los que los recursos no son consumidos uniformemente tal como se indica por las líneas 1103 y 1106. En cambio, cuando el proceso es asignado de modo que los recursos que son la CPU y la memoria son consumidos uniformemente, es posible evitar el desperdicio de los recursos que quedan sin consumir, tal como indica por la línea 1101.

(Procedimiento de evaluación de servidores)

[0068] Se describirá un procedimiento según la realización de la invención de evaluación de servidores ejecutado por el dispositivo de distribución de carga. La fig. 12 es una vista explicativa que muestra una estructura de un procedimiento según la realización de la invención de evaluación de los servidores ejecutado por el dispositivo de distribución de carga. En la fig. 12, el consumo de CPU está representado por la coordenada X y el consumo de memoria está representado por la coordenada Y en los gráficos 1202 y 1203. El gráfico 1202 muestra el contenido del consumo de recursos del servidor 101b y el gráfico 1203 muestra el contenido del consumo de recursos del servidor 101c.

[0069] Comparado con el gráfico 1203, el gráfico 1202 muestra una característica de un servidor que guarda recursos para los cuales el valor máximo del consumo de CPU es el doble de grande que el del gráfico 1203 y el valor máximo del consumo de memoria es la mitad de grande que el mostrado en el gráfico 1203. En el caso del gráfico 1202, los procesos 1204 y 1205 están asignados y ejecutados. Por otra parte, comparado con el gráfico 1202, el gráfico 1203 muestra una característica de un servidor que guarda recursos para los cuales el valor máximo del consumo de CPU es la mitad de grande que el del gráfico 1203 y el valor máximo del consumo de memoria es el doble de grande que el del gráfico 1203. En el caso del gráfico 1203, los procesos 1210 y 1211 ya están asignados y ejecutados. Más adelante se describirá un procedimiento de determinación de a cuál de los servidores que tienen respectivamente la característica de los gráficos 1202 y 1203 se le debería distribuir la carga del proceso 1201 para consumir óptimamente los recursos cuando es enviada una petición de ejecución del proceso 1201.

[0070] Cuando el proceso 1201 es ejecutado basándose en la característica del gráfico 1203, como se añade una cantidad del consumo de recursos que ha de estar causado por el proceso 1202, las coordenadas que representan la cantidad total de recursos consumidos (coordenadas de recursos consumidos) se cambian desde una cima 1211a del proceso 1211 a las coordenadas 1215. Una línea recta formada conectando las coordenadas 1213 que son la intersección de los valores máximos del consumo de CPU y el consumo de memoria, y el origen se define

como una “línea óptima de consumo de recursos” 1212 (en lo sucesivo, una línea recta formada conectado las coordenadas que son la intersección de los valores máximos del consumo de CPU y el consumo de memoria, y el origen se denominad “línea óptima de consumo de recursos”), y se dibuja una normal 1214 desde las coordenadas 1215 hasta esta línea óptima de consumo de recursos 1212.

[0071] Por otra parte, cuando el proceso 1201 es ejecutado basándose en la característica del gráfico 1202, como se añade una cantidad del consumo de recursos que ha de estar causado por el proceso 1201, las coordenadas que representan la cantidad total de recursos consumidos (coordenadas de recursos consumidos) se cambia de una cima 1205a del proceso 1205 a las coordenadas 1209. Una línea recta formada conectando las coordenadas 1207 que son la intersección de los valores máximos del consumo de CPU y el consumo de memoria, y el origen se define como una línea óptima de consumo de recursos 1206, y se dibuja una normal 1208 desde las coordenadas 1209 hasta esta línea óptima de consumo de recursos 1206.

[0072] Comparando las longitudes de la normal 1208 con la normal 1214, se muestra que la longitud de la normal 1208 es más corta. Esto indica que el equilibrio entre los consumos de la memoria y la CPU que son los recursos del servidor 101b está más cerca de ser uniforme cuando el proceso 1201 es ejecutado por el servidor 101b que cuando el proceso 1201 es ejecutado por el servidor 101c. Por lo tanto, como servidor para ejecutar el proceso 1201, no se selecciona el servidor 101c que guarda la característica del gráfico 1203 y se selecciona el servidor 101b que guarda la característica del gráfico 1202. Los recursos son consumidos óptimamente cuando el proceso 1201 es ejecutado por el servidor 101b.

[0073] Para lo anterior, puede existir el caso en el que las longitudes de las normales 1208 y 1214 sean iguales y puede juzgarse que los recursos pueden ser consumidos apropiadamente por igual cuando se ha seleccionado cualquiera de los servidores. En este caso, se comparan las distancias 1216 y 1217 entre las coordenadas de recursos consumidos 1209 y 1215 respectivamente, y el origen, y se asigna el servidor relacionado con la más corta de las distancias. Haciendo esto, puede asignarse un servidor que tiene un menor consumo de recursos cuando existen servidores que tienen respectivamente una distancia igual a la línea óptica de consumo de recursos de los mismos. Cuando se desea que se tengan en cuenta tanto el consumo de recursos como la distancia desde la línea óptima de consumo de recursos, este deseo puede realizarse seleccionando un servidor que tenga un menor valor de un área (1208x1216, 1214x1217) obtenido multiplicando la longitud de la normal (1208, 1214) por la distancia al origen (1216, 1217).

[0074] A continuación se explicará una función para el procedimiento de evaluación del servidor que ha de ser asignado según la realización de la presente invención. La fig. 13 es una vista explicativa que muestra un ejemplo de una función para el procedimiento de evaluación del servidor que ha de ser asignado según la realización de la presente invención. En la fig. 13, Xmax en la coordenada X mostrada representa el 100 % que es el valor máximo de un consumo de CPU e Ymax en la coordenada Y representa el 100 % que es el valor máximo de un consumo de memoria. En un espacio que tiene parámetros de recursos como los ejes del mismo, las coordenadas de la intersección del origen y las capacidades máximas disponibles de los parámetros, es decir, Xmax e Ymax se definen como (X_{max}, Y_{max}), y una línea recta formada conectando (X_{max}, Y_{max}) y el origen (0, 0) se define como una línea óptima de consumo de recursos 1301. Las características del gráfico mostrado en la fig. 13 se describirán más adelante suponiendo que el servidor 101b tiene estas características.

[0075] Un proceso 1302 que es ejecutado en primer lugar se muestra como un rectángulo que parte del origen y que tiene el consumo de memoria y el consumo de CPU consumidos por el proceso 1302 respectivamente como la longitud de los lados longitudinales y laterales del mismo. Este es un rectángulo que muestra los recursos consumidos por cada proceso. Una diagonal hasta la cima desde el origen del rectángulo mostrado por este proceso está representada como las coordenadas 1302a.

[0076] Un proceso 1303 que es ejecutado en segundo lugar se muestra como un rectángulo que tiene las coordenadas 1302a como puto de partida del mismo y el consumo de memoria y el consumo de CPU consumidos por el proceso 1303 respectivamente como la longitud de los lados longitudinal y lateral del mismo. Una diagonal hasta la cima desde las coordenadas 1302a que es el punto de partida de este rectángulo está representada como las coordenadas 1303a.

[0077] Cuando un tercer proceso que está recién asignado es un proceso 1304, el proceso 1304 se muestra como en la fig. 13 como un rectángulo que tiene la coordenada 1303a como punto de partida del mismo, el consumo de CPU como “c_x”, y el consumo de memoria como “c_y”. Una diagonal hasta la cima de la coordenada 1303a que es el punto de partida del rectángulo está representada como las coordenadas 1304a.

[0078] Se obtendrá la longitud de una normal 1309 dibujada desde las coordenadas 1304a hasta la línea óptima de consumo de recursos 1301. Los valores de las posiciones de x e y mostradas en el gráfico de la fig. 13 se expresan a continuación en la ecuación (1) y la ecuación (2).

$$y=f_{xy}(x)=(y_{max}/x_{max})Xx \quad (1)$$

$$x=f_{yx}(y)=(x_{max}/y_{max})Xy \quad (2)$$

[0079] Tal como se describió anteriormente, las coordenadas de recursos consumidos, consumidos por el primer proceso 1302 y el segundo proceso 1303, están representadas como las coordenadas 1303a. Las coordenadas de recursos consumidos, consumidos por el tercer proceso, están representadas como las coordenadas 1304a. Las coordenadas de las coordenadas 1304a están representadas como $(x_1,y_1)=(x_0+c_x,y_0+c_y)$.

[0080] Por otra parte, la intersección de una línea recta paralela al eje longitudinal y que cruza las coordenadas 1304a, y la línea óptima de consumo de recursos 1301 está representada como las coordenadas 1305. La intersección de una línea recta paralela al eje lateral y que cruza las coordenadas 1304a, y la línea óptima de consumo de recursos 1301 está representada como las coordenadas 1306. Una línea recta que conecta las coordenadas 1304a y las coordenadas 1305 está representada como una línea recta 1307 y la longitud de la línea recta 1307 viene dada por la ecuación (3) a continuación.

$$\Delta y=|f_{xy}(x_1)-y_1| \quad (3)$$

[0081] Una línea recta que conecta las coordenadas 1304a y las coordenadas 1306 está representada como una línea recta 1308 y la longitud de la línea recta 1308 viene dada por la ecuación (4) a continuación.

$$\Delta x=|f_{yx}(y_1)-x_1| \quad (4)$$

[0082] Una normal dibujada desde las coordenadas 1304a hasta la línea óptima de consumo de recursos 1301 está representada como una normal 1309 y la longitud de la normal 1309 viene dada por la ecuación (5) a continuación.

$$dis_{xy}=\Delta x\Delta y/\sqrt{(\Delta x^2+\Delta y^2)} \quad (5)$$

[0083] Esta dis_{xy} es la distancia desde la línea óptima de consumo de recursos 1301 hasta las coordenadas 1304a y es el valor que ha de ser evaluado cuando el dispositivo de distribución de carga 101a determina los servidores a los que se ha de hacer ejecutar un proceso desde una pluralidad de servidores.

[0084] Se describirá una ecuación para obtener la distancia desde las coordenadas de recursos consumidos hasta la línea óptima de consumo de recursos para el caso de una función de n dimensiones que incluye otros parámetros además de las CPU y las memorias descritas anteriormente. Una línea óptima de consumo de recursos que tiene n parámetros (x_i) (donde $1\leq i\leq n$) que se utilizan como recursos y pueden obtenerse en un espacio euclídeo de n dimensiones, se expresa por la ecuación (6) a continuación utilizando el caso para dos dimensiones.

$$x_{i+1}=f_{xix(i+1)\bmod n}(x_i)=(x_{(i+1)\bmod n\ max}/x_{i\ max})\times x_i \quad (6)$$

[0085] Suponiendo que los consumos de recursos de los servidores existentes son (x_{i0}) (donde $1\leq i\leq n$) y los consumos de recursos de un proceso son (c_i) (donde $1\leq i\leq n$),

$$x_{i1}=x_{i0}+c_i \quad (7)$$

[0086] En este caso, la ecuación (8) es,

$$\Delta x_{(i+1)\bmod n}=|f_{xi\ x(i+1)\bmod n}(x_{i1})-x_{(i+1)\bmod n\ 1}| \quad (8)$$

y la distancia entre un punto (x_{ii}) (donde $1\leq i\leq n$) en el espacio euclídeo de n dimensiones y la línea óptima de consumo de recursos 1301 puede expresarse como en la ecuación (9) a continuación.

$$dis = \prod_{i=1}^{n} (\Delta x_{(i+1) \bmod n}) / \sqrt{(\sum_{i=1}^{n} (\Delta x_{(i+1) \bmod n})^2)} \quad (9)$$

[0087] Por lo tanto, el dispositivo de distribución de carga 101a calcula la distancia (dis) expresada por la ecuación (5) o la ecuación (9) para cada uno de los servidores 101b a 101n a los que han de distribuirse las cargas; y asigna el procesamiento para la ejecución del proceso a un servidor que tiene la más corta de las distancias.

[0088] Una distancia 1310 entre un punto (x_{ii}) (donde $1 \leq i \leq n$) en el espacio euclídeo de n dimensiones y el origen puede expresarse en la ecuación (10).

$$diag = \sqrt{(\sum_{i=1}^{n} {x_{i1}}^2)} \quad (10)$$

[0089] Por lo tanto, cuando se selecciona un servidor de entre los servidores que tienen igual longitud dis de las normales del mismo, se selecciona un servidor que tiene el valor más pequeño de diag y el procesamiento para la ejecución del proceso es asignado a este servidor. Cuando se desea que la longitud dis de la normal y la distancia diag al origen sean evaluadas simultáneamente, el procesamiento es asignado al servidor que tiene el valor más pequeño de disxdiag.

[0090] Los recursos aparte de las CPU y las memorias descritas anteriormente son, por ejemplo, GPU, memorias de vídeo, tarjetas de interfaz de red, etc. Una GPU es un chip gráfico para ejecutar, por hardware, el procesamiento de representación y geometría en juegos que utilizan gráficos 3-D. Una memoria de vídeo es una memoria que es utilizada por una GPU cuando la GPU ejecuta el procesamiento de un proceso de imagen, y está contenida en la GPU. Una tarjeta de interfaz de red es, por ejemplo, un adaptador Ether que termina en 100Base, 1000Base-TX, etc.

[0091] Se describirán los detalles del procesamiento para evaluar y determinar un servidor según la realización de la presente invención mediante el dispositivo de distribución de carga. La fig. 14 es un diagrama de flujo que muestra un ejemplo de un procedimiento de evaluación y determinación de un servidor según la realización de la presente invención mediante el dispositivo de distribución de carga. En el diagrama de flujo mostrado en la fig. 14, se establecen los parámetros utilizados cuando el dispositivo de distribución de carga 101a evalúa los servidores 101b a 101n (etapa S1401). Los parámetros son, por ejemplo, un consumo de CPU, un consumo de memoria, un consumo de GPU, un consumo de memoria de vídeo incluida en una GPU, un consumo de banda de una tarjeta de interfaz de red, etc.

[0092] Se definen y establecen para cada parámetro las cantidades de recursos utilizados utilizadas respectivamente por una pluralidad de procesos cuando los procesos son ejecutados (etapa S1402). Para cada uno de los servidores 101b a 101n, se establecen los valores máximos de los recursos que son los parámetros definidos en la etapa S1402 (etapa S1403).

[0093] Para cada uno de los servidores 101b a 101n se obtienen y monitorizan los consumos de recursos de los parámetros establecidos en la etapa S1401 (etapa S1404). Se juzga si una petición de ejecución de un proceso ha sido enviada desde el dispositivo terminal 102 que es el cliente o el teléfono móvil 103 (etapa S1405). Cuando no ha sido enviada ninguna petición de ejecución del proceso (etapa S1405: NO), el procedimiento vuelve a la etapa S1404, y se prosigue la etapa de monitorizar los consumos de recursos consumidos, para cada uno de los servidores 101b a 101n (etapa S1404).

[0094] Por otra parte, cuando una petición de ejecución de un proceso ha sido enviada desde el dispositivo terminal 102 que es el cliente (etapa S1405: SÍ), los servidores candidatos a ejecutar este proceso son seleccionados de entre os servidores 101b a 101n (etapa S1406). Esta selección se lleva a cabo de modo que los servidores que han utilizado cantidades que no superan el valor máximo para uno cualquiera de los parámetros que han sido establecidos en la etapa S1401 son seleccionados de entre los servidores 101b a 101n. Además se lleva a cabo una evaluación obteniendo la distancia de la normal 1309 dibujada desde las coordenadas de recursos de consumo 1304 hasta la línea óptima de consumo de recursos 1301 de la fig. 13 descrita anteriormente, para los candidatos seleccionados en la etapa S1406, y el servidor que tiene la distancia más corta, por ejemplo, el servidor 101b es determinado como el destino de distribución de carga (etapa S1407).

[0095] En la selección de servidores candidatos en S1406, tal como se describió anteriormente, la distancia 1310 entre las coordenadas de recursos de consumo 1304 y el origen puede evaluarse simultáneamente además de la longitud de la normal 1309 dibujada desde las coordenadas de recursos de consumo 1304a hasta la línea óptima de consumo de recursos 1301. Cuando existen servidores que tienen respectivamente la normal de igual longitud, el servidor que tiene el consumo de recursos más pequeño puede seleccionarse seleccionando el servidor que tiene la distancia más corta 1310 al origen. Si no, seleccionando el servidor que tiene el valor más pequeño obtenido multiplicando la longitud de la normal 1309 por la distancia 1310 hasta el origen, puede llevarse a cabo el consumo óptimo de recursos y el proceso puede ser asignado al servidor que tiene el consumo de recursos más pequeño.

[0096] El dispositivo de distribución de carga 101a solicita la ejecución del proceso al servidor 101b que es el destino de distribución de carga determinado en la etapa S1407 (etapa S1408), y notifica el dispositivo terminal 102 de la dirección IP, el número de puerto de servicio, etc., que son colectivamente la información necesaria para acceder al servidor 110b que es el destino de distribución de carga para mostrar al dispositivo terminal 102 que es el cliente que el proceso ha sido ejecutado (etapa S1409). El dispositivo terminal 102 al que se le ha notificado la información necesaria para acceder al servidor 101b, puede acceder directamente al servidor 101b y el procesamiento finaliza.

[0097] Se describirán los detalles del procedimiento para evaluar y determinar un servidor según la realización de la presente invención mediante el dispositivo de distribución de carga. La fig. 15 es un esquema que muestra un ejemplo de un procedimiento de evaluación y determinación de un servidor según la realización de la presente invención mediante el dispositivo de distribución de carga. En la fig. 15, se muestra que el dispositivo de distribución de carga 101a distribuye la carga de la ejecución de un proceso solicitada por el dispositivo terminal 102a a 102n a los servidores 101b a 101n. En la fig. 15, el dispositivo de distribución de carga 101a ejecuta el procesamiento, siguiendo el procedimiento expuesto a continuación.

(1) Se establecen los parámetros utilizados para la evaluación llevada a cabo cuando se distribuye la carga.
(2) Se define y establece una cantidad utilizada para cada parámetro consumido por el proceso ejecutado en cada servidor. Este ajuste se efectúa cada vez que se añade un proceso. Por ejemplo, en el caso en que se utilizan cuatro parámetros tales como la CPU, la memoria, la GPU y la memoria de vídeo, cuando el proceso que ha de ser ejecutado es un juego 3-D, se consumen recursos para la GPU y la memoria de vídeo además de las CPU y las memorias. Cuando el proceso que ha de ser ejecutado es un juego 2-D, la GPU y la memoria de vídeo no se consumen. De esta manera, los recursos que se consumirán difieren según el contenido del proceso.
(3) Se establece el valor máximo de cada parámetro guardado respectivamente por los servidores. Este ajuste puede efectuarse manualmente en el dispositivo de distribución de carga 101a por un administrador, etc., o los servidores pueden monitorizarse respectivamente uno tras otro a través de la red 100 y puede hacerse que obtengan los valores que han sido establecidos. Como es probable que estén dispuestos varios servidores 101b a 101n según el propósito de uso de los mismos, puede establecerse la capacidad máxima para cada uno de los parámetros. Esta etapa está indicada por una flecha 1501a.
(4) Para los recursos establecidos como los parámetros, se monitorizan las cantidades consumidas para cada uno de los servidores 101b a 101n. Esta etapa está indicada por la flecha 1501a y los servidores 101b a 101n se monitorizan uno tras otro desde el dispositivo de distribución de carga 101a y se obtienen los consumos.
(5) Una petición de ejecución de un proceso es aceptada desde el dispositivo terminal 102a que es el cliente. Simultáneamente, se obtiene la cantidad de cada recurso que ha ser utilizado por el proceso aceptado. Esta etapa está indicada por una flecha 1502.
(6) El dispositivo de distribución de carga 101a selecciona servidores candidatos a ejecutar el proceso de los servidores 101b a 101n. Se comprueba si el consumo de los recursos que han de ser utilizados por el proceso para el cual ha sido aceptada la petición de ejecución, no supera los valores máximos de los recursos de cada servidor y los servidores que tienen los valores máximos que no son superados son seleccionados como los candidatos.
(7) El dispositivo de distribución de carga 101a evalúa además los candidatos seleccionados en (6) obteniendo cada distancia de la normal 1309 dibujada desde las coordenadas de recursos de consumo 1304a hasta la línea óptima de consumo de recursos 1301 o cada distancia 1310 entre las coordenadas de recursos de consumo 1304a y el origen mostrado en la fig. 13 descrita anteriormente; y, de esos candidatos, determina el servidor que tiene la normal del mismo que tiene la longitud más corta o, cuando las longitudes de las normales son iguales, el servidor que tiene la distancia más corta 1310 del mismo desde el origen o el servidor que tiene el valor más pequeño del mismo obtenido multiplicando la longitud de la normal por la distancia desde el origen, por ejemplo, el servidor 101b como el destino de distribución de carga.
(8) El dispositivo de distribución de carga 101a solicita la ejecución del proceso al servidor 101b que es el destino

de distribución de carga determinado en la etapa S1407. En este momento, cuando el servidor 101b acepta sólo accesos para la petición de la ejecución desde clientes que han sido autentificados, las ID, las direcciones IP, los números de puertos de servicio, las claves para cifrado, los testigos, etc., también son transmitidos desde el dispositivo terminal 102 hasta el servidor 101b. Esta etapa está indicada por la flecha 1501a.

(9) Se notifica la información sobre el servidor 101b que es el destino de distribución de carga para mostrar al dispositivo terminal 102a que es el cliente que el proceso ha sido ejecutado. Esta información es necesaria para el dispositivo terminal 102a cuando el dispositivo terminal 102a accede al servidor 101b. Por ejemplo, esta información se refiere a la dirección IP, el número de puerto de servicio, las claves y los testigos relacionados con la autentificación y el cifrado del servidor 101b necesarios para que el dispositivo terminal 102a acceda directamente al servidor 101b. Esta etapa está indicada por la flecha 1502.

(10) El dispositivo terminal 102a al que se ha notificado la información necesaria para acceder al servidor 101b, puede acceder al servidor 101b. Basándose en la información notificada en (9), se permite un acceso desde el dispositivo terminal 102a al servidor 101b. En el lado del servidor 101b, si un acceso procede o no del cliente que ha sido autentificado se verifica comprobando si la dirección IP, el ID, los testigos, etc., coinciden o no con los de la información notificada. La comunicación es cifrada cuando las claves de cifrado han sido intercambiadas de antemano. Esta etapa está indicada por una flecha 1503.

(Contenido de la información sobre el proceso)

[0098] Se describirá la información sobre los procesos guardada por el dispositivo de distribución de carga según la realización de la invención. La fig. 16 es una vista explicativa que muestra un ejemplo de una tabla que guarda la información sobre el proceso según la realización de la presente invención. Tal como se muestra en la fig. 16, una tabla de información de procesos 1600 almacena el contenido de datos para guardar la información sobre los procesos y consiste en columnas respectivamente para ID para identificar únicamente los procesos respectivamente, nombres de procesos que son nombres de los procesos, una cantidad utilizada de CPU utilizada cuando es ejecutado cada proceso, y la cantidad utilizada de memoria utilizada cuando es ejecutado cada proceso.

[0099] El atributo de datos del ID es un tipo de texto, o un tipo de número entero si un ID puede expresarse sólo en dígitos. El atributo de datos del nombre de proceso es un tipo de texto, etc. Los atributos de datos del consumo de CPU y el consumo de memoria se definen para que sean un atributo de un tipo de coma flotante, etc.

[0100] Según el contenido de datos de un proceso mostrado por un registro que tiene un ID de “100”, el nombre de proceso es Proceso 1, la cantidad de CPU que ha de utilizarse es 0,15, y la cantidad de memoria que ha de utilizarse es 0,2. Según el contenido de datos de un proceso mostrado por un registro que tiene un ID de “110”, el nombre de proceso es Proceso 2, la cantidad de CPU que ha de utilizarse es 0,2, y la cantidad de memoria que ha de utilizarse es 0,4, y la cantidad de CPU que ha de utilizarse y la cantidad de memoria que ha de utilizarse son iguales que el caso del proceso que tiene el ID de “100”. Esto muestra que, cuando son ejecutados respectivamente los procesos que tienen el ID de “100” y “110”, los valores del consumo de la CPU y la memoria consumida en un servidor que ejecuta cada uno de los procesos se incrementan en los valores de las cantidades utilizadas de la CPU y la memoria registradas respectivamente en la tabla.

[0101] Cuando el dispositivo de distribución de carga 101a ha recibido la petición de ejecución del proceso, el dispositivo de distribución de carga 101a selecciona un servidor al que distribuir la carga, utilizando la información de proceso en la tabla de información de proceso 1600. Cuando el número de servidores 101b a 101n a los que distribuir es pequeño, a veces es más sencillo que la información sea escrita en un archivo para referencia mediante un archivo de texto, etc., sin ninguna administración de tabla utilizando la base de datos tal como la tabla descrita anteriormente. En tal caso, el contenido de la tabla se crea como un archivo para referencia según sea necesario y el formato del archivo puede determinarse arbitrariamente.

(Información sobre la tabla)

[0102] Se describirá la información de servidor guardada de acuerdo con la realización de la presente invención por el dispositivo de distribución de carga. La fig. 17 es una vista explicativa que muestra un ejemplo de una tabla que guarda información sobre los servidores en el sistema servidor/cliente según la realización de la presente invención. En la fig. 17, una tabla de información de servidor 1700 almacena datos para guardar la información sobre los servidores y consta de columnas respectivamente para nombres de servidores de destino de conexión para identificar únicamente los servidores, el índice de uso de una CPU que se utiliza en cada servidor, y el índice de uso de una memoria que se utiliza en cada servidor.

[0103] El atributo de datos de los nombres de servidores de destino de conexión es un tipo de texto, o un tipo de número entero si un nombre puede expresarse sólo en dígitos. Los atributos de datos del índice de uso de una CPU y el índice de uso de una memoria son de tipo de coma flotante, etc. El índice de uso de una CPU o una memoria enumerado en la tabla de información de servidor 1700 es un índice, expresado como un valor, de una CPU o una memoria ya utilizada en cada uno de los servidores 101a, 101b y 101c. Los valores máximos de los índices de uso de una CPU y una memoria en los servidores 101b, 101c y 101d son respectivamente 1,0 (=100 %). El dispositivo de distribución de carga 101a puede monitorizar regularmente estos valores de cada uno de los servidores y establecer estos valores dentro de la tabla, o puede obtener estos valores según sea necesario cuando el dispositivo 101a distribuye la carga. A cuáles de los servidores 101b, 101c y 101d se distribuye la carga se evalúa como se indica a continuación.

[0104] Según un registro que tiene el nombre de servidor de destino de conexión, “101b” enumerado en la primera fila de la tabla de información de servidor 1700, el índice de uso de una CPU es 0,9 y el índice de uso de una memoria es 0,88. Los índices de uso de este servidor 101b son elevados tanto para una CPU como para una memoria, y cuando se añaden la cantidad de uso de una CPU y la cantidad de uso de una memoria del Proceso 1 mostrado en la fig. 16, superan respectivamente 1,0 (=100 %). Por lo tanto, se comprueba que no queda margen en los recursos para ejecutar el Proceso 1.

[0105] Según un registro que tiene el nombre de servidor de destino de conexión, “101c” enumerado en la segunda fila, el índice de uso de las CPU es 0,5 y el índice de uso de las memorias es 0,2. Cuando la cantidad de uso de una CPu y la cantidad de uso de una memoria del Proceso 1 mostrado en la fig. 16 se añaden a las de este servidor 101c, el índice de uso de una CPU es 0,65 y el índice de uso de una memoria es 0,4. Por lo tanto, ambos índices son inferiores a 1,0 (=100 %) y puede apreciarse que queda algo de margen de los recursos para ejecutar el Proceso 1.

[0106] Según un registro que tiene el nombre de servidor de conexión, “101d” enumerado en la tercera fila, el índice de uso de una CPU es 0,3 y el índice de uso de una memoria es 0,2. Cuando la cantidad de uso de una CPU y la cantidad de uso de una memoria del Proceso 1 mostrado en la fig. 16 se añaden a las de este servidor 101d, el índice de uso de una CPu es 0,45 y el índice de uso de una memoria es 0,4. Por lo tanto, ambos índices son inferiores a 1,0 (=100 %) y se comprueba que queda algo de margen en los recursos para ejecutar el Proceso 1.

[0107] Utilizando la ecuación (5), obteniendo la distancia dis de cada normal dibujad desde las coordenadas de recursos de consumo de estos servidores 101c y 101d hasta la línea óptima de consumo de recursos, la distancia para el servidor 101c es aproximadamente 0,177 y la distancia para el servidor 104c es aproximadamente 0,035. De ese modo, el dispositivo de distribución de carga 101a selecciona y asigna el servidor 101d para el cual la distancia es corta, como el servidor óptimo.

[0108] En cuanto a la distancia diag entre las coordenadas de recursos de consumo y el origen, a partir de la ecuación (10), esta distancia es aproximadamente 0,763 para el servidor 101c y aproximadamente 0,602 para el servidor 101d. El valor del producto de los dos valores (dis x diag) es aproximadamente 0,135 para el servidor 101c y aproximadamente 0,021 para el servidor 101d. Por lo tanto, aun cuando tanto la longitud de la línea óptima de consumo de recursos como la distancia entre las coordenadas de recursos de consumo y el origen se tienen en cuenta simultáneamente, el servidor 101d es seleccionado y asignado como el servidor óptimo.

[0109] Cuando los servidores a los que se ha de distribuir son pocos tales como 101b, 101c y 101d, a veces es más sencillo que la información sea escrita en un archivo para referencia mediante un archivo de texto, etc., y actualizado sin ninguna administración de tabla utilizando la base de datos tal como se describió anteriormente. En tal caso, el contenido de la tabla mostrada en la fig. 17 puede escribirse como un archivo de texto según sea necesario y el formato del archivo puede determinarse arbitrariamente y crearse. Aunque no se muestra, pueden establecerse según sea necesario no sólo los índices de uso de las CPU y las memorias sino también elementos tales como las direcciones IP, los números de puerto, etc. para acceder a cada servidor.

(Contenido de distribución regional de los servidores)

[0110] Se describirá la distribución de carga a los servidores administrados según la realización de la presente invención. La fig. 18 es una vista explicativa que muestra un ejemplo de una distribución regional de los servidores del sistema servidor/cliente según la presente invención. Tal como se muestra en la fig. 18, un dispositivo de distribución de carga 1801a y servidores 1803a a 1803n instalados en un centro 1803 están dispuestos en una región A; un dispositivo terminal 102b, un dispositivo de distribución de carga 1801b y servidores 1804a a 1804n

instalados en un centro 1804 están dispuestos en una región B; y un dispositivo terminal 102c, un dispositivo de distribución de carga 1801c y servidores 1805a a 1805n instalados en un centro 1805 están dispuestos en una tienda C. De este modo se muestra un ejemplo de un sistema servidor/cliente complejo que utiliza aparatos distribuidos en regiones o tiendas.

[0111] Una línea A1 que indica el flujo de una petición de ejecución de un proceso desde el dispositivo terminal 102a dispuesto en la región A, es transmitida por una navegación de usuario 1802 al dispositivo de distribución de carga 1801a a través de una línea A2. La navegación de usuario 1802 se refiere a un sistema que identifica la región a partir de las direcciones IP de un cliente y ejecuta el encaminamiento; identifica regiones a partir de direcciones de servidores DNS locales y ejecuta el encaminamiento; etc. De este modo puede identificarse la región de un cliente y se le puede asignar prioridad a un servidor dispuesto en una región más cercana (es decir, el retardo de la transmisión de datos es pequeño en términos de la red) al cliente. La petición de la ejecución del proceso transmitida al dispositivo de distribución de carga 1801a a través de la línea A2 hace que el dispositivo de distribución de carga 1801a distribuya la carga a, y transmita la petición de la ejecución del proceso al servidor 1803a dispuesto en la misma región a través de una línea A3.

[0112] Una línea B1 que indica una petición de ejecución de un proceso procedente del dispositivo terminal 102b dispuesto en la región B, es transmitida por la navegación de usuario 1802 al dispositivo de distribución de carga 1801b a través de una línea B2. Normalmente, la petición de la ejecución del proceso es transmitida a uno cualquiera de los servidores 1804a a 1804n dispuestos en la región B. Sin embargo, cuando el dispositivo de distribución de carga 1801a y el dispositivo de distribución de carga 1801b están organizados para intercambiar información de monitorización de recursos de grupos de servidores administrados respectivamente por los dispositivos de distribución de carga 1801a y 1801b, se permite que la carga sea distribuida al servidor 1803a dispuesto en otra región que es la región A a través de una línea B3. Cuando se hace que el dispositivo de distribución de carga 1801b monitorice los recursos de todos los servidores de las regiones A y B y la tienda C, los servidores a los que se ha de distribuir la carga pueden seleccionarse de entre todos los servidores además del servidor 1803a dispuesto en la región A.

[0113] Una línea C1 que identifica una petición de ejecución de un proceso procedente del dispositivo terminal 102c dispuesto en la tienda C, es transmitida por la navegación de usuario 1802 al dispositivo de distribución de carga 1801c a través de una línea C2. El dispositivo de distribución de carga 1801c evalúa los servidores a los que se puede distribuir la carga de los servidores 1805a a 1805c en la tienda C. Sin embargo, cuando no está presente ningún servidor que guarde recursos que puedan ser asignados en la misma tienda C que el dispositivo terminal 102c, la carga puede ser distribuida y asignada al servidor 1804n en la siguiente región B a través de una línea C3.

[0114] Se sabe que, en cuanto a los accesos a muchas aplicaciones de Internet incluyendo juegos, etc., la frecuencia de los accesos varía según la hora del día. Por ejemplo, se supone una aplicación, para la cual el máximo de los accesos a la misma es a las 10:00 p.m. y la frecuencia de accesos se reduce en varias decenas de porcientos respecto a la de la hora de máxima demanda en las horas anteriores y posteriores al máximo, por ejemplo, 9:00 p.m. y 11:00 p.m. Teniendo en cuenta la diferencia horaria entre el este y el oeste, se permite una distribución de carga regional este-oeste, según la cual, cuando los recursos de los servidores en la región que está en la banda de hora de máxima demanda están agotados, los servidores dispuestos en regiones que tienen respectivamente una diferencia horaria de una hora más temprano y más tarde aceptan la petición de procesar el proceso procedente de clientes que están en la banda de hora de máxima demanda y procesan el proceso como recursos extra. Incluso en el ecuador, la diferencia horaria media de una hora corresponde a un poco menos de 1700 km y el tiempo de ida y vuelta generado por el retardo de propagación es aproximadamente 17 ms. Según lo anterior, el retardo es un retardo a un nivel que puede utilizarse satisfactoriamente incluso en aplicaciones para las cuales la transmisión se lleva a cabo frecuentemente entre un servidor y un cliente, tal como un juego, etc., y la carga de aplicaciones que tienen máximos de las mismas dependiendo de la hora del día puede distribuirse eficazmente.

[0115] Cuando sólo queda la capacidad de recursos que es consumida por el proceso que se ha solicitado que sea ejecutado, los servidores que han de ser destinos de distribución de carga pueden estar dispuestos en cualquier parte independientemente de la región. Así, suscribiendo contratos para alquilar parcialmente recursos tales como servidores de alojamiento utilizados como servidores web que generalmente no consumen recursos significativamente, una pluralidad de PC servidores instalados en cibercafés en diversos lugares, etc., y para utilizar esos ordenadores como servidores a los que se ha de distribuir la carga, se permite una distribución de carga redundante.

[0116] Tal como se describió anteriormente, según la realización, evaluando el uso del consumo de recursos consumido cuando es ejecutado el proceso que se ha solicitado que sea ejecutado desde el dispositivo terminal 102, los servidores que pueden ejecutar el proceso pueden seleccionarse de entre los servidores 101b a 101n. Cada vez que el proceso que ha de ser ejecutado es añadido o modificado nuevamente, la ejecución del proceso siempre puede ser asignada al servidor óptimo registrando el último consumo de recursos. Como los servidores son seleccionados utilizando el consumo de recursos del proceso que se ha solicitado que sea ejecutado y la cantidad restante de los recursos guardados por los servidores 101b a 101n que siempre están siendo monitorizados, puede impedirse la discordancia de la cantidad de recursos consumidos por el proceso y la cantidad restante del recurso de los servidores.

[0117] Evaluando con la función que hace que los recursos de los servidores sean consumidos sin desperdiciar ninguno de ellos utilizando el consumo de recursos del proceso que se ha solicitado que sea ejecutado y la cantidad restante de los recursos guardados por los servidores 101b a 101n que siempre están siendo monitorizados, pueden seleccionarse los servidores óptimos que consumen los recursos sin desperdiciar ninguno de ellos, y pueden seleccionarse flexiblemente los servidores para ejecutar juegos y programas de aplicación, etc., para los cuales la cantidad de recursos consumidos difiere según el proceso.

[0118] De entre una pluralidad de servidores que guardan respectivamente diferentes cantidades de recursos, pueden seleccionarse los servidores que tienen las cantidades de recursos restantes que corresponden a la cantidad de recursos consumidos por el proceso que se desea que sea ejecutado. De ese modo, los recursos de los servidores son consumidos idealmente, pueden maximizarse los procesos (el número de juegos, etc.) que son procesados simultáneamente por la pluralidad de servidores conjuntamente, y puede utilizarse totalmente la capacidad sin desperdiciar nada de ella para cada parámetro de recursos de cada servidor.

[0119] Cuando se utiliza el dispositivo de distribución de carga 101a de la realización, puede impedirse la distribución de la carga a servidores que no pueden ejecutar el proceso debido a escasez de cantidades de recursos restantes, o que es probable que pongan el proceso en cola de espera. Por lo tanto, puede reducirse el tiempo de espera después de la petición de ejecución de un proceso.

[0120] El procedimiento de distribución de carga descrito en la realización puede realizarse ejecutando un programa preparado de antemano, en un ordenador tal como un ordenador personal, una estación de trabajo, etc. Este programa está grabado en un medio de grabación legible por ordenador tal como un disco duro, un disco flexible, un CD-ROM, un MO, un DVD, etc., y es ejecutado siendo leído del medio de grabación por el ordenador. Este programa puede ser un medio de transmisión capaz de ser distribuido a través de una red tal como Internet, etc.

APLICACIÓN INDUSTRIAL

[0121] Tal como se describió anteriormente, el sistema servidor/cliente, el procedimiento de distribución de carga y el programa de distribución de carga resultan útiles para hacer que los recursos de los servidores que son destinos de distribución de carga que han de ser consumidos estén bien equilibrados, y resultan particularmente adecuados para asignar procesos, tales como juegos, etc. que muchos usuarios solicitan que sean ejecutados, a una pluralidad de servidores.

REIVINDICACIONES

1. Un sistema servidor/cliente en el cual una pluralidad de servidores (101a, 101b a 101n) y una pluralidad de clientes (102) están conectados a través de una red (100), y los servidores (101b a 101n) ejecutan un proceso basándose en una petición de proceso procedente de los clientes (102) y transmiten un resultado de proceso a los clientes (102), en el que
al menos uno de los servidores (101a) incluye

una unidad de recepción de información de proceso (401) configurada para recibir información sobre el proceso procedente de los clientes (102) a través de la red (100);
una unidad de determinación (402) configurada para determinar un servidor (101b a 101n) para ejecutar el proceso de entre los servidores (101b a 101n) basándose en la información sobre el proceso; y
una unidad de transmisión de información de servidor (403) configurada para transmitir la información sobre el servidor determinado (101b a 101n) a los clientes (102), y

cada uno de los clientes (102) incluye

una unidad de recepción de información de servidor (412) configurada para recibir la información sobre el servidor (101b a 101n); y
una unidad de transmisión de petición de proceso (413) configurada para transmitir la petición de proceso a los servidores determinados (101b a 101n); **caracterizado porque**

la unidad de determinación (402) incluye

una primera unidad de cálculo (404) configurada para calcular, para cada uno de los servidores, una primera distancia desde un punto de estimación que indica un consumo estimado hasta una línea de consumo ideal, el consumo estimado obtenido añadiendo una cantidad de recursos que han de ser consumidos por la ejecución del proceso hasta una punto que indica una cantidad de recursos que han sido consumidos por cada uno de los servidores (101b a 101n), siendo la línea de consumo ideal una línea recta que conecta un origen y un punto que indica una capacidad de recursos máxima de cada uno de los servidores (101b a 101n) expresada en un espacio donde los ejes son los parámetros de recursos; y
la unidad de determinación (402) está configurada para determinar el servidor (101b a 101n) con la primera distancia más corta.

2. El sistema servidor/cliente según la reivindicación 1, en el que los parámetros incluyen al menos uno de una cantidad de carga de una unidad de procesamiento central, una cantidad de carga de una memoria de sistema, una cantidad de carga de una unidad de procesamiento gráfico, una cantidad de carga de una memoria de vídeo y una cantidad de carga de una tarjeta de interfaz de red.

3. Un procedimiento de distribución de carga utilizado en un sistema servidor/cliente, en el cual una pluralidad de servidores y una pluralidad de clientes están conectados a través de una red, y los servidores ejecutan un proceso basándose en una petición de proceso procedente de los clientes y transmiten un resultado de proceso a los clientes, que comprende las etapas de
recibir información sobre el proceso procedente del cliente a través de la red;
determinar un servidor para ejecutar el proceso de entre los servidores basándose en la información sobre el proceso; y
transmitir la petición de proceso a los servidores determinados; **caracterizado porque** la determinación incluye
calcular, para cada uno de los servidores, una primera distancia desde un punto de estimación que indica un consumo estimado hasta una línea de consumo ideal, el consumo estimado obtenido añadiendo una cantidad de recursos que han de ser consumidos por la ejecución del proceso a un punto que indica una cantidad de recursos que han sido consumidos por cada uno de los servidores, siendo la línea de consumo ideal una línea recta que conecta un origen y un punto que indica una capacidad de recursos máxima de cada uno de los servidores expresada en un espacio donde los ejes son los parámetros de recursos; y

la determinación incluye la determinación del servidor con la primera distancia más corta nce.

4. El procedimiento de distribución de carga según la reivindicación 3, en el que los parámetros incluyen al menos uno de una cantidad de carga de una unidad de procesamiento central, una cantidad de carga de una memoria de sistema, una cantidad de carga de una unidad de procesamiento gráfico, una cantidad de carga de una

memoria de vídeo y una cantidad de carga de una tarjeta de interfaz de red.

5. Un programa de distribución de carga para distribuir cargas de servidores en un sistema servidor/cliente en el cual una pluralidad de servidores y una pluralidad de clientes están conectados a través de una red, y los servidores ejecutan un proceso basándose en una petición de proceso procedente del cliente y transmiten un resultado de proceso al cliente, haciendo el programa de distribución de carga que los servidores ejecuten las etapas de:

recibir información sobre el proceso procedente del cliente a través de la red;
determinar un servidor para ejecutar el proceso de entre los servidores basándose en la información sobre el proceso; y
transmitir la petición de proceso a los servidores determinados; **caracterizado porque** la determinación incluye calcular, para cada uno de los servidores, una primera distancia desde un punto de estimación que indica un consumo estimado hasta una línea de consumo ideal, el consumo estimado obtenido añadiendo una cantidad de recursos que han de ser consumidos por la ejecución del proceso a un punto que indica una cantidad de recursos que han sido consumidos por cada uno de los servidores, siendo la línea de consumo ideal una línea recta que conecta un origen y un punto que indica una capacidad de recursos máxima de cada uno de los servidores expresada en un espacio donde los ejes son los parámetros de recursos; y

la determinación incluye determinar el servidor con la primera distancia más corta.

6. El programa de distribución de carga según la reivindicación 5, en el que los parámetros incluyen al menos uno de una cantidad de carga de una unidad de procesamiento central, una cantidad de carga de una memoria de sistema, una cantidad de carga de una unidad de procesamiento gráfico, una cantidad de carga de una memoria de vídeo y una cantidad de carga de una tarjeta de interfaz de red.

FIG.1

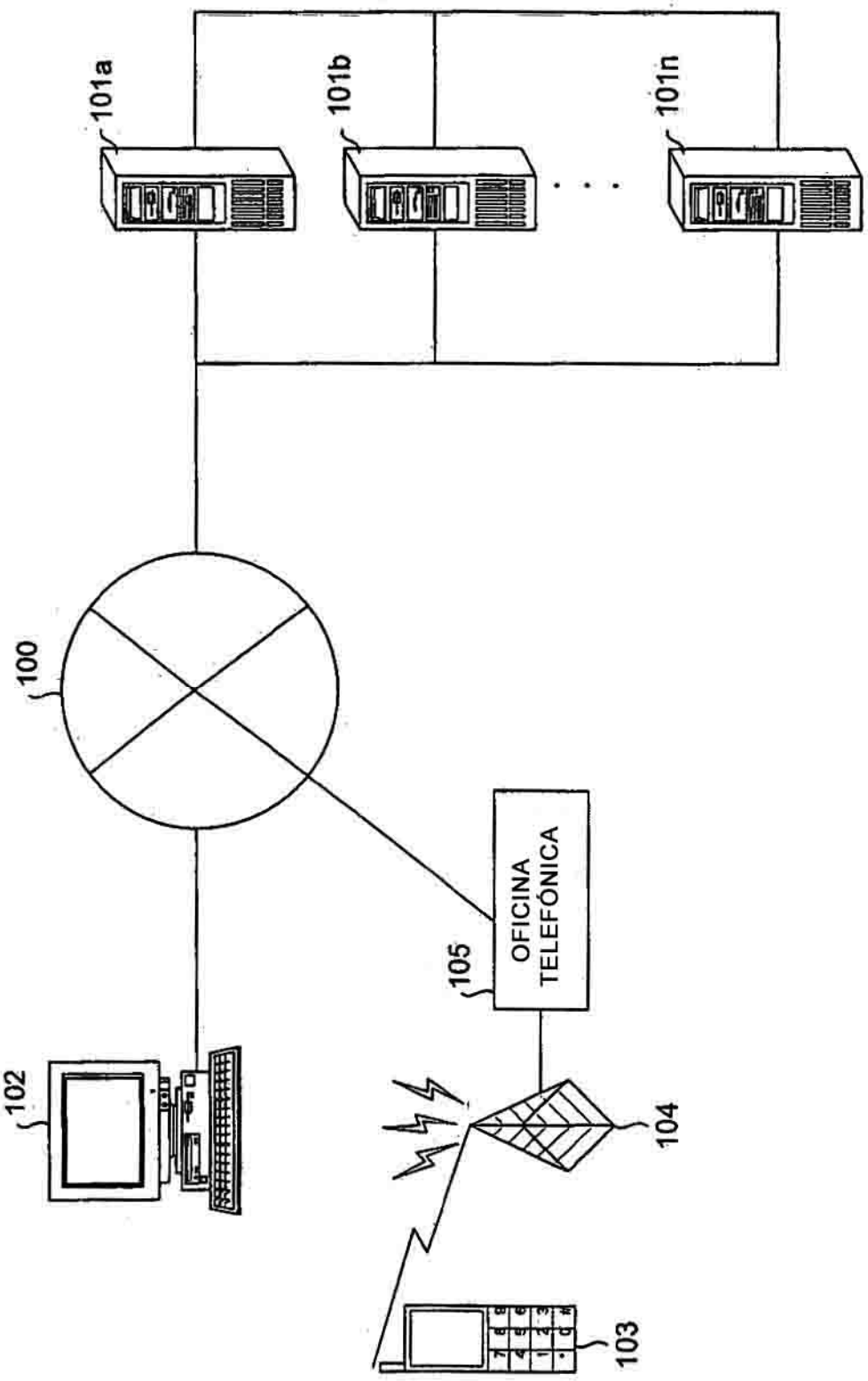


FIG.2

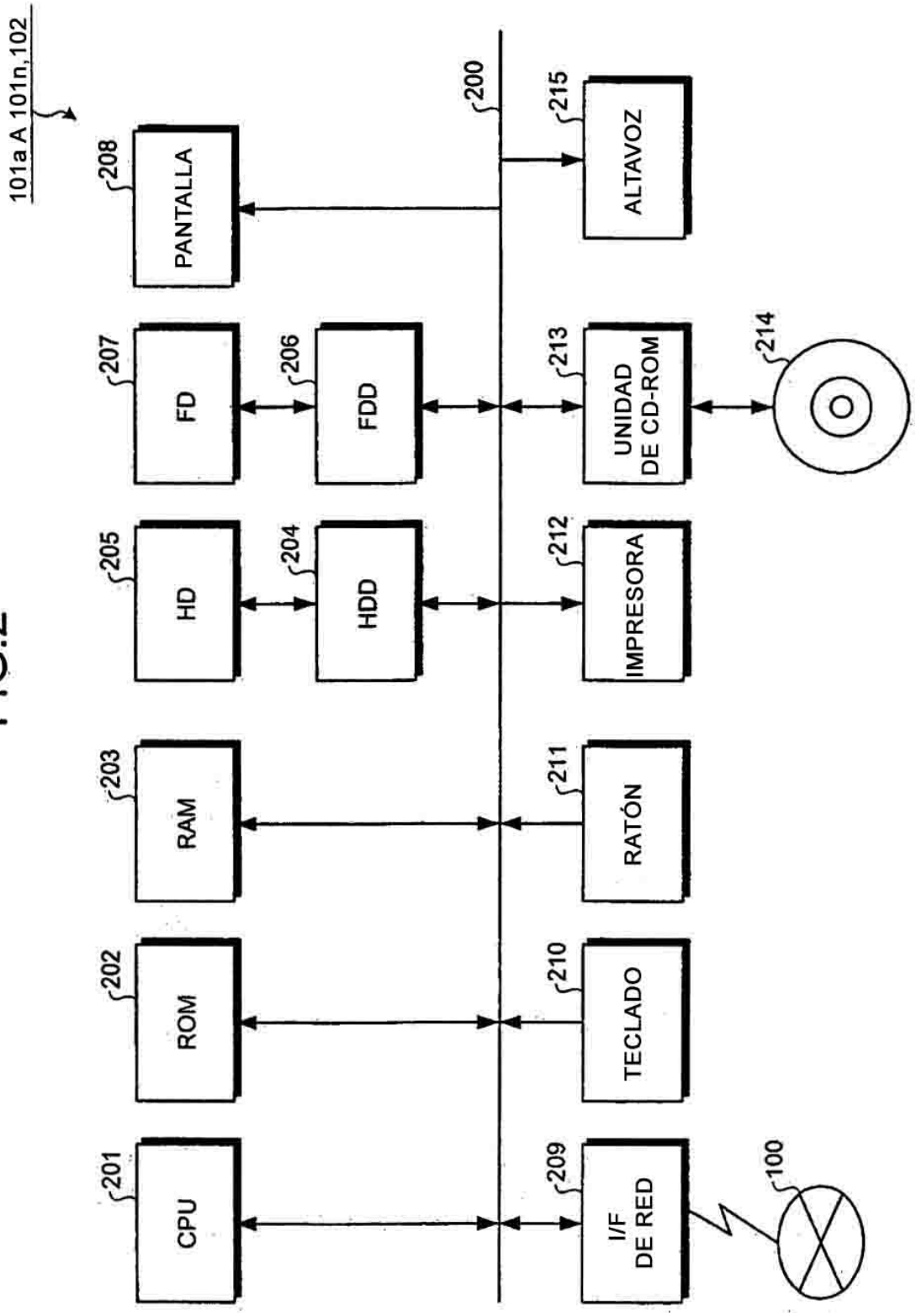

101a A 101n,102
200
208
PANTALLA
215
ALTAVOZ
207
FD
206
FDD
213
UNIDAD DE CD-ROM
214
205
HD
204
HDD
212
IMPRESORA
203
RAM
211
RATÓN
202
ROM
210
TECLADO
201
CPU
209
I/F DE RED
100

FIG.3

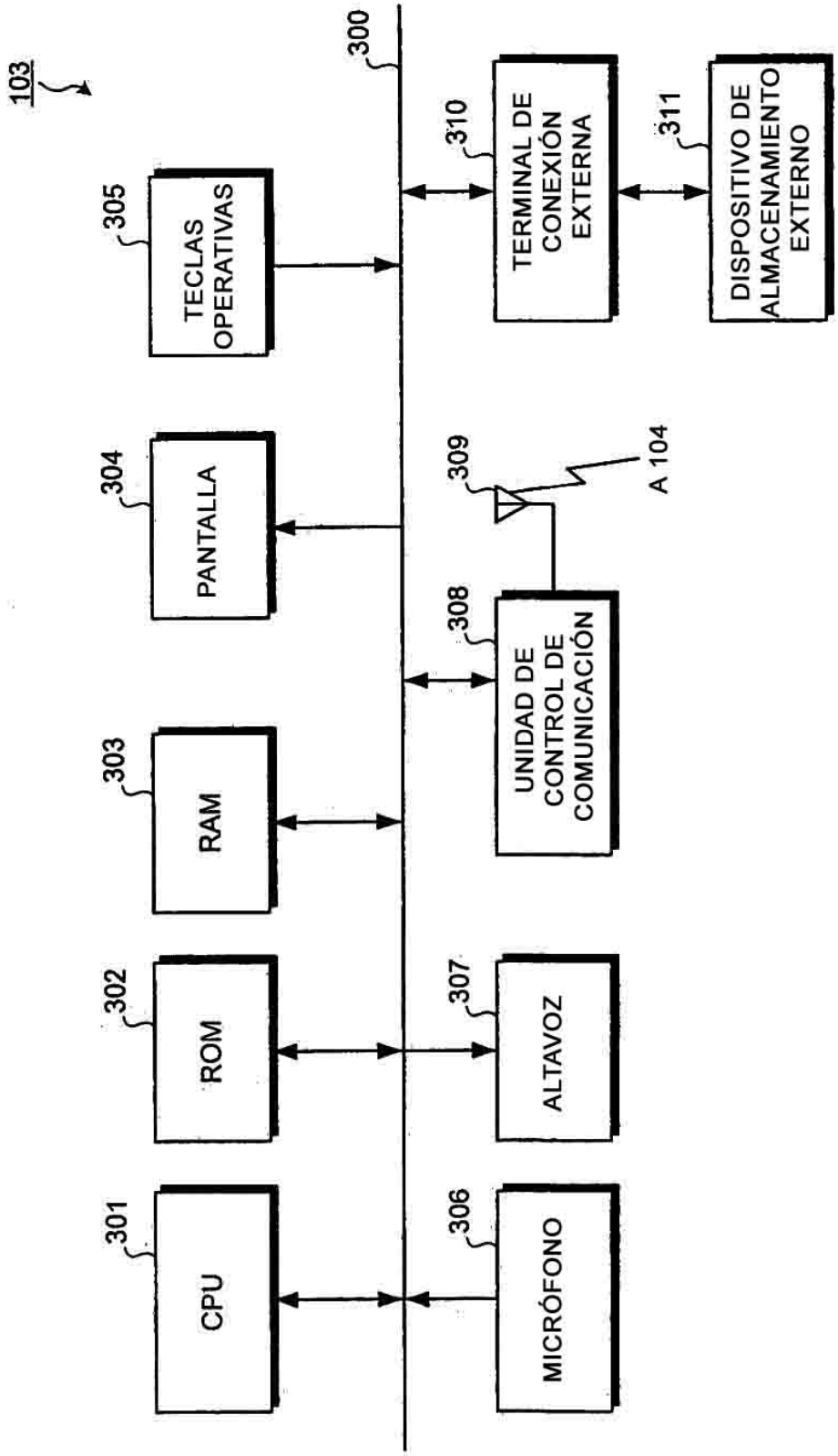

103
300
305
TECLAS OPERATIVAS
310
TERMINAL DE CONEXIÓN EXTERNA
311
DISPOSITIVO DE ALMACENAMIENTO EXTERNO
304
PANTALLA
309
A 104
308
UNIDAD DE CONTROL DE COMUNICACIÓN
303
RAM
302
ROM
307
ALTAVOZ
301
CPU
306
MICRÓFONO

FIG.4

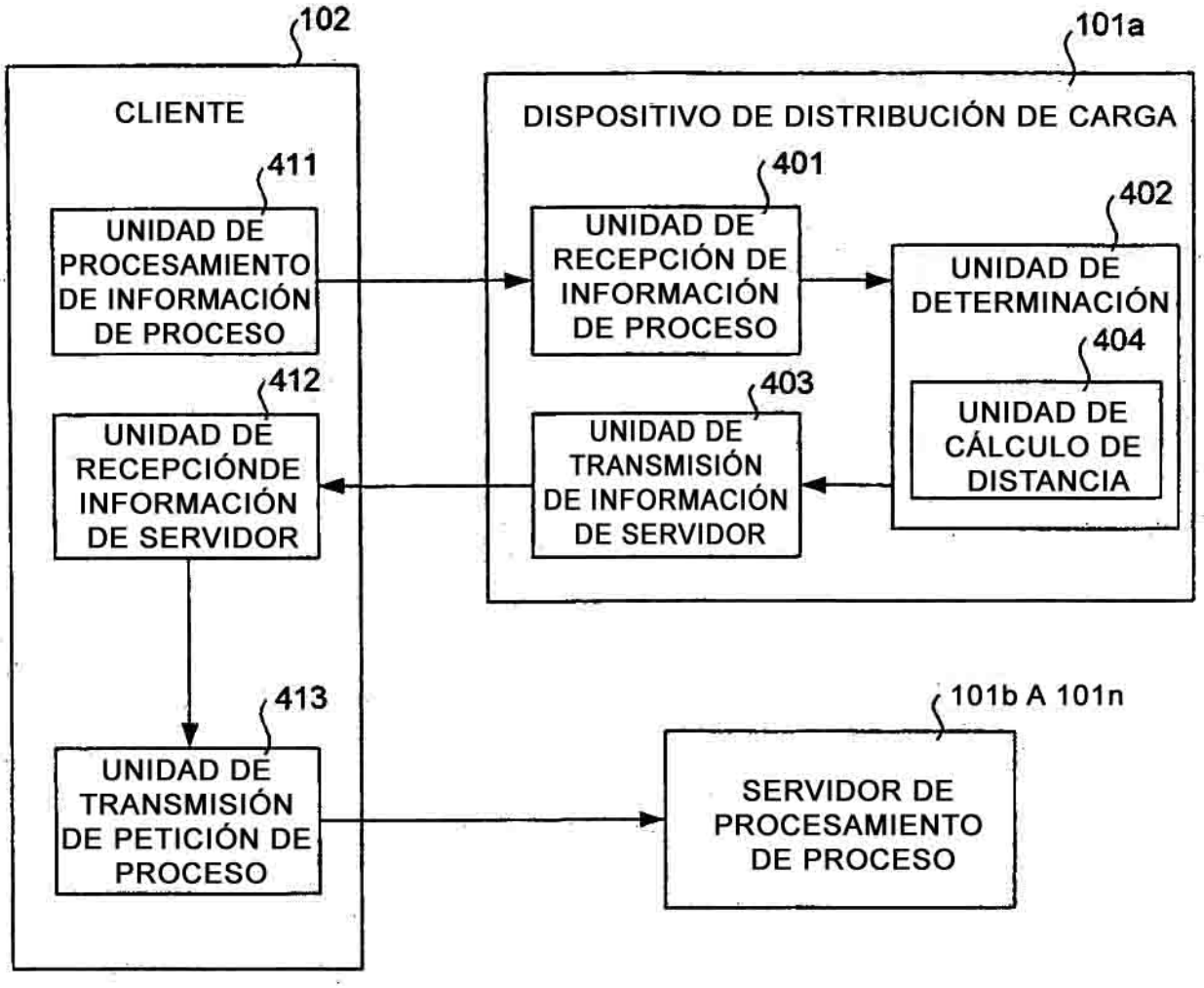

102
CLIENTE
411
UNIDAD DE PROCESAMIENTO DE INFORMACIÓN DE PROCESO
412
UNIDAD DE RECEPCIÓNDE INFORMACIÓN DE SERVIDOR
413
UNIDAD DE TRANSMISIÓN DE PETICIÓN DE PROCESO
101a
DISPOSITIVO DE DISTRIBUCIÓN DE CARGA
401
UNIDAD DE RECEPCIÓN DE INFORMACIÓN DE PROCESO
402
UNIDAD DE DETERMINACIÓN
404
UNIDAD DE CÁLCULO DE DISTANCIA
403
UNIDAD DE TRANSMISIÓN DE INFORMACIÓN DE SERVIDOR
101b A 101n
SERVIDOR DE PROCESAMIENTO DE PROCESO

FIG.5

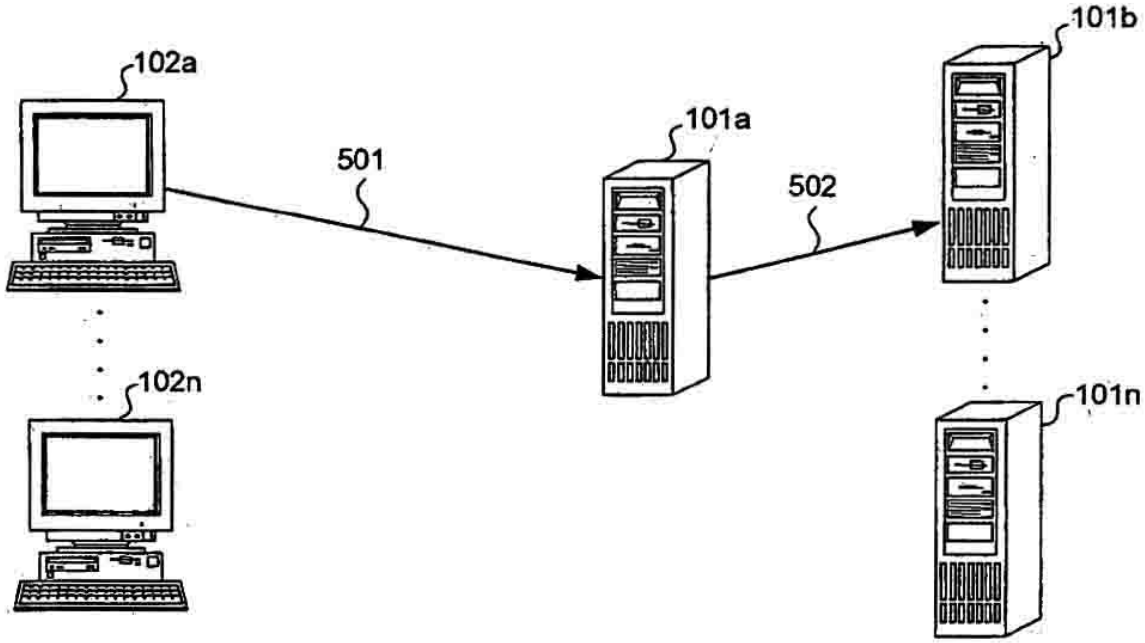

102a
501
101a
502
101b
102n
101n

FIG.6

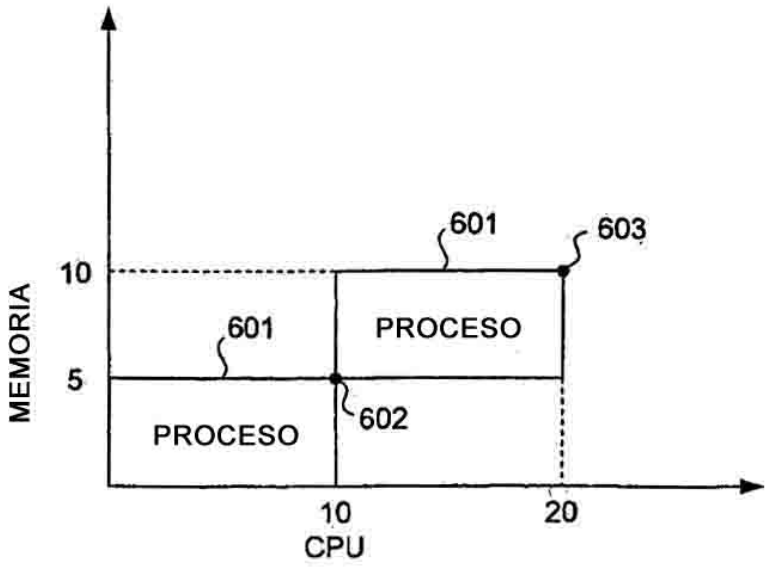

MEMORIA
10
5
601
603
601
PROCESO
602
PROCESO
10
20
CPU

FIG.7

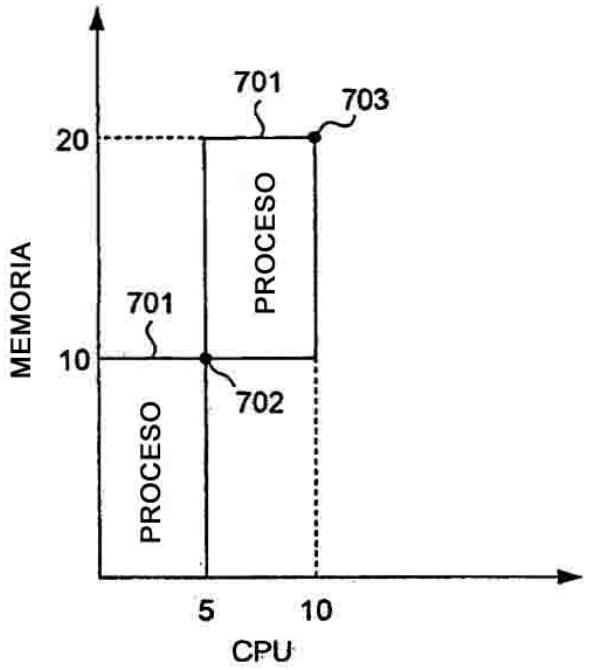

701
703
20
PROCESO
MEMORIA
701
10
702
PROCESO
5
10
CPU

FIG.8

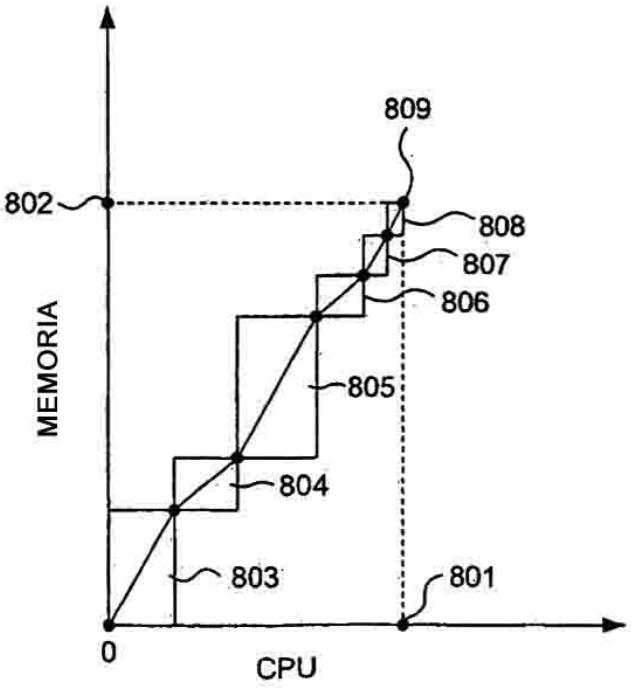

809
802
808
807
806
MEMORIA
805
804
803
801
0
CPU

FIG.9

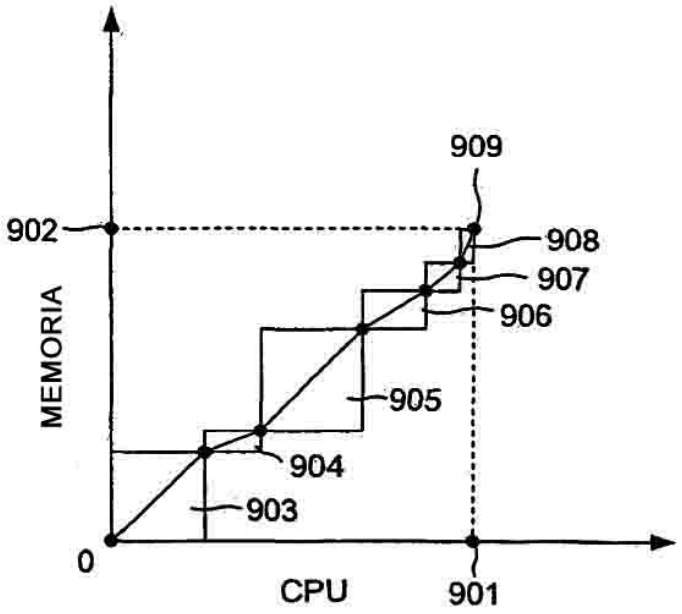

909
902
908
907
906
MEMORIA
905
904
903
0
CPU
901

FIG.10

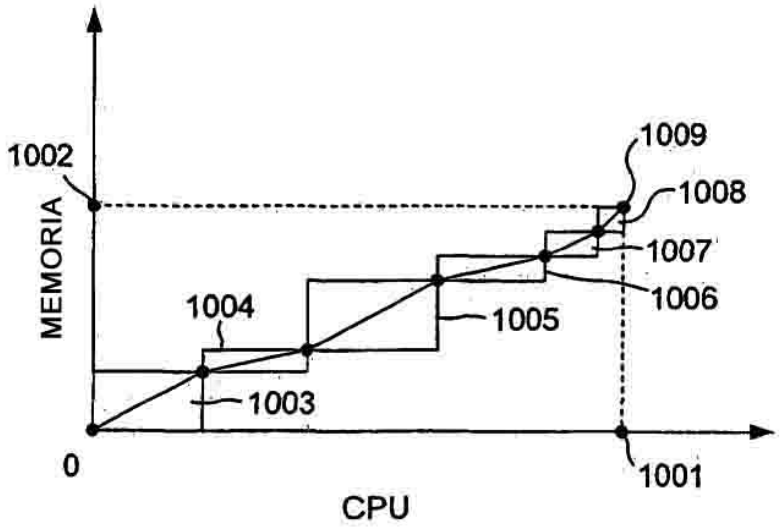

1009
1002
1008
1007
1006
1004
1005
MEMORIA
1003
0
1001
CPU

FIG.11

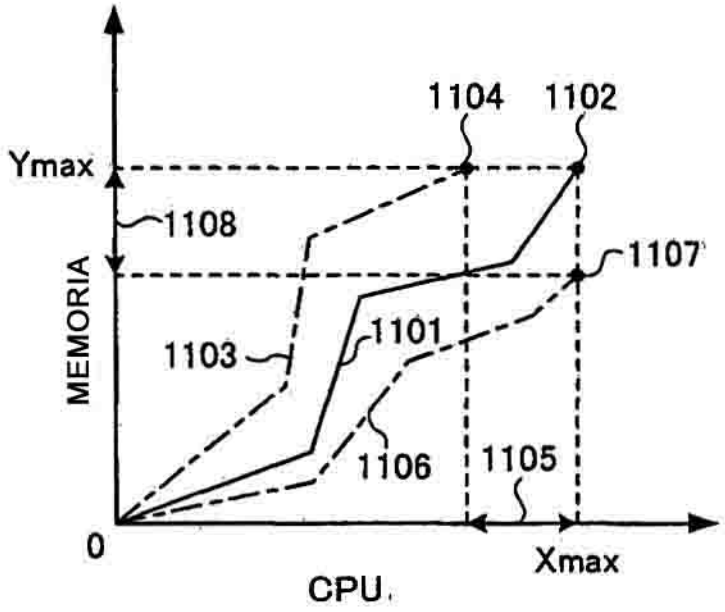

1104
1102
Ymax
1108
1107
MEMORIA
1103
1101
1106
1105
0
Xmax
CPU

FIG.12

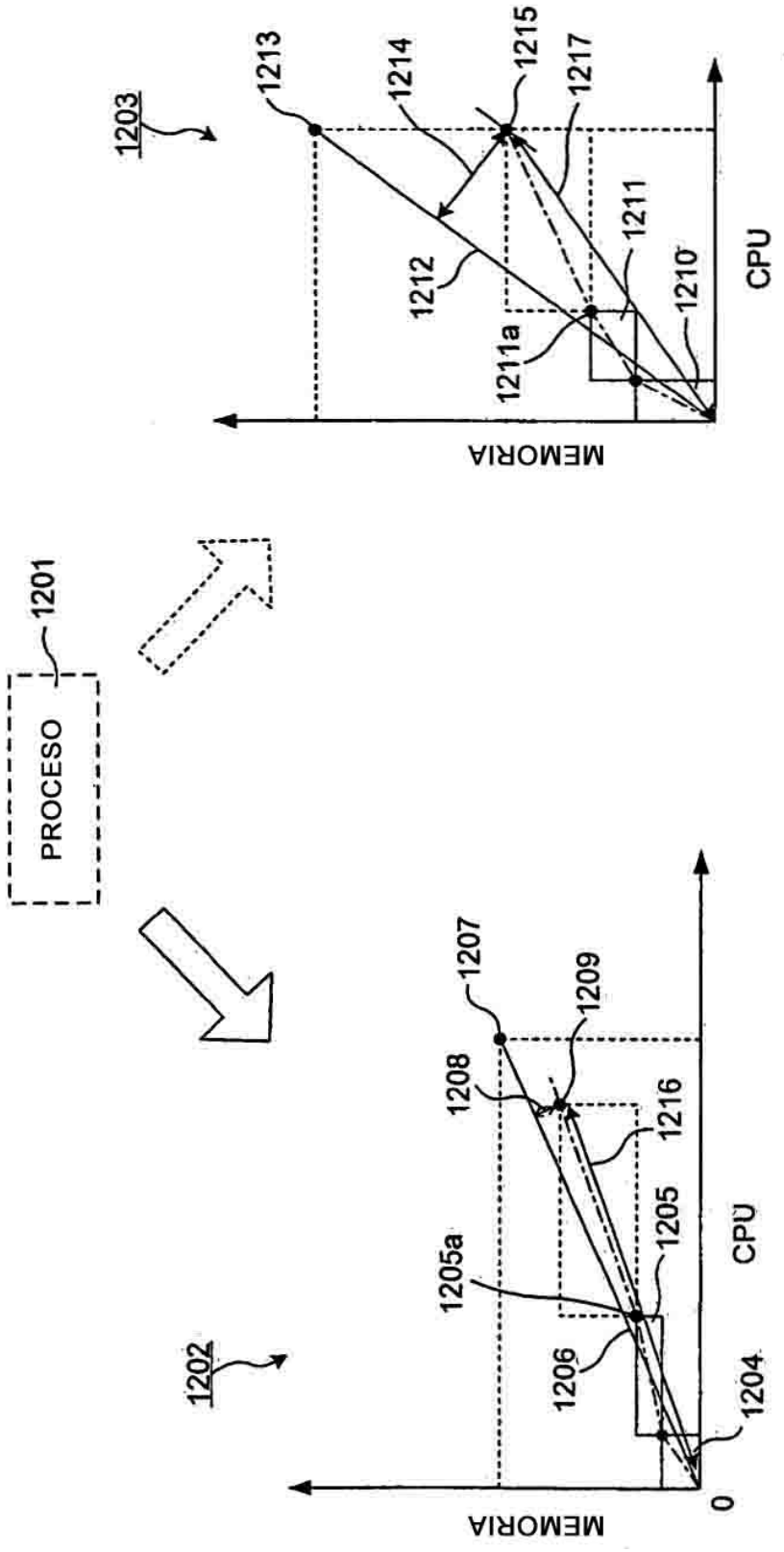

PROCESO
1201
1202
1203
MEMORIA
CPU
0
1204
1205
1205a
1206
1207
1208
1209
1210
1211
1211a
1212
1213
1214
1215
1216
1217

FIG.13

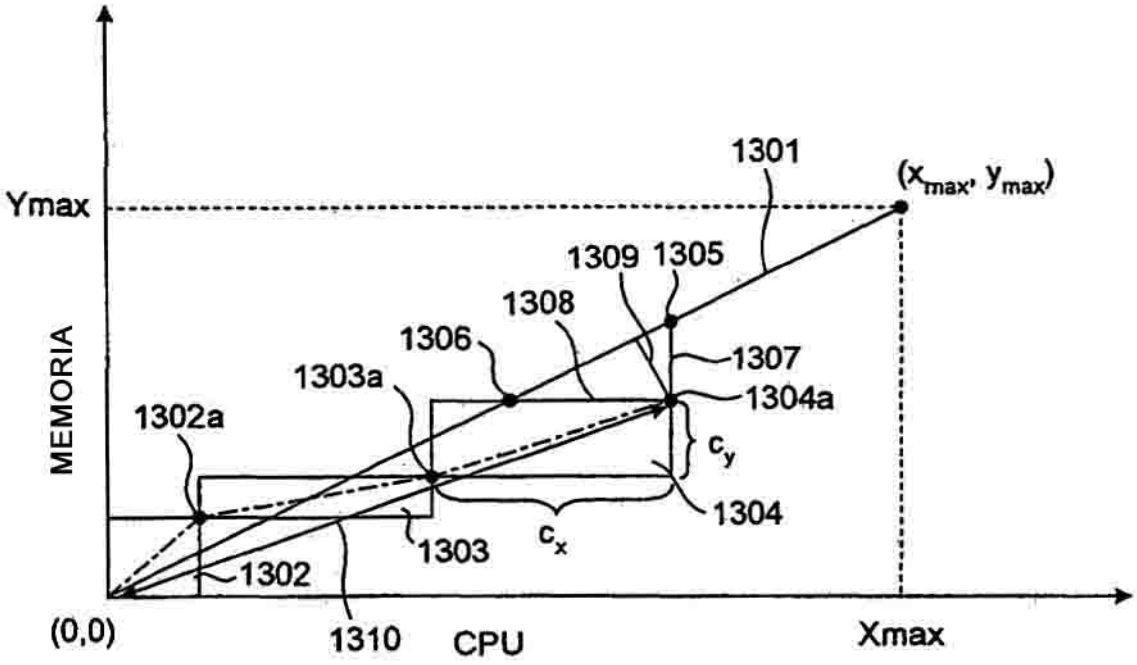

Ymax
MEMORIA
1301
(x_{max}, y_{max})
1309
1305
1308
1306
1307
1303a
1304a
1302a
c_y
1304
c_x
1303
1302
(0,0)
1310
CPU
Xmax

FIG.14

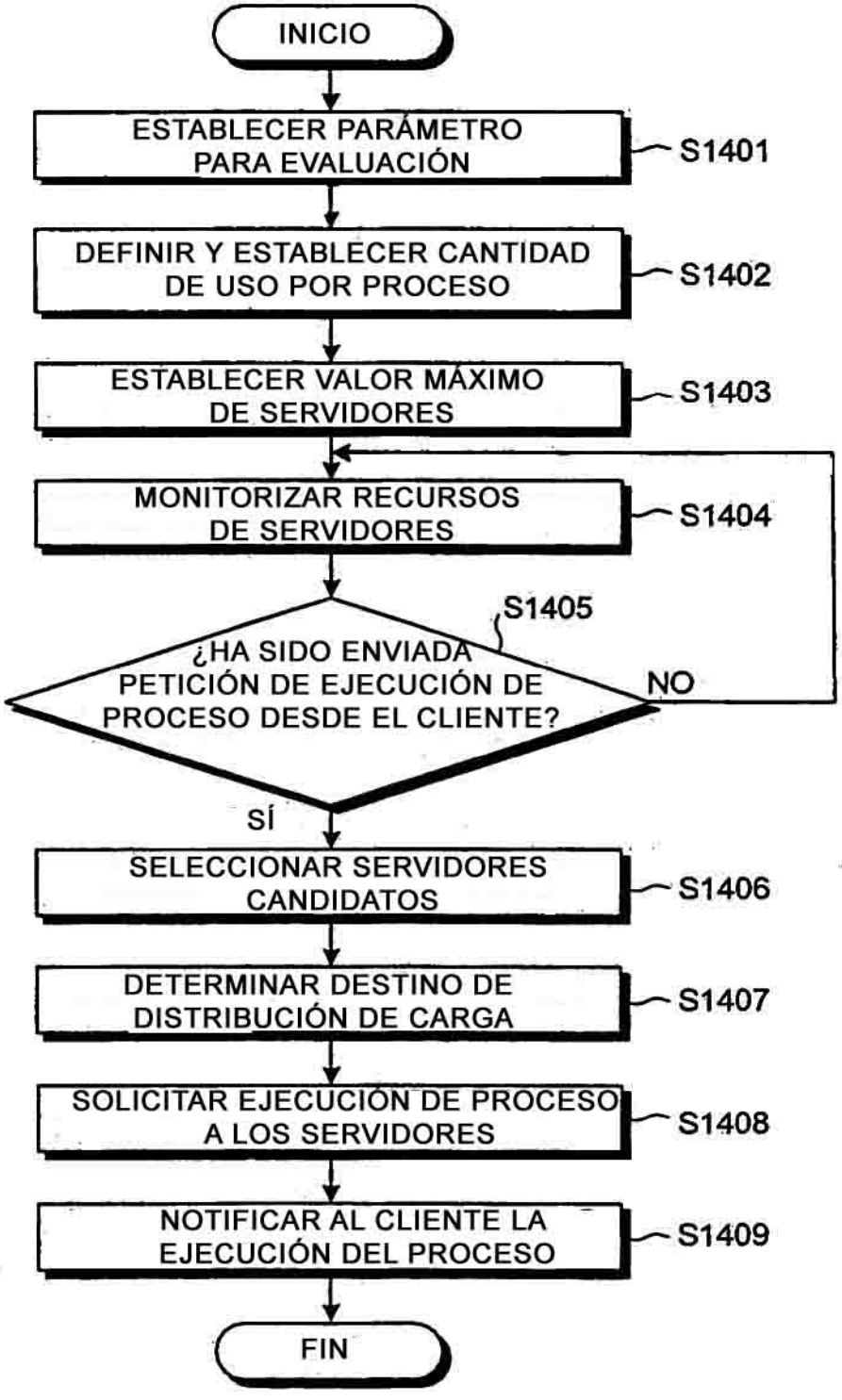

INICIO
ESTABLECER PARÁMETRO PARA EVALUACIÓN
S1401
DEFINIR Y ESTABLECER CANTIDAD DE USO POR PROCESO
S1402
ESTABLECER VALOR MÁXIMO DE SERVIDORES
S1403
MONITORIZAR RECURSOS DE SERVIDORES
S1404
S1405
¿HA SIDO ENVIADA PETICIÓN DE EJECUCIÓN DE PROCESO DESDE EL CLIENTE?
NO
SÍ
SELECCIONAR SERVIDORES CANDIDATOS
S1406
DETERMINAR DESTINO DE DISTRIBUCIÓN DE CARGA
S1407
SOLICITAR EJECUCIÓN DE PROCESO A LOS SERVIDORES
S1408
NOTIFICAR AL CLIENTE LA EJECUCIÓN DEL PROCESO
S1409
FIN

FIG.15

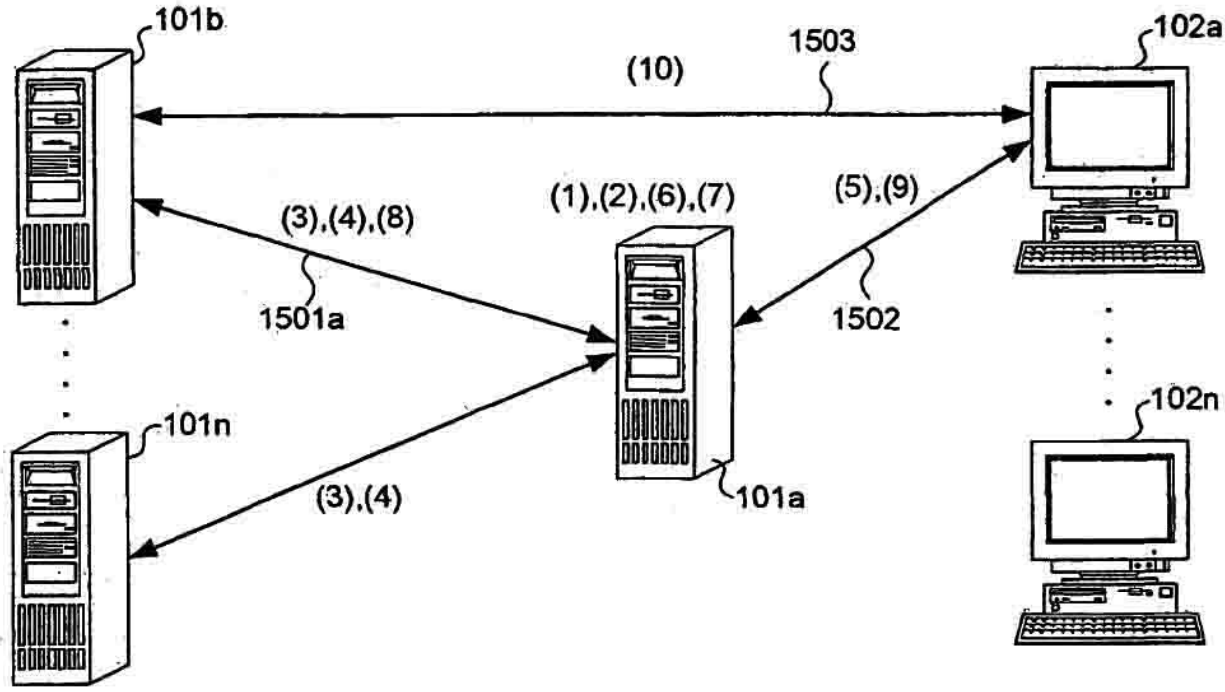


FIG.16

1600

ID	NOMBRE DE PROCESO	CANTIDAD DE CPU UTILIZADA	CANTIDAD DE MEMORIA UTILIZADA
100	PROCESO 1	0,15	0,2
110	PROCESO 2	0,2	0,4
. . .			

FIG.17

1700

NOMBRE DE SERVIDOR DE DESTINO DE CONEXIÓN	ÍNDICE DE USO DE CPU	ÍNDICE DE USO DE MEMORIA
101b	0,9	0,88
101c	0,5	0,2
101d	0,3	0,2
. . .		

FIG.18

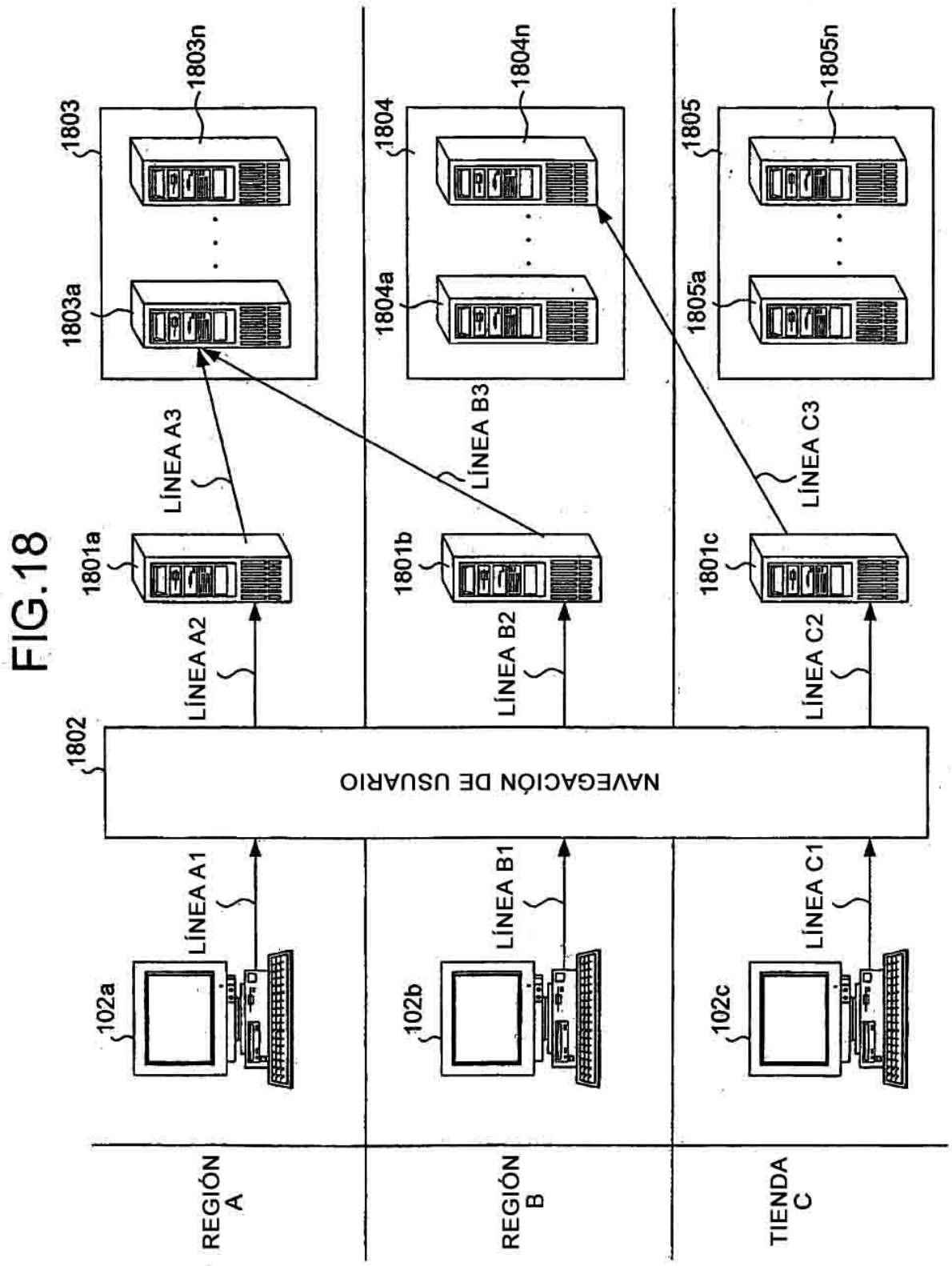

REGIÓN A
REGIÓN B
TIENDA C
NAVEGACIÓN DE USUARIO
1802
102a
102b
102c
LÍNEA A1
LÍNEA B1
LÍNEA C1
1801a
1801b
1801c
LÍNEA A2
LÍNEA B2
LÍNEA C2
LÍNEA A3
LÍNEA B3
LÍNEA C3
1803
1803a
1803n
1804
1804a
1804n
1805
1805a
1805n

FIG.19

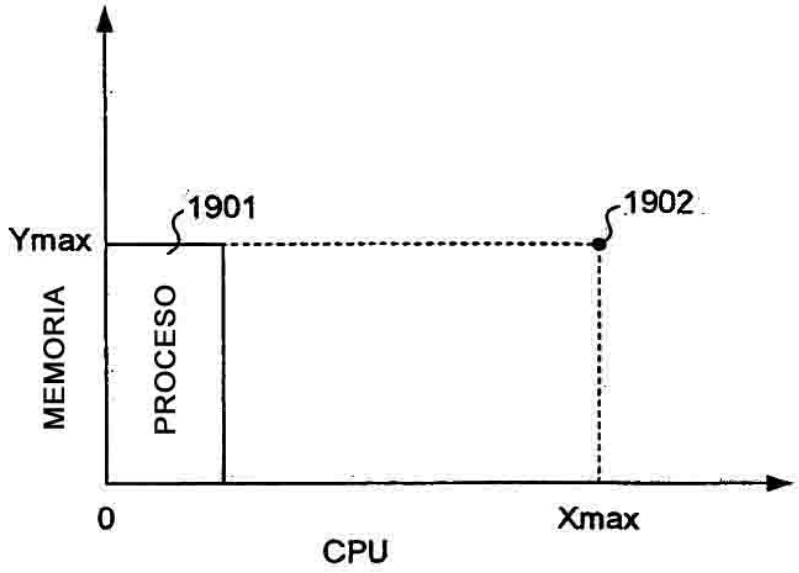
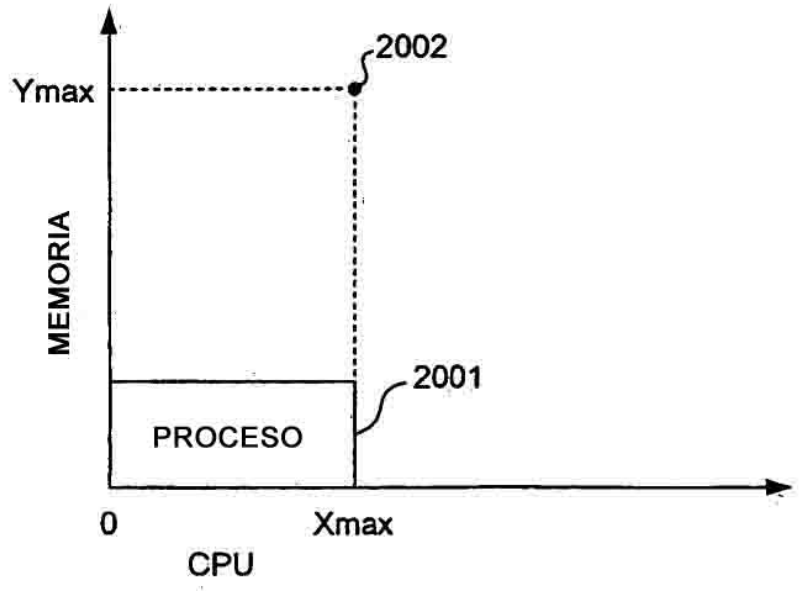

1901
1902
Ymax
MEMORIA
PROCESO
0
Xmax
CPU

FIG.20


2002
Ymax
MEMORIA
2001
PROCESO
0
Xmax
CPU