(19) OFICINA ESPAÑOLA DE PATENTES Y MARCAS

ESPAÑA

(11) Número de publicación: **2 993 459**

(51) Int. CI.:

H04W 28/08 (2013.01)
G06N 20/00 (2009.01)
G06N 3/006 (2013.01)
G06N 3/044 (2013.01)
H04W 24/02 (2009.01)
H04W 28/086 (2013.01)
H04W 36/22 (2009.01)
G06N 3/088 (2013.01)

(12) TRADUCCIÓN DE PATENTE EUROPEA T3

(86) Fecha de presentación y número de la solicitud internacional: **04.11.2021 PCT/KR2021/015939**

(87) Fecha y número de publicación internacional: **12.05.2022 WO22098131**

(96) Fecha de presentación y número de la solicitud europea: **04.11.2021 E 21889596 (9)**

(97) Fecha y número de publicación de la concesión europea: **24.07.2024 EP 4150956**

(54) Título: **Aparato y procedimiento para el equilibrio de la carga en un sistema de comunicación inalámbrica**

(30) Prioridad:

06.11.2020 US 202063110515 P
30.06.2021 US 202117363918

(45) Fecha de publicación y mención en BOPI de la traducción de la patente:
30.12.2024

(73) Titular/es:

SAMSUNG ELECTRONICS CO., LTD. (100.00%)
129, Samsung-ro, Yeongtong-gu
Suwon-si, Gyeonggi-do 16677, KR

(72) Inventor/es:

KANG, JIKUN;
CHEN, XI;
WU, DI;
XU, YI TIAN;
LIU, XUE;
DUDEK, GREGORY LEWIS;
LEE, TAESEOP y
PARK, INTAIK

(74) Agente/Representante:

GONZÁLEZ PECES, Gustavo Adolfo

DESCRIPCIÓN

Aparato y procedimiento para el equilibrio de la carga en un sistema de comunicación inalámbrica

Campo técnico

La presente divulgación está relacionada con la mejora del equilibrio de carga en un sistema de comunicación utilizando inteligencia artificial (IA).

Antecedentes de la técnica

La presente divulgación se refiere a el equilibrio de carga de un sistema de comunicación inalámbrica. En un sistema de comunicaciones, existen equipos de usuario (EU) en modo inactivo y equipos de usuario en modo activo. El sistema de comunicación puede experimentar una carga desequilibrada entre las celdas del sistema de comunicación. La carga desequilibrada hace que algunas celdas tengan demasiados usuarios mientras que otras están poco cargadas. Los usuarios de las celdas más sobrecargadas pueden sufrir retrasos, un bajo rendimiento o altas tasas de error. Además, es posible que el operador del sistema desee utilizar una fracción mayor del ancho de banda del sistema para los datos de carga útil del usuario en lugar de para la señalización relacionada con la reselección de celda y/o la señalización relacionada con el traspaso.

Dado que existen usuarios en modo inactivo y usuarios en modo activo, el problema es complejo. Por ejemplo, después de cambiar un usuario de modo inactivo a una segunda celda, el usuario de modo inactivo puede convertirse en un usuario de modo activo más ventajosamente asociado con una celda distinta de la segunda celda. Provocar un traspaso en esta situación aumenta el ancho de banda del sistema dedicado a la señalización y puede ser perceptible para el usuario.

Divulgación de la invención

Solución al Problema

Existe un problema en la asignación de recursos radioeléctricos para el acceso múltiple. En el escenario de acceso múltiple, algunos equipos de usuario (EUs) están transmitiendo/recibiendo activamente en un sistema de comunicación (por ejemplo, una red de radio) y algunos otros EUs están asociados con el sistema de comunicación, pero los otros EUs no están transmitiendo/recibiendo activamente. Cuando un EU inactivo pasa a un estado activo, puede convertirse en una carga para la celda de radio en la que estaba acampado anteriormente. Además, debido al problema, un usuario del sistema de comunicación experimenta retrasos, bajo rendimiento y/o alta tasa de errores. Este problema puede reducirse desplazando a algunos usuarios en modo inactivo a otras celdas e influyendo en algunos traspasos de usuarios en modo activo.

Existe un segundo problema en el consumo de energía eléctrica de las estaciones base. A veces una estación base está funcionando, consumiendo energía eléctrica, y la estación base no necesita estar funcionando.

La solución utiliza un paradigma de aprendizaje automático denominado aprendizaje por refuerzo. El aprendizaje por refuerzo modela un sistema como un Proceso de Decisión de Markov (PDM) que incluye acciones y recompensas a medida que un sistema cambia de un estado a otro. Una política elige acciones para aumentar la acumulación de recompensas a lo largo de una secuencia de acciones.

La solución ajusta la asociación de los EU inactivos con las celdas de radio y la asociación de los EU activos con las celdas de radio formando una política inferior para los EU inactivos y una política superior para los EU activos.

Para el problema de la energía eléctrica, la política inferior se dirige a decidir si se apaga un aspecto de una estación base. Por ejemplo, una acción de la política inferior puede apagar un dispositivo de radio de una estación base concreta.

En la aplicación, la política superior se utiliza para seleccionar una primera acción que afecta a los EUs activos. A continuación, la solución establece un objetivo basado en la primera acción. Se elige una segunda acción basada en la política inferior modificada. De este modo, se acoplan las políticas que controlan los EU activos y los inactivos, y se mejoran las características medibles del sistema. Ejemplos de características medibles del sistema son la desviación estándar del rendimiento (equidad), el rendimiento medio del sistema (número único que caracteriza el rendimiento del sistema) y la moderación de la tasa de traspaso (número único relacionado con la sobrecarga del traspaso).

Una realización proporciona un procedimiento para equilibrar la carga en un sistema de comunicación inalámbrica, el procedimiento comprende: recibir una utilización de ancho de banda, un número de equipos de usuario activos (EUs), y un rendimiento medio de al menos una estación base (EB); obtener primeros datos de estado basados en la utilización de ancho de banda, el número de EUs activos, y el rendimiento medio; obtener primeros datos de acción asociados con un equilibrio de carga de EU activo (ECEA) y primeros datos de objetivo para obtener segundos datos de acción, introduciendo los primeros datos de estado y los primeros datos de recompensa en un primer modelo de aprendizaje por refuerzo; obtener segundos datos de acción asociados con un equilibrio de carga de EU inactivo (ECEI), introduciendo los primeros datos de estado, los primeros datos de objetivo y los segundos datos de

recompensa en un segundo modelo de aprendizaje por refuerzo; determinar terceros datos de acción asociados tanto con el ECEA como con el ECEI, basándose en los primeros datos de acción y los segundos datos de acción; y transmitir los terceros datos de acción a la al menos una estación base (EB).

Otras realizaciones proporcionan un aparato para equilibrar la carga en un sistema de comunicación inalámbrica, el aparato comprende: un transceptor; una memoria que almacena una o más instrucciones; y al menos un procesador conectado al transceptor y configurado para ejecutar la una o más instrucciones almacenadas en la memoria para: recibir una utilización de ancho de banda, un número de equipos de usuario activos (EUs), y un rendimiento medio de al menos una estación base (EB); obtener unos primeros datos de estado basados en la utilización de ancho de banda, el número de EUs activos, y el rendimiento medio; obtener unos primeros datos de acción asociados con un equilibrio de carga de EU activo (ECEA) y unos primeros datos de objetivo para obtener unos segundos datos de acción, introduciendo los primeros datos de estado y unos primeros datos de recompensa en un primer modelo de aprendizaje por refuerzo; obtener unos segundos datos de acción asociados con un equilibrio de carga de EU inactivo (ECEI), introduciendo los primeros datos de estado, los primeros datos de objetivo y los segundos datos de recompensa en un segundo modelo de aprendizaje por refuerzo; determinar unos terceros datos de acción asociados tanto con el ECEA como con el ECEI, basándose en los primeros datos de acción y los segundos datos de acción; y transmitir los terceros datos de acción a la al menos una estación base (EB).

Otras realizaciones proporcionan un medio de almacenamiento legible por ordenador que almacena un programa ejecutable por al menos un procesador para realizar un procedimiento para equilibrar la carga, el procedimiento comprende: recibir una utilización de ancho de banda, un número de equipos de usuario activos (EUs), y un rendimiento medio desde al menos una estación base (EB); obtener primeros datos de estado basados en la utilización de ancho de banda, el número de EUs activos, y el rendimiento medio; obtener primeros datos de acción asociados con un equilibrio de carga de EU activo (ECEA) y primeros datos de objetivo para obtener segundos datos de acción, introduciendo los primeros datos de estado y los primeros datos de recompensa a un primer modelo de aprendizaje por refuerzo; obtener segundos datos de acción asociados con un equilibrio de carga de EU inactivo (ECEI) introduciendo los primeros datos de estado, los primeros datos de objetivo y los segundos datos de recompensa en un segundo modelo de aprendizaje por refuerzo; determinar terceros datos de acción asociados tanto con el ECEA como con el ECEI, basándose en los primeros datos de acción y los segundos datos de acción; y transmitir los terceros datos de acción a la al menos una estación base (EB).

La invención se define por medio de las reivindicaciones adjuntas.

Breve descripción de los dibujos

El texto y las figuras se proporcionan únicamente como ejemplos para ayudar al lector a comprender la invención. No se pretende ni debe interpretarse que limitan en modo alguno el alcance de esta invención. Aunque se han proporcionado ciertas realizaciones y ejemplos, será evidente para los expertos en la materia basados en las divulgaciones contenidas en la presente memoria descriptiva que pueden realizarse cambios en las realizaciones y ejemplos mostrados sin apartarse del alcance de las realizaciones contenidas en la presente memoria descriptiva.

La Figura 1A ilustra un flujo lógico para determinar conjuntamente una primera política y una segunda política y aplicar una acción unida a una red de comunicación, de acuerdo con algunas realizaciones.
La Figura 1B ilustra un sistema 1-19 de un servidor de determinación de políticas para determinar conjuntamente una primera política y una segunda política y aplicar una acción unida a una red de comunicación, de acuerdo con algunas realizaciones.
Similar a FIG. 1B, la FIG. 1C ilustra un sistema 1-99 que muestra los EUs de acampada 1-91, los EUs activos 1-92, siendo servidos por medio de las celdas 1-93 siendo influenciados, controlados u optimizados por medio de un servidor de determinación de políticas 1-9, de acuerdo con algunas realizaciones.
La Figura 1D ilustra un ejemplo de sistema de comunicación 1-4, de acuerdo con algunas realizaciones.
La Figura 2A ilustra un flujo lógico 2-8, de acuerdo con algunas realizaciones.
La Figura 2B ilustra una arquitectura ejemplar 2-9 de aprendizaje jerárquico de políticas con el sistema de comunicación 1-4 como entorno de ejemplo.
La Figura 3 ilustra un diagrama de rebote de mensajes 3-1, de acuerdo con algunas realizaciones.
La Figura 4 ilustra un flujo de algoritmo 4-1, de acuerdo con algunas realizaciones.
La Figura 5 ilustra un aparato para implementar una o más de las realizaciones, por ejemplo el servidor de determinación de políticas de la FIG. 1B.
La FIG. 6A ilustra un Proceso de Decisión de Markov (PDM) 6-1, de acuerdo con algunas realizaciones.
La Figura 6B ilustra el pseudocódigo ejemplar 6-11, un ejemplo más específico del flujo lógico 2-8, de acuerdo con algunas realizaciones.
La Figura 7A ilustra la lógica básica de las políticas de cooperación 7-8, de acuerdo con algunas realizaciones.
La Figura 7B ilustra un sistema 7-19 asociado a la lógica de la FIG. 7A, de acuerdo con algunas realizaciones.
La Figura 8 ilustra un sistema ejemplar 8-4 en el que interactúan el dispositivo de aprendizaje por refuerzo 8-5 y el entorno 8-1, de acuerdo con algunas realizaciones.

Modo para la invención

Las comunicaciones celulares han penetrado en todos los rincones de nuestra vida cotidiana. Para responder a nuestra creciente demanda de comunicaciones, se han desplegado celdas por todo el territorio para ofrecer mejores servicios. Sin embargo, debido a restricciones normativas y de ingeniería, las celdas no pueden desplegarse arbitrariamente.

Esto conduce a un desajuste entre la distribución relativamente uniforme de las celdas y la desigual distribución demográfica de las personas. Como resultado, un sistema celular suele ser testigo de cargas muy desequilibradas entre distintas celdas, lo que se traduce en usuarios insatisfechos y recursos desaprovechados.

Se han realizado grandes esfuerzos para equilibrar la carga por medio de la migración de equipos de usuario (EU) entre celdas. Los procedimientos de equilibrio de carga (Load Balancing, EC) existentes se pueden dividir en dos categorías, el EC-EU activo (ECEA) y el EC-EU inactivo (ECEI). Los procedimientos ECEA utilizan el mecanismo de traspaso (T) para descargar los equipos de usuario en modo activo (es decir, los equipos de usuario que actualmente transmiten señales) de las celdas de servicio ocupadas a celdas vecinas menos ocupadas. Estos procedimientos consiguen resultados instantáneos de equilibrio de carga, pagando el precio de una mayor sobrecarga del sistema. Los procedimientos ECEI aprovechan el mecanismo de reselección de celda (RC) para trasladar los EU en modo inactivo (es decir, los EU conectados pero que no transmiten señales) de las celdas de acampada congestionadas a otras celdas. Estos procedimientos son más ligeros, ya que RC requiere menos sobrecarga del sistema que T. Sin embargo, el beneficio sólo se obtiene cuando los EU inactivos migrados pasan a estar activos.

Las acciones de ECEA pueden entrar en conflicto con las acciones de ECEI (y viceversa), resultando en una degradación inesperada del rendimiento del sistema.

Para mejorar el rendimiento del sistema, las realizaciones proporcionan un procedimiento de Aprendizaje Jerárquico de Políticas (AJP), que integra tanto ECEA como ECEI en una estructura de Aprendizaje por Refuerzo (AR) de dos niveles. Concretamente, el nivel superior ajusta las acciones ECEA, y el nivel inferior controla las acciones ECEI. El nivel superior pretende optimizar el rendimiento del sistema directamente como recompensa AR y, al mismo tiempo, aprende a establecer un objetivo para el nivel inferior. Dado que este objetivo se refiere en particular al nivel inferior y no al objetivo global del sistema, el objetivo del nivel inferior es una especie de subobjetivo.

El objetivo para el nivel inferior es un estado AR deseado, que potencia aún más la recompensa de nivel superior (y sin embargo no se puede lograr sólo con acciones de nivel superior). Al acercarse a este objetivo, el nivel inferior 1) mejora indirectamente el rendimiento del sistema, y 2) se ve influido para alinearse con el nivel superior.

De este modo, se establece una colaboración entre la ECEA y la ECEI, reduciendo o eliminando los conflictos.

Las realizaciones proporcionan AJP, el primer procedimiento de aprendizaje jerárquico que integra diferentes mecanismos EC (es decir, ECEA e ECEI) de forma colaborativa.

Con diferentes configuraciones de densidad de EU, nuestro procedimiento AJP siempre supera a los procedimientos SOTA. En concreto, en comparación con una combinación directa de SOTA ECEA e ECEI, AJP mejora el rendimiento medio hasta en un 24%, al tiempo que reduce la desviación estándar del rendimiento hasta en un 31%.

El término "carga" se refiere al número de EUs que están siendo servidos. Una EB se encuentra en un emplazamiento físico, en el que se colocan los dispositivos de acceso radioeléctrico. Consideremos una red celular con NB EBs, cada una de las cuales consta de NS sectores no solapados. Un sector da servicio a los EU situados en una determinada dirección de su EB anfitriona. Un sector admite frecuencias portadoras NC, cada una de las cuales corresponde a una celda. Una celda es una entidad de servicio que atiende a los EU dentro de una determinada dirección de una EB y en una determinada frecuencia portadora.

La Figura 1A ilustra el flujo lógico 1-8. El flujo lógico 1-8 es una lógica de una política inferior y una política superior. En 1-10, el flujo lógico determina conjuntamente una política superior 1-1 y una política inferior 1-2. Basándose en la política superior 1-1 y la política inferior 1-2, se proporciona una acción unida 1-3 a un sistema de comunicación 1-4. En 1-20, la acción unida 1-3 se aplica al sistema de comunicación 1-4.

La Figura 1B ilustra un sistema 1-19 y la evolución de la política. El sistema 1-19 es un sistema influenciado por la política superior 1-1 y la política inferior 1-2. El sistema incluye un servidor de determinación de políticas 1-5 y el sistema de comunicación 1-4. Una serie de hechos temporales se indexa con números entre paréntesis. Se puede decir que todos los eventos (1), (2), (3) y (4) ocurren en una época, y que eventos similares con datos diferentes ocurren en la siguiente época. Una época también puede denominarse iteración.

En el evento (1), se proporciona un conjunto de parámetros del sistema 1-6 desde el sistema de comunicación 1-4 al servidor de determinación de políticas 1-5. En el evento (2), se realiza la lógica de 1-10 y se determinan o evolucionan (revisan o entrenan) la política superior 1-1 y la política inferior 1-2. En el evento (3), la acción unida 1-3 se proporciona al sistema de comunicación 1-4. En el evento (4), se ejecuta la lógica 1-20 y se aplica la acción unida 1-3 al sistema de comunicación 1-4. La evolución continúa entonces en el evento (5) con un conjunto actualizado de parámetros del sistema 1-6 proporcionado al servidor de determinación de políticas 1-5. Como en la época anterior, el servidor de determinación de políticas 1-5 ejecutará la lógica 1-10, y así sucesivamente.

La Figura 1C ilustra un sistema 1-99 incluyendo más detalles con respecto a las FIGS. 1A y 1B de un sistema de comunicación ejemplar 1-4 siendo influenciado, controlado u optimizado por el servidor de determinación de políticas 1-5. La Figura 1C es una representación esquemática que muestra los EUs de acampada denominados EUs 1-91 y los EUs activos denominados EUs 1-92. Los EUs acampados 1-91 incluyen un EU_j acampado en la celda k. Los EUs activos 1-92 incluyen un EU_m que transmite y recibe datos de usuario en la celda n.

La Figura 1D ilustra una porción de ejemplo del sistema de comunicación 1-4. En el ejemplo de la FIG. 1D, siete estaciones base soportan 21 celdas ilustradas. Cada celda tiene forma de tarta. La Figura 1D es de naturaleza esquemática, las celdas en la práctica se solapan geográficamente para proporcionar cobertura solapada de forma que cualquier EU pueda obtener servicio. Los puntos indican los EUs acampados 1-91 y los EUs activos 1-92. Se indica un ejemplo de EU_j de los EUs de acampada 1-91. También se indica un ejemplo de EU_m de los EUs activos 1-92.

La Figura 2A proporciona la lógica 2-8 que ilustra detalles lógicos adicionales ejemplares para realizar la lógica 1-8 de la FIG. 1A. Un sumario de la lógica 2-8 incluye determinar un conjunto de parámetros del sistema; actualizar una política superior basada en el conjunto de parámetros del sistema y la pérdida superior; seleccionar una acción superior basada en la política superior; seleccionar un objetivo para la política inferior basado en el conjunto de parámetros del sistema y la acción superior; actualizar una política inferior basada en el conjunto de parámetros del sistema, el objetivo y la pérdida inferior; seleccionar una acción inferior basada en el objetivo y la política inferior; y aplicar una acción unida al sistema de comunicación. La acción unida se basa en la acción inferior y la acción superior.

La lógica 2-8 se discute ahora en detalle. En 2-10, se obtiene un conjunto de parámetros del sistema 1-6. El conjunto de sistemas puede ser una versión actualizada como se indica en el evento (5) de la FIG. 1B. El conjunto de parámetros del sistema 1-6, a modo de ejemplo, incluye el número de EUs acampados en cada celda, el número de EUs activos en cada celda y estadísticas en tiempo real (o con bajo retardo) del caudal de datos de usuario para cada celda. Las estadísticas de rendimiento se utilizan para estimar las recompensas por acciones futuras. Véase la línea 5 del pseudocódigo de la FIG. 6B (analizado más adelante).

También pueden obtenerse recompensas en 2-10 del flujo lógico 2-8 (no mostrado). Se puede obtener una recompensa de nivel inferior utilizando, por ejemplo, la Ec. (10) (analizado más adelante). Se puede obtener una recompensa de mayor nivel utilizando el rendimiento medio durante un tiempo T de la Ec. (1) (analizado más adelante).

En 2-12, la política inferior 1-2 se actualiza basándose en el conjunto de parámetros del sistema 1-6 y en una pérdida de nivel superior 2-5. La política superior puede entrenarse, por ejemplo, utilizando un gradiente de una función de pérdida definida para la política superior. Un ejemplo de función de pérdida para expresar la pérdida de nivel superior 2-5 lo proporciona la Ec. (12) (analizado más adelante).

En 2-14, se selecciona una primera acción 2-1 basada en la política superior 1-1. La primera acción 2-1 está relacionada con los EUs activos 1-92.

En 2-16, se selecciona un objetivo 2-3 para la política inferior 1-2 basándose en el conjunto de parámetros del sistema 1-6 y basándose en la primera acción 2-1. Puede utilizarse una red neuronal de memoria a corto plazo (LSTM), por ejemplo, para seleccionar el objetivo 2-3. Un ejemplo de una LSTM 2-91 se indica en la FIG. 2B. En la FIG. se dan ejemplos de objetivos indexados por tiempo. 2B (véanse las variables g_0, g_1, ..., g_{k-1}, g_k, ...).

En 2-18, la política inferior 1-2 se actualiza basándose en el conjunto de parámetros del sistema 1-6 y en una pérdida de nivel inferior 2-4. La pérdida de nivel inferior 2-4 depende del objetivo 2-3. La política inferior 1-2 puede entrenarse, por ejemplo, utilizando un gradiente de una función de pérdida. Un ejemplo de función de pérdida, para expresar la pérdida de nivel inferior 2-4, se da a continuación en la Ec. (11).

En 2-20, se selecciona una segunda acción 2-2 basada en el objetivo 2-3. La segunda acción 2-2 está relacionada con los EUs de acampada 1-92.

La primera acción 2-1 y la segunda acción 2-2 se combinan entonces como una acción unida 1-3, por ejemplo, por concatenación. La concatenación puede indicarse algebraicamente por medio del símbolo"$\oplus$". Para un vector u = [a,b,c] y un vector v =[d,e,f], u $\oplus$v= [a,b,c,d,e,f].

En 2-22, la acción unida 1-3 se aplica al sistema de comunicación 1-4.

Como resultado de la acción unida 1-3, se mejoran las métricas de rendimiento 1-5 del sistema de comunicación 1-4 (esto se muestra en 2-24). Ejemplos de métricas de rendimiento se indican en la Ec.. (1). Ec. (2). Ec. (3) (analizado más adelante). Otra medida de rendimiento es la reducción del número de traspasos por celda y por unidad de tiempo.

La lógica entonces fluye de regreso como se indica en 2-26 a 2-10 para una siguiente iteración o época de tiempo.

La Figura 2B ilustra la evolución de la política superior y la política inferior en función del tiempo. En la parte inferior de la FIG. 2B, con el tiempo avanzando de izquierda a derecha. Se muestran las recompensas que contribuyen a la evolución de la política (r^H y r^L). Las acciones a^H de la política superior se muestran producidas por la política superior

en cada momento del tiempo (época, o iteración k). Las acciones a^L de la política inferior se muestran salidas por la política inferior en cada momento en el tiempo. Las acciones se combinan y luego se aplican al sistema de comunicación 1-4, indicado en la FIG. 2B como "entorno" Los objetivos son producidos por la LSTM 2-91 y las capas ocultas de la LSTM 2-91 se indican como h 2-90. Las capas ocultas proporcionan continuidad entre los objetivos a medida que avanza el tiempo. Más información sobre FIG. 2B se visitan de nuevo con un debate adicional a continuación.

De acuerdo con una realización, la arquitectura 2-9 puede incluir N (N>2) políticas. La política 1 puede significar la política más alta, y la política N puede significar la política más baja. El objetivo N-1 puede obtenerse basándose en la acción para la política N-1 y el estado actual, es decir, el generador de subobjetivos (descrito a continuación) puede incluirse en el módulo de refuerzo N-1 como la política superior 1-1.

La Figura 3 ilustra un diagrama de rebote 3-1 ejemplar que ilustra eventos del servidor de determinación de políticas 1-5 y el sistema de comunicación 1-4 de las FIGS. 1B y 1C.

Un eje de tiempo en el lado izquierdo de la FIG. 3 muestra el avance del tiempo de arriba a abajo de la figura. Generalmente, el servidor de determinación de políticas 1-5 configura el servidor de red 1-9. El servidor de red 1-9 configura las celdas 1-93. Los EUs de acampada 1-91 y los EUs activos 1-92 observan los niveles de potencia y las identidades de las celdas 1-93.

En la FIG. 3, entonces, el servidor de red 1-9 proporciona una versión actual del conjunto de parámetros del sistema 1-6 al servidor de determinación de políticas 1-5 en un mensaje 3-10. Este mensaje 3-10 puede pasar por un retorno de red, que puede ser cableado o inalámbrico.

Como se muestra en 3-11, el servidor de determinación de políticas 1-5 realiza entonces la lógica 1-10 para actualizar las políticas y generar la acción unida 1-3. La acción unificada 1-3 se envía en un mensaje de acción unificada 3-6 como evento 3-12 al servidor de red 1-9.

A continuación, el servidor de red 1-9 configura las celdas 1-93. Una configuración general ejemplar como la indicada como 3-21 para este caso de acción unida 1-3. La acción unida 1-3 incluye la primera acción 2-1 (relacionada con los EUs activos 1-91) y la segunda acción 2-2 (relacionada con los EUs de acampada 1-92).

A continuación, las celdas 1-93 actualizan los parámetros de reselección 3-50 y los EUs de acampada 1-91 se enteran de la actualización (por ejemplo, a través de un bloque de información del sistema (BIS) convencional, o similar). Véanse las expresiones (8) y (9) a continuación para más detalles de los parámetros de reselección de ejemplo 3-50.

De forma similar, las celdas 1-93 actualizan los parámetros de traspaso 3-60 y los EUs activos 1-91 se enteran de la actualización (por ejemplo, a través del bloque de información del sistema (BIS) convencional, o similar). Consulte la Expresión (4) a continuación para obtener más detalles de los parámetros de traspaso de ejemplo 3-60. En algunas realizaciones, los EUs activos 1-91 obtienen el parámetro de traspaso α de un mensaje BIS transmitido por las celdas 1-93 (reenviado a través del servidor 1-5 y/o el servidor 1-9) mientras están activos. Los EUs 1-91 en acampada (por ejemplo, en inactividad) obtienen los parámetros de reselección de celda β y γ a partir de un mensaje BIS transmitido por las celdas 1-93 (reenviado a través del servidor 1-5 y/o el servidor 1-9) mientras están en acampada.

Como se muestra en el evento 3-15, una porción 3-2 de los EUs de acampada 1-91 realiza la reselección 3-3 basándose en los parámetros de reselección 3-50.

Además, en 3-16, una porción 3-4 de los EUs activos 1-92 realiza el traspaso 3-5 basándose en los parámetros de traspaso 3-60.

El eje temporal no está a escala, y puede haber varios retrasos entre la configuración de las celdas, el aprendizaje de los parámetros actualizados por parte de los equipos de usuario, la observación de los niveles de potencia por parte de los equipos de usuario (RSRP, como se explica más adelante) y las acciones de reselección o traspaso por parte de los equipos de usuario.

El sistema sigue evolucionando, y el servidor de determinación de políticas 1-5 recibe en 3-17 un conjunto de parámetros del sistema 1-6 actualizado. En 3-18, se ejecuta la lógica 1-10, y en 3-19, se envía un mensaje de acción unida 3-6 para la siguiente iteración.

Los procesos continúan repitiéndose como muestran las elipses verticales (...) en la parte inferior de la FIG. 3.

Basándose en la evolución de las políticas y las acciones tomadas, se mejora la utilización de los recursos de radio del sistema de comunicación 1-4, mostrada como 3-20. Esta mejora se produce en cada iteración, aunque 3-20 se muestra como un único caso en la FIG. 3.

La Figura 4 ilustra el flujo del algoritmo 4-1, que es una vista general de la evolución de la FIG. 1B, la lógica 2-8 de la FIG. 2A y el diagrama de rebote de la FIG. 3.

En la Figura 4, en el estado 1 del algoritmo, se obtiene el conjunto de parámetros 1-6 del sistema. A modo de ejemplo, el conjunto de parámetros 1-6 del sistema puede incluir tres conjuntos: un conjunto que identifique los equipos de usuario 1-91 en acampada y en qué celda está acampado cada equipo de usuario, un conjunto que identifique los equipos de usuario 1-92 activos y en qué celda está transmitiendo/recibiendo cada equipo de usuario (puede tratarse de más de una celda simultáneamente), y un conjunto de datos de rendimiento que indiquen las estimaciones de rendimiento de cada celda para un periodo de tiempo reciente.

El algoritmo pasa entonces al estado de algoritmo 2, y la lógica de determinar conjuntamente 1-10 (ver FIG. 1A, véase también FIG. 2A puntos 2-12, 2-14, 2-16, 2-18 y 2-20).

A continuación, el algoritmo pasa al estado de algoritmo 3 en el que los parámetros de reselección 3-50 y los parámetros de traspaso 3-60 se actualizan en el sistema de comunicación 1-4 por medio de las acciones 2-1 y 2-2 (en conjunto, la acción unida 1-3).

Posteriormente al estado 3 del algoritmo, la segunda acción 2-2 es la causa de que una porción 3-2 de los EUs de acampada 1-91 realice la reselección, denominada colectivamente reselección 3-3; véase también la FIG. 3 a 3-15. También con posterioridad al estado 3 del algoritmo, la primera acción 2-1 es la causa de que una parte 3-4 de los EUs activos 1-92 inicien el traspaso, denominados colectivamente traspaso 3-5; véase también la FIG. 3 a 3-16.

El hardware para realizar las realizaciones proporcionadas en la presente memoria descriptiva se describe ahora con respecto a la FIG. 5.

La Figura 5 ilustra un aparato ejemplar 5-1 para la implementación de las realizaciones desveladas en la presente memoria descriptiva. El aparato 5-1 puede ser un servidor, un ordenador, un ordenador portátil, un dispositivo de mano o un dispositivo de tableta, por ejemplo. El aparato 5-1 puede incluir uno o más procesadores de hardware 5-9. El uno o más procesadores de hardware 5-9 puede incluir un CIAE (circuito integrado de aplicación específica), CPU (por ejemplo CISC o dispositivo RISC), y / o hardware personalizado. El aparato 5-1 también puede incluir una interfaz de usuario 5-5 (por ejemplo, una pantalla de visualización y/o un teclado y/o un dispositivo señalador tal como un ratón). El aparato 5-1 puede incluir una o más memorias volátiles 5-2 y una o más memorias no volátiles 5-3. La una o más memorias no volátiles 5-3 pueden incluir un medio legible por ordenador no transitorio que almacena instrucciones para su ejecución por el uno o más procesadores de hardware 5-1 para hacer que el aparato 5-1 realice cualquiera de los procedimientos de las realizaciones desveladas en la presente memoria descriptiva. En una realización de la divulgación, un transmisor y un receptor pueden denominarse colectivamente transceptor 5-4, y el transceptor 5-4 puede transmitir o recibir una señal hacia o desde un EU, una EB o una entidad de red. La señal transmitida o recibida puede incluir información y datos de control. Con este fin, el transceptor 5-4 puede incluir un transmisor de radiofrecuencia (RF) para convertir una frecuencia de y amplificar una señal para ser transmitida, y un receptor de RF para amplificar con bajo ruido y reducir la conversión de las señales recibidas Sin embargo, esto no es más que un ejemplo del transceptor 5-4, y por lo tanto, los elementos del transceptor 5-4 no se limitan al transmisor de RF y al receptor de RF. Además, el transceptor 5-4 puede recibir señales a través de canales alámbricos o inalámbricos y emitir las señales al procesador 5-4, y puede transmitir señales emitidas por el procesador 5-9 a través de canales alámbricos o inalámbricos.

La Figura 6A ilustra un ejemplo de punto de decisión conceptual en el progreso de un Proceso de Decisión de Markov (PDM) 6-1. Una primera decisión se denomina decisión "i" en la FIG. 6A, y una decisión alternativa se denomina decisión "w" en FIG. 6A. El estado se define como una tupla (S; A; R; P) (que se analiza con más detalle a continuación).

Algunos atributos de un estado, "sistema de comunicación 1-4 estado 1," se muestran en la FIG. 6A. Estos atributos son el número medio de equipos de usuario que acampan en la celda k, el número medio de equipos de usuario activos que transmiten/reciben con la celda k, la utilización media del ancho de banda para la celda k y el rendimiento medio para la celda k.

A modo de ejemplo, se muestran cuatro estados siguientes que pueden alcanzarse desde el estado 1. Se puede alcanzar el estado A(i1) y el estado B(i2) eligiendo la decisión i. Se puede alcanzar los estados X(w1) e Y(w2) eligiendo la decisión w. El aspecto Markov se demuestra por el hecho de que tomar una decisión no siempre conduce a un mismo estado sucesor, sino que hay una Probabilidad i1 de llegar a A(i1) y una Probabilidad i2 de llegar a B(i2). Del mismo modo, tras la elección de tomar una decisión w, existe una Probabilidad w1 de llegar a X(w1) y una Probabilidad w2 de llegar a Y(w2). Del mismo modo, hay cuatro recompensas posibles, Recompensa i1, Recompensa i2, Recompensa w1 y Recompensa w2.

Después de tomar la decisión i, las políticas 1-1 y 1-2 se actualizan de acuerdo con el estado al que se ha llegado y cualquiera que haya sido la recompensa. Del mismo modo, tras tomar la decisión w, las políticas 1-1 y 1-2 se actualizan en función del estado al que se ha llegado y de cuál haya sido la recompensa.

En el ejemplo específico de aplicación al sistema de comunicación 1-4, el estado se asocia con el funcionamiento de las celdas (rendimiento).

La Figura 6B ilustra el pseudocódigo de un ejemplo de realización. La línea 3 corresponde a la aplicación de la acción unida 1-3 al sistema de comunicación 1-4.

Línea 4 de la FIG. 6B corresponde a la obtención del conjunto de parámetros 1-6 del sistema.

La línea 5 se refiere al cálculo de las recompensas. Las recompensas se discuten más adelante, por ejemplo, véase la discusión de la Ec. (10) a continuación para la recompensa de nivel inferior y la Ec. (1) a continuación como ejemplo de recompensa de nivel superior.

La línea 6 se refiere a las funciones de ventaja. Una función de ventaja es una medida de hasta qué punto una determinada acción es una buena o mala decisión dado un determinado estado. La función de ventaja da una medida de la ventaja de seleccionar una determinada acción a partir de un determinado estado.

Las líneas 7-9 se refieren a la Ec. (11). Ec. (12) y la Ec. (13) a continuación.

En la línea 10, se elige la acción con la mayor recompensa de acuerdo con lo indicado por la política de nivel superior.

En la línea 11, un generador de subobjetivos (basado en un LSTM) genera una nueva subobjetivo. Un subobjetivo también se denomina en la presente memoria descriptiva objetivo.

En la línea 12, se genera una acción de nivel inferior, basada en parte en el objetivo.

La línea 13 representa la formación de la acción unida 1-3 como una concatenación de acciones de cada política.

La línea 14 indica el retorno a la línea 1 (esto es similar al bucle de retorno 2-26 de la FIG. 2A).

Las Figuras 7A y 7B ilustran una amplitud de aplicación el enfoque de política jerárquica indicado por las realizaciones.

Por ejemplo, la política inferior 1-2 está asociada con el movimiento hacia un objetivo 2-3 asociado con la reselección de celda, por ejemplo, la reselección 3-3 de la FIG. 3 evento 3-25

Sin embargo, las realizaciones no limitan la política inferior 1-2 a la asociación con la reselección de celdas.

Por ejemplo, la Figura 7A ilustra el flujo lógico 7-8. El flujo lógico 7-8 es una lógica de una primera política y una segunda política. En 7-10, el flujo lógico determina conjuntamente una primera política 7-1 y una segunda política 7-2. Basándose en la primera política 7-1 y la segunda política 7-2, se proporciona una acción unida 7-3 a un sistema de comunicación 7-4. En 7-20, la acción unida 1-3 se aplica al sistema de comunicación 1-4.

Además de la FIG. 7A, la FIG. 7B ilustra un sistema 7-19 y la evolución de la política. El sistema 7-19 es un sistema influenciado por la primera política 7-1 y la segunda política 7-2. El sistema incluye un servidor de determinación de políticas 7-5 y el sistema de comunicación 7-4. Una serie de hechos temporales se indexa con números entre paréntesis. Se puede decir que todos los sucesos (1), (2), (3) y (4) ocurren en una época, y que eventos similares con datos diferentes ocurren en la siguiente época.

En el evento (1), se proporciona un conjunto de parámetros del sistema 7-6 desde el sistema de comunicación 7-4 al servidor de determinación de políticas 7-5. En el evento (2), se ejecuta la lógica de 7-10 y se determinan o evolucionan (revisan o entrenan) la política inferior 7-1 y la política superior 7-2. En el evento (3), la acción unida 7-3 se proporciona al sistema de comunicación 7-4. En el evento (4), se ejecuta la lógica 7-20 y se aplica la acción unida 7-3 al sistema de comunicación 7-4. La evolución continúa entonces en el evento (5) con un conjunto actualizado de parámetros del sistema 7-6 proporcionado al servidor de determinación de políticas 7-5. Como en la época anterior, el servidor de determinación de políticas 7-5 ejecutará la lógica 7-10, y así sucesivamente.

Así, los parámetros 7-6, las recompensas, las acciones y el sistema de comunicación 7-4 no se limitan a los EUs de acampada 1-91. Por ejemplo, el conjunto de parámetros del sistema 7-6 de la lógica 7-8 y el sistema 7-19, en algunas realizaciones son datos de sensores. Las recompensas de determinar conjuntamente 7-10 aplicar el Procedimiento 1 utilizan métricas para el rendimiento del sistema. Así pues, el Procedimiento 1 no se limita a que la política μ^L se dirija a la reselección de celda de los EU de acampada. Las acciones que forman la acción unida 7-3 son, en algunas realizaciones, parámetros controlables que pueden ajustarse manual o automáticamente. El sistema de comunicación 7-4, es en algunas realizaciones, el sistema de comunicación 1-4. Una interfaz de comunicación, por ejemplo entre el servidor de determinación de políticas 1-5 de la FIG. 1C y el servidor de red 1-9 (véanse también los mensajes 3-10 y 3-12 de FIG. 3) están adaptados para transferir el conjunto de parámetros del sistema 7-6 al servidor de determinación de políticas 1-5 y la acción unida 7-3 al servidor de red 1-9 para que se den más órdenes al agente de control.

En una realización adicional, las políticas jerárquicas de la FIG. 1A se utilizan para ayudar a los servidores a ahorrar energía. En este caso, los estados del PDM de las FIG. 6A incluyen el uso del BFR (Bloque Físico de Recursos) de la celda además de los valores de estado descritos anteriormente (promedio de EUs acampados por celda, promedio de EUs activos por celda, ancho de banda y/o rendimiento). La acción inferior en esta realización es el control de un interruptor de encendido/apagado de radio a nivel de celda (para cada celda de las celdas 1-93 de la FIG. 1C, apagar un dispositivo de radio de una estación base concreta) y la acción superior incluye el ajuste de α para el equilibrio de

carga de EU activos (EUs 1-92 de FIG. 1C). La recompensa en el nivel superior es la métrica de energía del sistema, además de la métrica de rendimiento IP original; ambas se observan directamente en el entorno. La recompensa de nivel inferior evalúa si se ha alcanzado el objetivo, y se calcula con una función de recompensa.

La Figura 8 ilustra un ejemplo de sistema 8-4 en el que interactúan el aparato de aprendizaje por refuerzo 8-5 y el entorno 8-1, de acuerdo con algunas realizaciones de la divulgación.

Con referencia a la FIG. 8, el aparato de aprendizaje por refuerzo 8-5 puede transmitir la acción unida 8-10 como una acción al entorno 8-1, de forma que pueda realizarse un procedimiento de traspaso entre una EB y un EU en el entorno 8-1. El entorno 8-1 puede transmitir, al aparato de aprendizaje por refuerzo 8-5, un estado 8-3 del entorno 8-1. El estado 8-3 puede incluir una utilización de ancho de banda para una estación base (EB) del entorno 8-1, un número de EUs activos para la estación base, y un rendimiento medio para la EB. La acción unida puede incluir el umbral de traspaso/ECEA y el umbral de reselección de celda/ECEI. Por ejemplo, la acción unida 8-10 es la acción unida 1-3 mostrada en la Fig. 1A.

En una realización de la divulgación, el aparato de aprendizaje por refuerzo 8-5 puede controlar los datos de entrada basándose en una regla de operación predefinida (por ejemplo, un flujo lógico 2-8) o un modelo de IA, derivándose los datos de entrada de una señal de control recibida y una señal de datos recibida.

La regla de funcionamiento predefinida o el modelo de IA pueden crearse por medio de entrenamiento. En este caso, cuando la regla de funcionamiento predefinida o el modelo de IA se crean por medio de entrenamiento, puede significar que un modelo básico de IA se entrena utilizando múltiples datos de entrenamiento basados en un algoritmo de aprendizaje para ejecutar las características deseadas (o propósito), creando así la regla de funcionamiento predefinida o el modelo de IA. Dicho entrenamiento puede ser realizado por una EB o entidad de red en la que se implemente la IA de acuerdo con la divulgación o por un servidor y/o un sistema independiente.

En una realización de la divulgación, un primer EU activo 8-20 puede acceder a una red a través de una primera EB 8-30 que es una EB de servicio y puede ser provista con un servicio, y luego puede transmitir un mensaje de informe de medición a la primera EB 8-30. El mensaje de informe de medición puede incluir información para que el primer EU 8-20 activo realice un procedimiento de traspaso. Por ejemplo, el mensaje de informe de medición puede incluir información que indique que una señal recibida de la primera EB 8-30 es inferior o igual a la potencia preestablecida. La potencia preestablecida puede actualizarse en función del umbral de traspaso y del umbral de reselección de celda. La primera EB 8-30 puede transmitir un primer caudal de datos actual 8-2 y un caudal de datos, un número de EUs activos, y una utilización de ancho de banda como primer estado 8-3 al aparato de aprendizaje por refuerzo 8-5. El aparato de aprendizaje por refuerzo 8-5 puede obtener la acción unida 8-10, en respuesta a una entrada del primer estado 8-3 y/o el primer caudal de datos 8-2.

En una realización de la divulgación, cada una de una pluralidad de EBs puede recibir la acción unida 8-10 y puede transmitir toda o parte de la acción unida a un EU. Basándose en la acción unida 8-10, el primer EU activo 8-20 puede realizar un procedimiento de traspaso (por ejemplo, por la condición Ec. (4)) con la segunda EB 8-31.

En una realización de la divulgación, similar al primer EU activo 8-20, basado en la acción unida 8-10, un primer EU en inactividad 8-21 puede realizar un procedimiento de reselección de celda (por ejemplo, por medio de la condición Ec. (5)) con la primera EB 8-30.

Tras un periodo de tiempo preestablecido, el entorno 8-1 puede transmitir, al aparato de aprendizaje por refuerzo 8-5, un segundo estado y un segundo caudal de datos en respuesta al traspaso de al menos un EU. El aparato de aprendizaje por refuerzo 8-5 puede identificar una recompensa como + cuando el rendimiento de los segundos datos aumenta en comparación con el rendimiento de los primeros, y puede identificar una recompensa como - cuando el rendimiento de los segundos datos disminuye en comparación con el rendimiento de los primeros. 1A, 1B, 1C, 1D, 2A, 2B, 3, 4, 6A y 6B).

Supongamos que hay N_U EUs (EUs de acampada1-91 y EUs activos 1-92 tomados en conjunto). Un EU inactivo en el momento actual puede pasar a estar activo en el futuro, y viceversa.

Un objetivo de las realizaciones es equilibrar la distribución de todos los EUs a través de diferentes celdas, para maximizar las siguientes métricas.

$$G_{aver} = \frac{1}{N_U} \sum_k \sum_i \frac{A_{i,k}}{T} \quad \text{Ec. (1)}$$

en la que T es el periodo de tiempo de interés, y $A_{i,k}$ es el tamaño total de los paquetes recibidos por $u_{i,k}$ dentro de T.

La segunda métrica es el rendimiento mínimo G_{min}, es decir,

$$G_{min} = min_{i,k}\left(\frac{A_{i,k}}{T}\right) \quad \text{Ec. (2)}$$

La tercera métrica es la desviación estándar (DE) del rendimiento, G_{sd}.

$$G_{sd} = \sqrt{\frac{1}{N_U}\sum_k \sum_i \left(\frac{A_{i,k}}{T} - G_{aver}\right)^2} \quad \text{Ec. (3)}$$

La minimización de G_{sd}. reduce la brecha entre los rendimientos de los diferentes EUs y por lo tanto proporciona servicios más justos a través de los EUs colectivamente.

Algunos procedimientos EC se basan en el traspaso (T) de EUs activos. En cuanto a cómo se producen la reselección y el traspaso, consideremos un ejemplo de un mecanismo de T basado en la potencia recibida de la señal de referencia (RSRP1). Un equipo de usuario observa la potencia de la celda; en concreto, un equipo de usuario compara un valor RSRP de su celda servidora con los valores de sus celdas vecinas. Si se cumple la siguiente condición, el equipo de usuario activo se transferirá a una celda vecina, es decir,

$$RSRP_j > RSRP_i + \alpha_{i,j} + H \quad \text{Expresión (4)}$$

en la que $RSRP_i$ representa el RSRP del EU de la celda de servicio i, $RSRP_j$ denota el RSRP del EU de una celda vecina j, $\alpha_{i,j}$ es el umbral T de la celda i a la celda j, y H es la histéresis T. Este umbral T $\alpha i_{,j}$ es una variable direccional de pares (por ejemplo, $\alpha i_{,j} \neq \alpha i_{j}$). Al cambiar $\{\alpha_{i,j}\}$, las realizaciones ajustan los límites de T entre celdas y, por tanto, equilibran el número de EU activos entre celdas.

Otra categoría de procedimientos EC depende de la Reselección de celda (RC) de los EUs inactivos. Cuando un EU se enciende, primero entra en modo inactivo y "acampa" en una celda. Un EU inactivo está listo para iniciar un posible servicio dedicado o para recibir un servicio de difusión. Una vez activo, el equipo de usuario suele permanecer en la misma celda en la que ha acampado durante el modo inactivo.

Un EU inactivo puede acampar en otra celda por medio del procedimiento de reselección de celda (RC), para permanecer conectado cuando se desplaza. Este procedimiento RC se activará si se cumple la siguiente condición para un EU inactivo:

$$RSRP_i < \beta_{i,j}, \quad \text{y} \quad RSRP_j > \gamma_{i,j} \quad \text{Expresión (5)}$$

en la que $\beta i_{,j}$ y $\gamma_{i,j}$ son umbrales RSRP por pares y direccionales para activar la RC desde una celda de acampada i a una celda vecina j. El mecanismo de RC representado por la condición mostrada en la Ec. (5) es una generalizada. Ajustando$\{\beta_{i,j}\}$ y $\{\gamma_{i,j}\}$, las realizaciones consiguen una distribución equilibrada de los EU inactivos entre las celdas. Esto ayuda a reducir la congestión cuando los EU inactivos se activan.

Las realizaciones resuelven un problema EC híbrido, en el que se aplican tanto ECEA como ECEI para conseguir una carga equilibrada y un mejor rendimiento del sistema. Las realizaciones definen un problema EC híbrido como sigue.

$$\max_{\{\alpha_{i,j}\},\{\beta_{i,j}\},\{\gamma_{i,j}\}} G \quad \text{Expresión (6)}$$

de forma que

$$\alpha_{i,j} \in [\alpha_{min}, \alpha_{max}] \quad \textbf{Expresión (7)}$$

$$\beta_{i,j} \in [\beta_{min}, \beta_{max}] \quad \textbf{Expresión (8)}$$

$$\gamma_{i,j} \in [\gamma_{min}, \gamma_{max}] \quad \textbf{Expresión (9)}$$

en las que G es el rendimiento del sistema, α_{min} y α_{max} definen el intervalo controlable de las acciones ECEA, y β_{min}, β_{max}, γ_{min} y γ_{max} definen el intervalo controlable de las acciones ECEI.

Aunque el ECEA y el ECEI individuales proporcionan alguna mejora a una red tomada por separado, no es trivial fusionarlos.

Como ejemplo de la dificultad, considérense dos celdas coubicadas, la celda 1 y la celda 2 (misma estación base), que residen en frecuencias portadoras diferentes. Como ejemplo, un procedimiento ECEA puede establecer $\alpha_{1,2} = \alpha_{2,1}$ = 2dB y H = 1dB, mientras que un procedimiento ECEI puede establecer $\beta_{1,2}$ = - 100dBm y$\gamma_{1,2}$ = - 106dBm. Un sistema puede establecer $\gamma_{i,j} < \beta_{i,j}$ y $\gamma_{j,i} > \beta_{j,i}$ para que los EU puedan equilibrarse proporcionalmente al ancho de banda (u otros recursos) de las celdas.

En este ejemplo, un EU de ejemplo está al principio inactivo, y acampa en la celda 1 con $RSRP_1$ = - 101dB y $RSRP_2$ = - 105dB. Se cumple la condición RC (Expresión 5), es decir, $RSRP_1 < \beta_{1,2}$ y $RSRP_2 > \gamma_{1,2}$ Por lo tanto, en este ejemplo el EU vuelve a seleccionar la celda 2 y acampa en ella. En una ampliación del ejemplo, inmediatamente después, el EU del ejemplo pasa a estar activo (transmitiendo y recibiendo datos de usuario), y utiliza la celda 2 como celda de servicio. En este momento, el equipo de usuario de ejemplo observa que la condición T (Expresión 4) de la celda 2 a la celda 1 se cumple, es decir, $RSRP_1 > RSRP_2 + \alpha_{1,2} + H$. En consecuencia, el equipo de usuario de ejemplo se traslada de nuevo a la celda 1 por medio de T (traspaso). Por lo general, estas rápidas oscilaciones de los equipos de usuario ("ping-pong") entre celdas provocan una degradación del rendimiento y un despilfarro de recursos.

Las realizaciones proporcionan un marco AR (profundo), que es eficaz y eficiente para los problemas EC. La primera etapa para aplicar la AR consiste en formular el problema EC híbrido como un modelo PDM, como se ha comentado anteriormente, por ejemplo FIG. 6A. Un modelo PDM se define como una tupla (S; A; R; P) como sigue.

(i) S: es el espacio de estados. Cada estado es una variable multidimensional continua, que contiene el número medio de EU activos en cada celda, la utilización del ancho de banda de cada celda y el rendimiento medio de cada celda.
(ii) A: es el espacio de acción. Cada acción contiene dos partes. La primera parte a^H corresponde a los parámetros T que controlan las acciones ECEA (es decir, $\alpha_{i,j}$). La segunda parte a^L corresponde a los parámetros RC que controlan las acciones ECEI (es decir, $\beta_{i,j}$ y $\gamma_{i,j}$). En algunas realizaciones, como alternativa, la segunda parte a^L está dirigida a encender o apagar un aspecto de una estación base (por ejemplo, apagar un dispositivo de radio de una estación base particular, como una fuente de alimentación, un amplificador de potencia, un transmisor y/o un receptor).
(iii) R: es la recompensa. Los ejemplos utilizan el rendimiento como recompensa. Otras recompensas son posibles. Cualquier métrica deseable del sistema puede configurarse como recompensa. En algunas realizaciones, la recompensa de nivel superior depende, al menos en parte, del consumo de energía eléctrica de las celdas 1-93.
(iv) P: es la función de probabilidad de transición; véanse las probabilidades de ejemplo indicadas en la FIG. 6A.

Para resolver la falta de cooperación entre ECEA e ECEI (por ejemplo, ping-pongneando), las realizaciones proporcionan una estructura de aprendizaje de políticas jerárquica de dos niveles, que se muestra en la FIG. 2B. El nivel superior controla las acciones ECEA a^H (acciones 2-1) con la política μ^H (política superior 1-1) y el nivel inferior controla las acciones ECEI a^L (acciones 2-2) con la política μ^L (política inferior 1-2). Las acciones a^H y a^L se introducen en las celdas 1-93 para ajustar los parámetros T y RC (descritos en detalle más adelante). La acción 2-1 para la política superior en el momento t puede denotarse como a_t^H o a^H_t. A continuación, se recopila el rendimiento del sistema resultante para proporcionar la recompensa AR actualizada.

En cada etapa de tiempo t, tanto las políticas de nivel superior como las de nivel inferior reciben un estado s_t del entorno. Basándose en este estado, la política de nivel superior $\mu^H(s_t)$ produce una acción de control de nivel superior a^H (acción 2-1). Esta acción de nivel superior se utiliza de dos maneras. 1) se introduce en el sistema para el control del traspaso, y 2) esta acción de nivel superior también se utiliza para producir el objetivo 2-3 para el nivel inferior. El objetivo se denomina $g \in \mathbb{R}^d$ en el que d es la dimensión de *g*. El objetivo $g_t \in \mathbb{R}^d$ (también denominada g_t) se calcula por medio de la función de transición de meta $g_t = f(s_t, a_t^H)$. En otras palabras, en cada etapa de tiempo, esta función genera un objetivo de acuerdo con el estado actual s_t y la acción de nivel superior en el tiempo t: a^H_t. Las realizaciones utilizan una red LSTM para implementar una función de transición de objetivo, es decir, $g_t = LSTM(s_t, a_t^H)$. El uso de LSTM garantiza que el objetivo actual generado sea coherente con los objetivos anteriores.

Basándose en el estado actual s_t y el objetivo g_t, la política de nivel inferior $\mu^L(s_t; g_t)$ produce una acción ECEI a^L_t. Dado que el objetivo engloba las acciones de nivel superior, la política de nivel inferior se ve obligada a alinearse con la política de nivel superior a la hora de alcanzar este objetivo. Las acciones superiores e inferiores combinadas, que pueden representarse como la concatenación aHtEBaLtare se aplican al sistema, de modo que el entorno puede devolver el siguiente estado s_{t+1} y la recompensa r_t.

La recompensa de cada nivel es diferente. En la etapa de tiempo t + 1, la política de nivel superior recibe la recompensa r_t directamente del entorno, es decir, $r^H_t = r_t$, que mide el rendimiento del sistema. La recompensa de nivel inferior r^L_t evalúa si se ha alcanzado el objetivo, y se calcula con una función de recompensa $r^L_t = \eta(g_t; s_{t+1})$.

Un objetivo se define como un estado objetivo que se espera que proporcione una recompensa mayor que el estado actual. Normalmente, un estado objetivo no es alcanzable sólo con las acciones de nivel superior. De este modo, el nivel inferior entra en juego, de modo que la recompensa del nivel superior mejora aún más. En consecuencia, las realizaciones definen la función de recompensa de nivel inferior basándose en la distancia entre el estado actual y el estado objetivo, es decir,

$$r_t^L = \eta(g_t; s_{t+1}) = -\|\emptyset(g_t) - \emptyset(s_{t+1})\|_2 \quad \text{Ec. (10)}$$

en el que Ø(-) es una función de incrustación para asignar el espacio de alta dimensión al espacio de baja dimensión; las realizaciones utilizan la distancia euclidiana de baja dimensión para describir lo cerca que están dos estados de alta dimensión. La política de nivel inferior es recompensada por tomar acciones que producen estados s_{t+1} cercanos al objetivo deseado g_t.

Ambas políticas pueden ser entrenadas usando procedimientos avanzados de AR, incorporando g_t como una entrada adicional en las funciones de valor y política. Algunas realizaciones utilizan un procedimiento de aprendizaje sobre políticas Optimización de Políticas Proximales (OPP) como nuestro procedimiento de entrenamiento de políticas, debido a su robustez. OPP es un ejemplo no limitativo de algoritmo de aprendizaje de política superior e inferior. Otros algoritmos de aprendizaje de políticas encajan en el marco aquí descrito.

Por ejemplo, el algoritmo de aprendizaje para ambas políticas puede ser adoptado de actor crítico suave. Actor crítico suave es un algoritmo de AR profunda basado en el marco de aprendizaje de refuerzo de máxima entropía. El actor pretende maximizar la recompensa esperada al tiempo que maximiza la entropía.

Una técnica adicional que puede aplicarse aborda el error de aproximación de funciones en los procedimientos actor-críticos por medio de una estrategia de regularización que incluye el suavizado de la política objetivo.

Dada la recompensa de nivel inferior de la Ec. (10), la función de valor Q de nivel inferior es minimizar la pérdida:

$$L(\mu^L, D) = \mathbb{E}_{(s_t, a_t, g_t, s_{t+1}, a_{t+1}, g_{t+1}, r_t) \sim D}\left[Q\mu^L(s_t, a_t, g_t) - r(g_t, s_{t+1}) - \gamma Q\mu^L(s_{t+1}, a_{t+1}, g_{t+1})\right]$$

Ec. (11)

Arriba, $Q\mu^L$ indica la función de valor de ventaja del nivel inferior, y D es el amortiguador de repetición. La pérdida de nivel inferior, Ec. 11, obliga a que las acciones aprendidas acerquen el estado al objetivo.

Un amortiguador de repetición almacena un número de transiciones recogidas más recientemente y las utiliza para mejorar el entrenamiento. Un amortiguador de repetición puede implementarse como un amortiguador circular, en el que la transición más antigua del amortiguador se elimina para dejar espacio a una transición que se acaba de recoger. Las transiciones se muestrean desde el amortiguador de repetición para su uso en el entrenamiento.

La función de recompensa de nivel superior se muestra en la Ec. (12). La política aprendida pretende maximizar las recompensas colectivas futuras basándose en el estado actual. En otras palabras, la política de nivel superior genera un objetivo que se espera que mejore el rendimiento del sistema (que es un objetivo principal).

$$L(\mu^H, D) = \mathbb{E}(s_t, a_t, s_{t+1}, a_{t+1}, r_t) \sim D \left[Q\mu^H (s_t, a_t) - r(s_{t+1}) - \gamma Q\mu^H (s_{t+1}, a_{t+1}) \right]$$

Ec. (12)

En el que $Q\mu^H$ es la función de valor de ventaja del nivel superior.

Combinando las dos políticas propuestas, las acciones aprendidas para el ECEA híbrido y el ECEI trabajan en colaboración en términos de mejora del rendimiento del sistema sin conflicto entre ellas. La política ECEA de nivel superior da el paso principal hacia el rendimiento óptimo del sistema eligiendo sus propias acciones y estableciendo el objetivo para la política ECEI de nivel inferior. Al cumplir el objetivo, la política de nivel inferior ayuda a mejorar aún más el rendimiento del sistema con respecto a lo conseguido por el nivel superior.

Algunas técnicas de AR Jerárquico (ARJ) utilizan la política de *nivel superior* para generar el objetivo directamente. A diferencia de ellos, las realizaciones emplean un LSTM como generador de objetivos. Las ventajas de este diseño pueden resumirse del siguiente modo. En primer lugar, la LSTM puede aproximar una función de transición de objetivos no lineal y de alta dimensión, lo que no es posible con la función de transición de objetivos bidimensional existente. En segundo lugar, al generar los objetivos, es importante mantener un cierto nivel de coherencia entre el objetivo actual y los anteriores. Al utilizar el modelo LSTM, el estado de la celda oculta calculado en cada etapa temporal tendrá en cuenta los estados anteriores y generará un objetivo coherente. Véase h 2-90 en la Figura 2B.

La LSTM 2-91 es guiada por el entrenamiento para generar g_t que puede mejorar aún más r_t. Para ello, la pérdida de entrenamiento de este modelo LSTM se establece como el opuesto de la función de valor de ventaja (nótese que la función de valor de ventaja captura el incremento en la recompensa) es decir,

$$L_{generador} = -Q\mu^L(s_t, a_t, g_t) \quad \textbf{Ec. (13)}$$

Minimizando esta pérdida ($L_{generador}$), la LSTM es entrenada para producir un estado objetivo g_t que puede mejorar aún más r_t. Este generador de objetivos basado en LSTM se entrena junto con las políticas de control.

El procedimiento completo de AJP se resume como Procedimiento 1 en la FIG. 6B.

El procedimiento AJP ha sido comparado con otros enfoques de equilibrio de carga y para diferentes métricas.

Los otros enfoques de equilibrio de carga son:

ECEA solo, ECEI solo, ECEA secuencial y luego ECEI, ECEA e ECEI juntos. Las diferentes métricas son la desviación estándar del rendimiento de la Ec. (3), el rendimiento medio de la Ec. (1), el rendimiento mínimo de la Ec. (2), y número de traspasos por celda y hora. El número de EUs por celda ha variado de 10 EUs a 30 EUs. El sistema de comunicación de ejemplo se muestra en la FIG. 1D.

Utilizando AJP, la SD de rendimiento se reduce entre un 20 y un 30% con respecto a los enfoques de comparación. El rendimiento medio aumenta entre un 0,3% y un 24% con respecto a los procedimientos de comparación. El rendimiento mínimo aumenta del 0,17% al 13%. El número de traspasos se reduce del 2% al 9%.

REIVINDICACIONES

1. Un procedimiento para equilibrar la carga en un sistema de comunicación inalámbrica, comprendiendo el procedimiento:

recibir (3-10) una utilización de ancho de banda, un número de equipos de usuario activos, EUs, y un rendimiento medio de al menos una estación base, EB, ;
obtención (3-11) de los primeros datos de estado basados en la utilización del ancho de banda, el número de EUs activos y el rendimiento medio;
obtener (3-12) primeros datos de acción asociados con un equilibrio de carga de EU activo, ECEA, y primeros datos de objetivo para obtener segundos datos de acción, introduciendo los primeros datos de estado y los primeros datos de recompensa en un primer modelo de aprendizaje por refuerzo;
obtener (3-14) segundos datos de acción asociados con un equilibrio de carga de EU inactiva, ECEI, introduciendo los primeros datos de estado, los primeros datos de objetivo y los segundos datos de recompensa en un segundo modelo de aprendizaje por refuerzo;
determinar (3-18) terceros datos de acción asociados tanto con la ECEA como con la ECEI, basándose en los primeros datos de acción y en los segundos datos de acción; y
transmitir (3-19) los datos de la tercera acción a la al menos una estación base, EB.

2. El procedimiento de la reivindicación 1, en el que la obtención de los primeros datos de acción asociados con el ECEA y los primeros datos de objetivo para obtener los segundos datos de acción comprende:

obtener los primeros datos de acción introduciendo los primeros datos de estado y los primeros datos de recompensa en el primer modelo de aprendizaje por refuerzo; y
obtener los primeros datos de objetivo introduciendo los primeros datos de estado y los primeros datos de acción en una red de memoria a largo plazo, LSTM.

3. El procedimiento de la reivindicación 1, en el que los primeros datos de recompensa comprenden al menos uno de rendimiento medio, rendimiento mínimo o desviación estándar del rendimiento.

4. El procedimiento de la reivindicación 1, en el que los primeros datos de acción comprenden al menos un umbral asociado con el ECEA y los segundos datos de acción comprenden al menos un umbral asociado con el ECEI.

5. El procedimiento de la reivindicación 4, en el que el tercer dato de acción es una concatenación del primer dato de acción y del segundo dato de acción.

6. El procedimiento de la reivindicación 1, en el que la segunda recompensa se obtiene basándose en los primeros datos de estado y los primeros datos de objetivo.

7. El procedimiento de la reivindicación 2, que comprende además: actualizar la red LSTM basándose en los primeros datos de estado, los primeros datos de objetivo y los terceros datos de acción.

8. Un aparato para equilibrar la carga en un sistema de comunicación inalámbrica, comprendiendo el aparato:

un transceptor (5-4);
una memoria que almacena una o más instrucciones (5-2, 5-3); y
al menos un procesador (5-9) conectado al transceptor y configurado para ejecutar la una o más instrucciones almacenadas en la memoria para
recibir una utilización de ancho de banda, un número de equipos de usuario activos, EUs y un rendimiento medio de al menos una estación base, EB, ;
obtener los primeros datos de estado basados en la utilización del ancho de banda, el número de EU activos y el rendimiento medio;
obtener unos primeros datos de acción asociados con un equilibrio de carga de EU activo, ECEA, y unos primeros datos de objetivo para obtener unos segundos datos de acción, introduciendo los primeros datos de estado y unos primeros datos de recompensa en un primer modelo de aprendizaje por refuerzo;
obtener unos segundos datos de acción asociados con un equilibrio de carga de EU inactiva, ECEI, introduciendo los primeros datos de estado, los primeros datos de objetivo y los segundos datos de recompensa en un segundo modelo de aprendizaje por refuerzo;
determinar un tercer dato de acción asociado tanto a la ECEA como a la ECEI, basándose en el primer dato de acción y el segundo dato de acción; y transmitir el tercer dato de acción a la al menos una estación base, EB.

9. El aparato de la reivindicación 8, en el que el al menos un procesador está configurado además para ejecutar la una o más instrucciones para:

obtener los primeros datos de acción introduciendo los primeros datos de estado y los primeros datos de recompensa en el primer modelo de aprendizaje por refuerzo; y
obtener los datos del primer objetivo introduciendo los datos del primer estado y los datos de la primera acción en una red de memoria a largo plazo, LSTM.

10. El aparato de la reivindicación 8, en el que los primeros datos de recompensa comprenden al menos uno de rendimiento medio, rendimiento mínimo o desviación estándar del rendimiento.

11. El aparato de la reivindicación 8, en el que los primeros datos de acción comprenden al menos un umbral asociado con el ECEA y los segundos datos de acción comprenden al menos un umbral asociado con el ECEI.

12. El aparato de la reivindicación 8, en el que el tercer dato de acción es una concatenación del primer dato de acción y del segundo dato de acción.

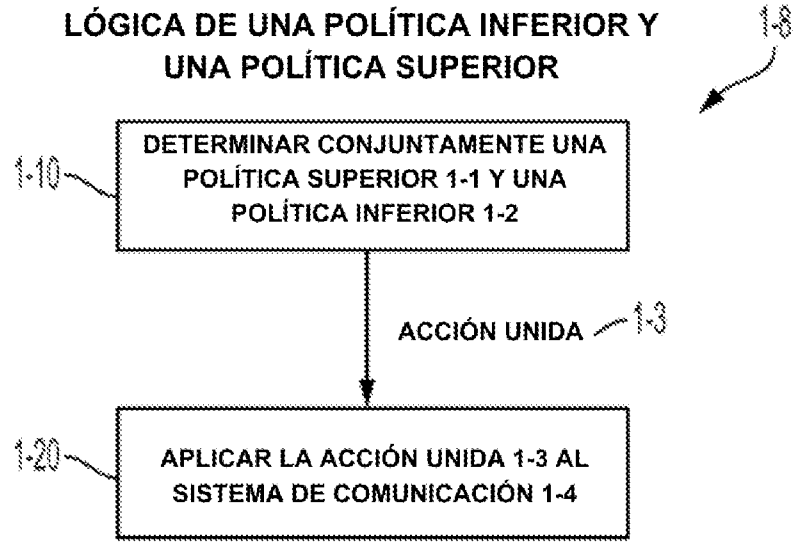
13. El aparato de la reivindicación 8, en el que la segunda recompensa se obtiene basándose en los primeros datos de estado y los primeros datos de objetivo.

14. El aparato de la reivindicación 8, en el que el al menos un procesador está configurado además para ejecutar la una o más instrucciones para: actualizar la red LSTM basándose en los primeros datos de estado, los primeros datos de objetivo y los terceros datos de acción.

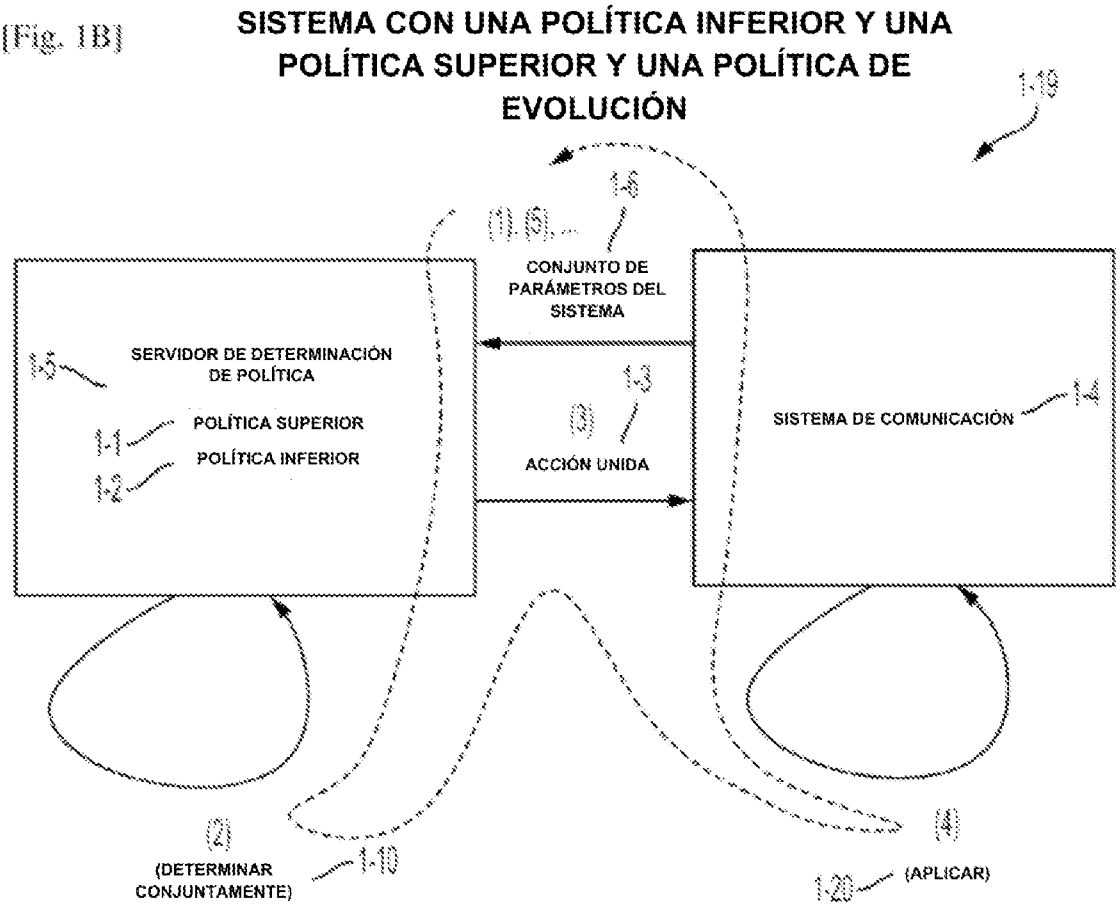
15. Un medio de almacenamiento legible por ordenador que almacena un programa ejecutable por al menos un procesador para llevar a cabo un procedimiento de equilibrado de carga:

recibir (3-10) una utilización de ancho de banda, un número de equipos de usuario activos, EUs, y un rendimiento medio de al menos una estación base, EB;
obtención (3-11) de los primeros datos de estado basados en la utilización del ancho de banda, el número de EUs activos y el rendimiento medio;
obtener (3-12) los primeros datos de acción asociados con el equilibrio de carga activo de EU, ECEA, y los primeros datos de objetivo para obtener los segundos datos de acción, introduciendo los primeros datos de estado y los primeros datos de recompensa en un primer modelo de aprendizaje por refuerzo;
obtener (3-14) segundos datos de acción asociados con un equilibrio de carga de EU inactiva, ECEI, introduciendo los primeros datos de estado, los primeros datos de objetivo y los segundos datos de recompensa en un segundo modelo de aprendizaje por refuerzo;
determinar (3-18) terceros datos de acción asociados tanto con la ECEA como con la ECEI, basándose en los primeros datos de acción y en los segundos datos de acción; y
transmitir (3-19) los datos de la tercera acción a la al menos una estación base, EB.

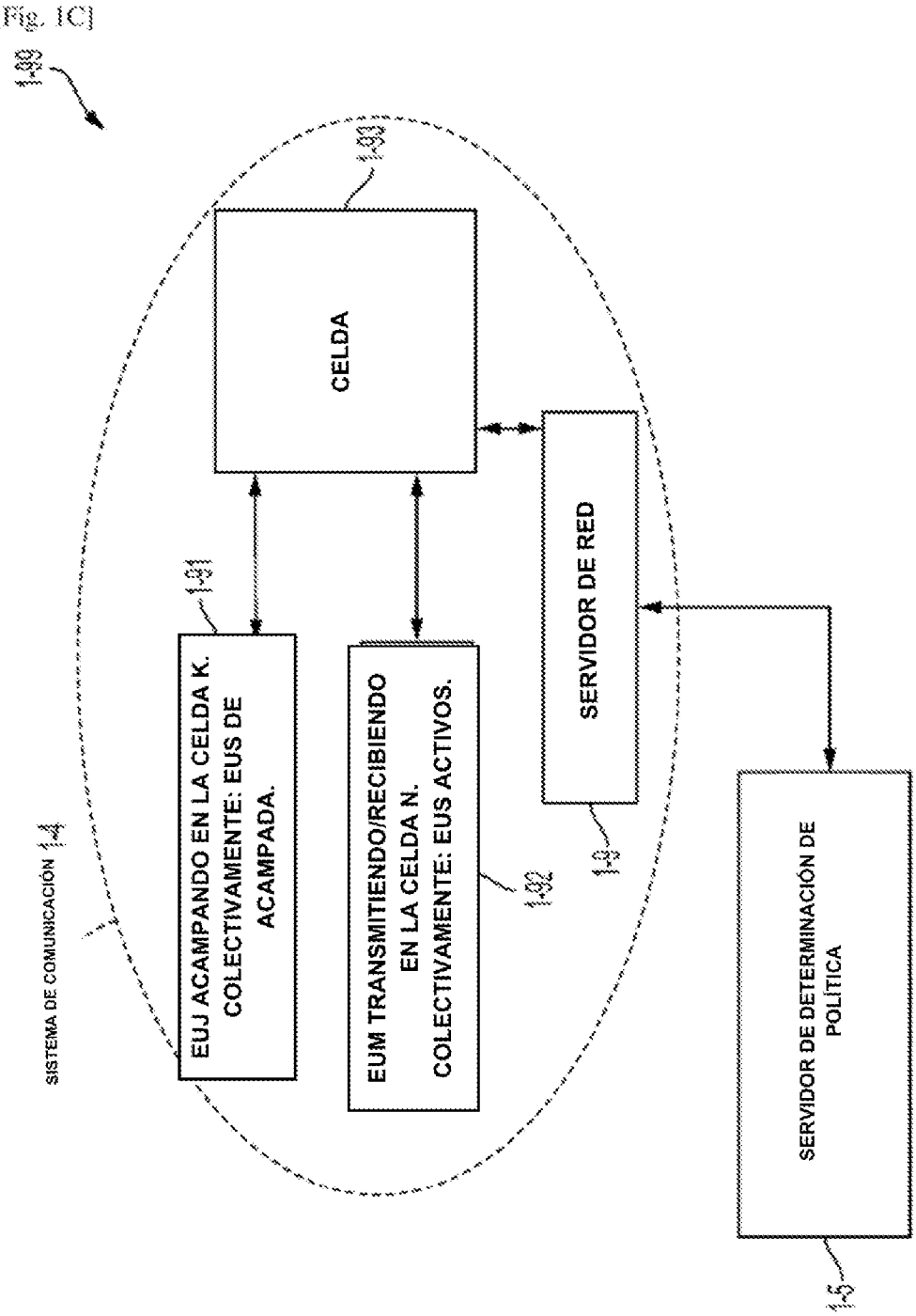
[Fig. 1A]


LÓGICA DE UNA POLÍTICA INFERIOR Y UNA POLÍTICA SUPERIOR
1-8
DETERMINAR CONJUNTAMENTE UNA POLÍTICA SUPERIOR 1-1 Y UNA POLÍTICA INFERIOR 1-2
1-10
ACCIÓN UNIDA
1-3
APLICAR LA ACCIÓN UNIDA 1-3 AL SISTEMA DE COMUNICACIÓN 1-4
1-20

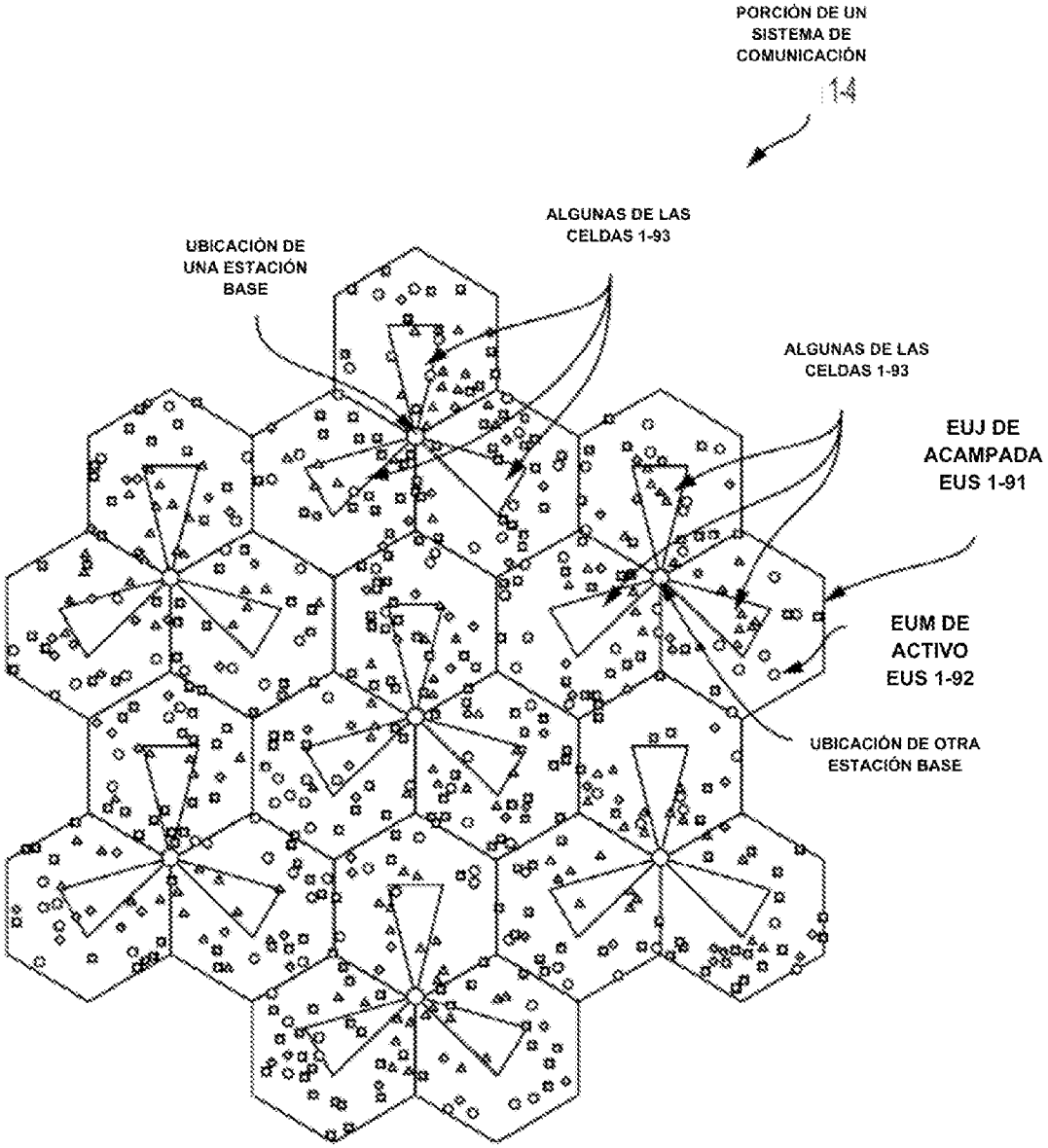
[Fig. 1B]


SISTEMA CON UNA POLÍTICA INFERIOR Y UNA POLÍTICA SUPERIOR Y UNA POLÍTICA DE EVOLUCIÓN
1-19
(1), (5)
1-6
CONJUNTO DE PARÁMETROS DEL SISTEMA
SERVIDOR DE DETERMINACIÓN DE POLÍTICA
1-5
POLÍTICA SUPERIOR
1-1
POLÍTICA INFERIOR
1-2
(3)
1-3
ACCIÓN UNIDA
SISTEMA DE COMUNICACIÓN
1-4
(2)
(DETERMINAR CONJUNTAMENTE)
1-10
(4)
1-20
(APLICAR)

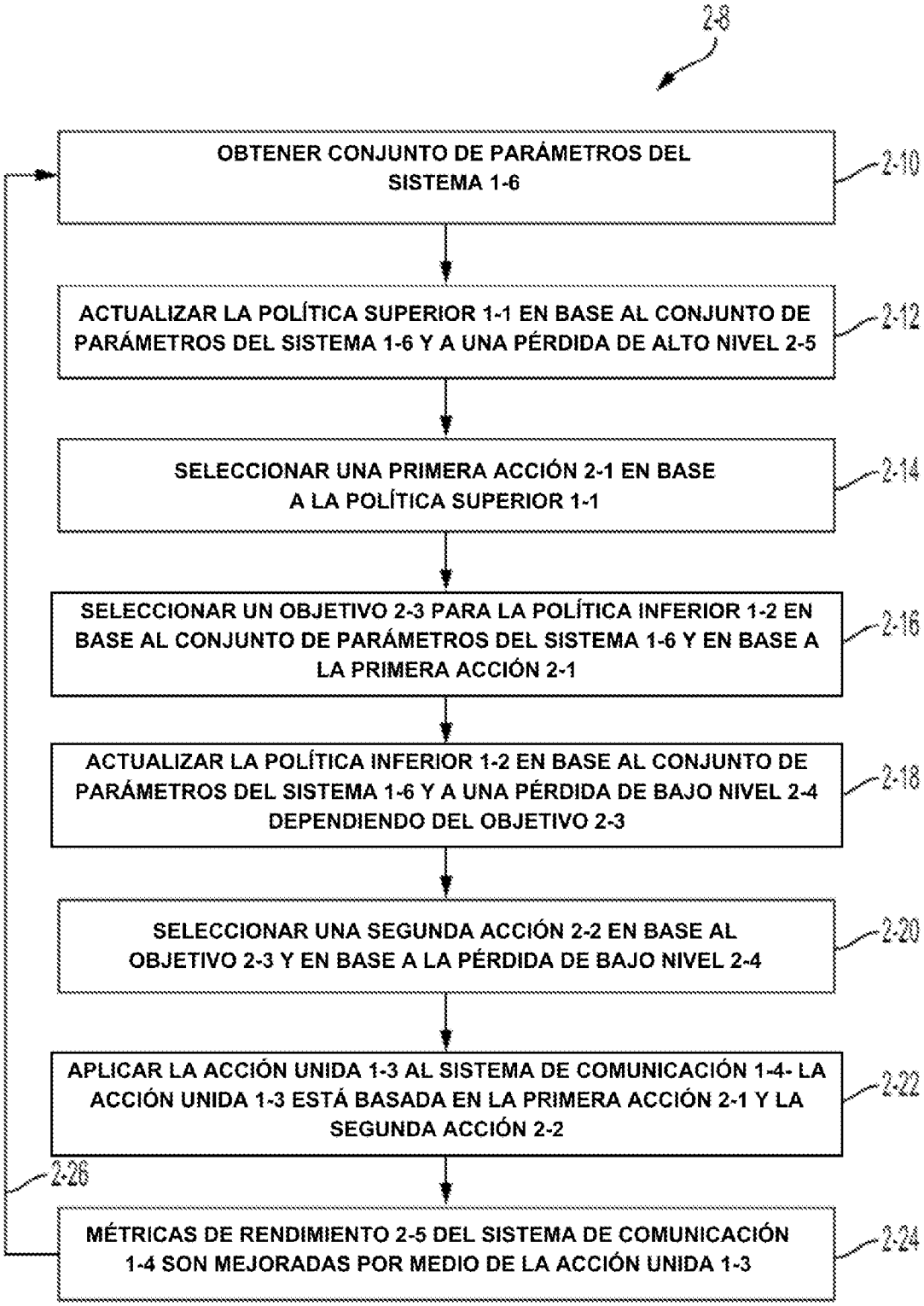
[Fig. 1C]


100
SISTEMA DE COMUNICACIÓN
CELDA
SERVIDOR DE RED
SERVIDOR DE DETERMINACIÓN DE POLÍTICA
EUJ ACAMPANDO EN LA CELDA K. COLECTIVAMENTE: EUS DE ACAMPADA.
EUM TRANSMITIENDO/RECIBIENDO EN LA CELDA N. COLECTIVAMENTE: EUS ACTIVOS.
101
102
103
104
105

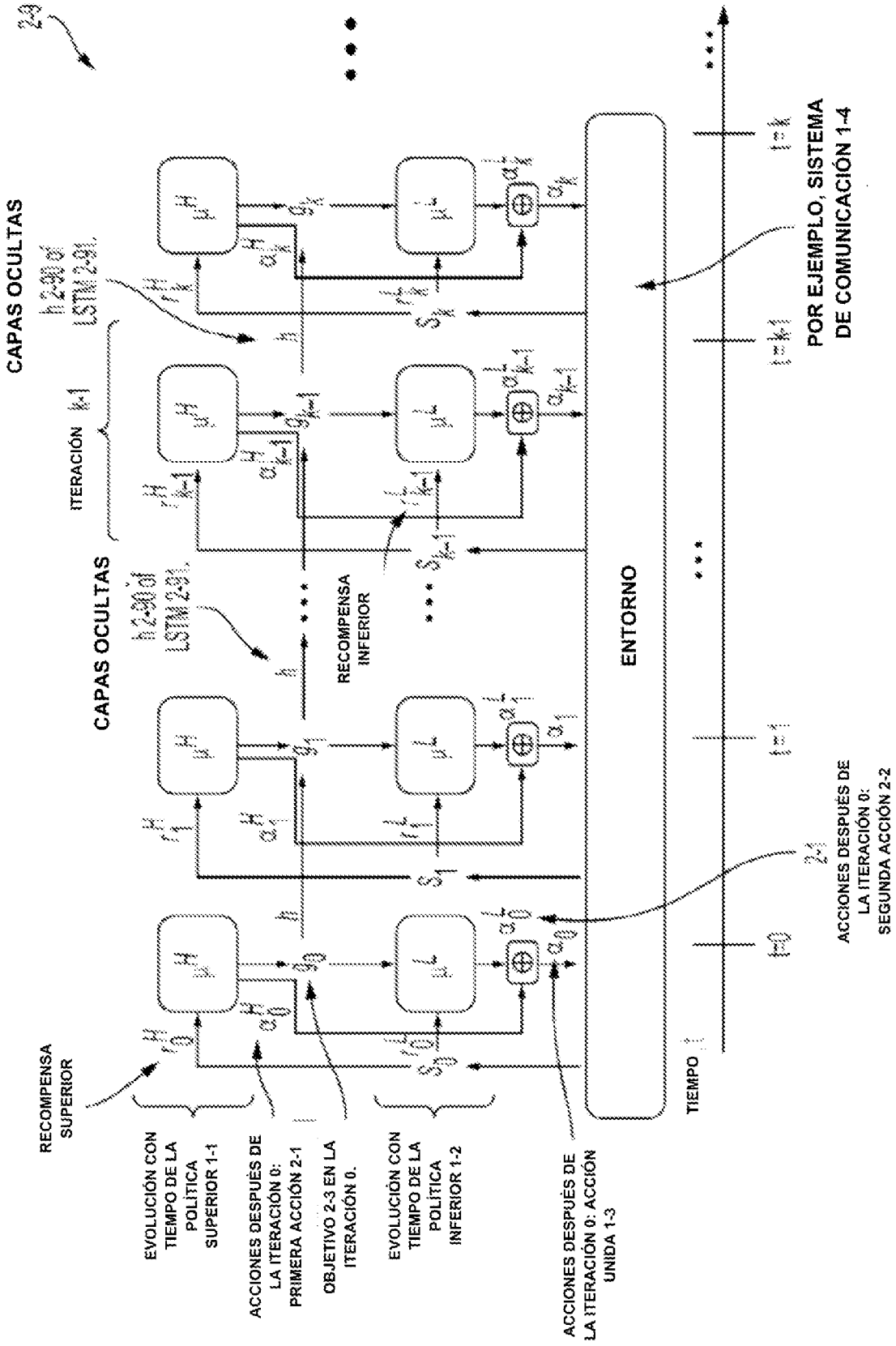
[Fig. 1D]


PORCIÓN DE UN SISTEMA DE COMUNICACIÓN
14
ALGUNAS DE LAS CELDAS 1-93
UBICACIÓN DE UNA ESTACIÓN BASE
ALGUNAS DE LAS CELDAS 1-93
EUJ DE ACAMPADA EUS 1-91
EUM DE ACTIVO EUS 1-92
UBICACIÓN DE OTRA ESTACIÓN BASE

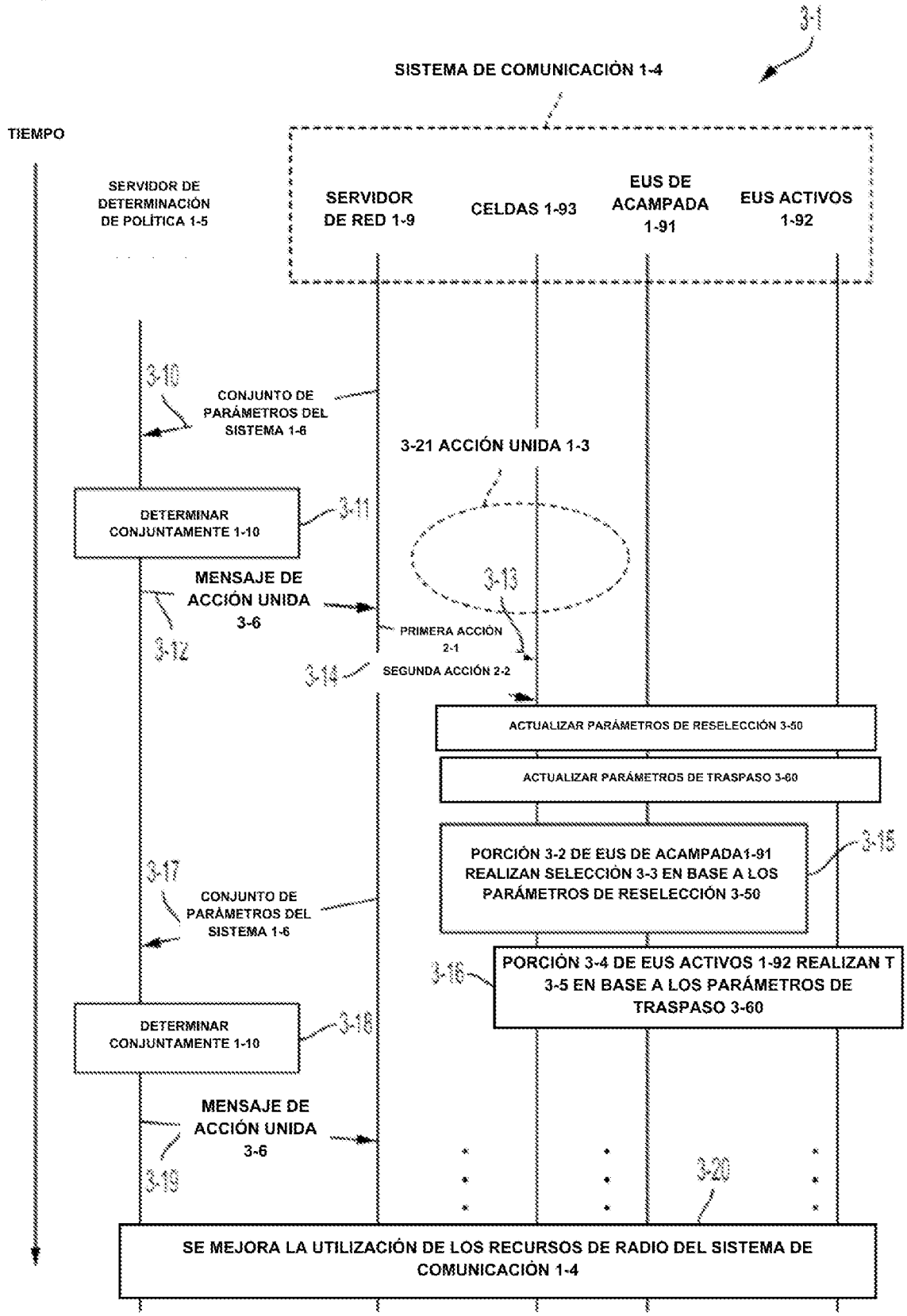
[Fig. 2A]


2-8
OBTENER CONJUNTO DE PARÁMETROS DEL SISTEMA 1-6
2-10
ACTUALIZAR LA POLÍTICA SUPERIOR 1-1 EN BASE AL CONJUNTO DE PARÁMETROS DEL SISTEMA 1-6 Y A UNA PÉRDIDA DE ALTO NIVEL 2-5
2-12
SELECCIONAR UNA PRIMERA ACCIÓN 2-1 EN BASE A LA POLÍTICA SUPERIOR 1-1
2-14
SELECCIONAR UN OBJETIVO 2-3 PARA LA POLÍTICA INFERIOR 1-2 EN BASE AL CONJUNTO DE PARÁMETROS DEL SISTEMA 1-6 Y EN BASE A LA PRIMERA ACCIÓN 2-1
2-16
ACTUALIZAR LA POLÍTICA INFERIOR 1-2 EN BASE AL CONJUNTO DE PARÁMETROS DEL SISTEMA 1-6 Y A UNA PÉRDIDA DE BAJO NIVEL 2-4 DEPENDIENDO DEL OBJETIVO 2-3
2-18
SELECCIONAR UNA SEGUNDA ACCIÓN 2-2 EN BASE AL OBJETIVO 2-3 Y EN BASE A LA PÉRDIDA DE BAJO NIVEL 2-4
2-20
APLICAR LA ACCIÓN UNIDA 1-3 AL SISTEMA DE COMUNICACIÓN 1-4- LA ACCIÓN UNIDA 1-3 ESTÁ BASADA EN LA PRIMERA ACCIÓN 2-1 Y LA SEGUNDA ACCIÓN 2-2
2-22
2-26
MÉTRICAS DE RENDIMIENTO 2-5 DEL SISTEMA DE COMUNICACIÓN 1-4 SON MEJORADAS POR MEDIO DE LA ACCIÓN UNIDA 1-3
2-24

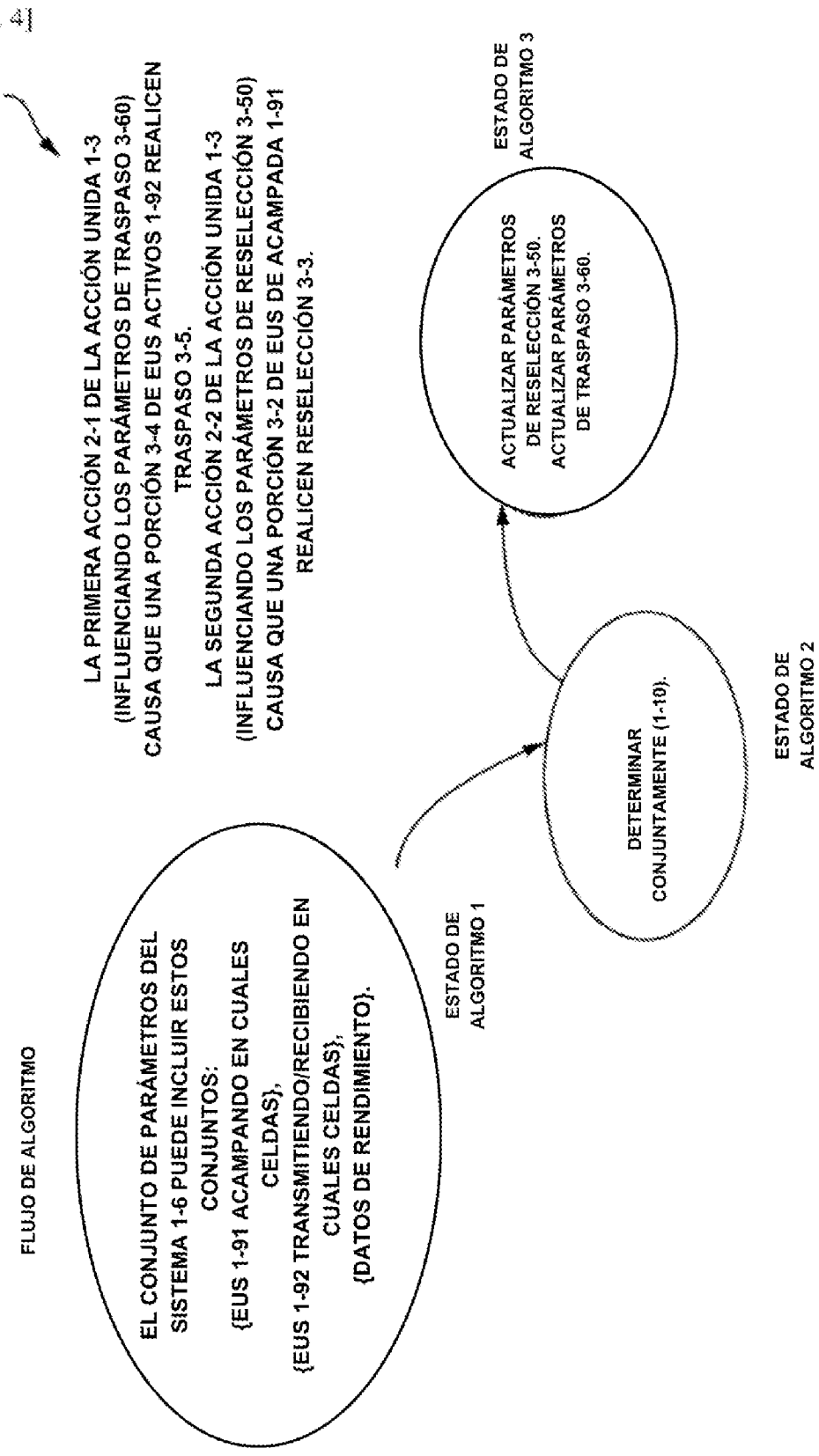
[Fig. 2B]


2-4
CAPAS OCULTAS
h 2-40 de LSTM 2-41.
ITERACIÓN k-1
CAPAS OCULTAS
h 2-30 de LSTM 2-31.
RECOMPENSA INFERIOR
ENTORNO
RECOMPENSA SUPERIOR
EVOLUCIÓN CON TIEMPO DE LA POLÍTICA SUPERIOR 1-1
ACCIONES DESPUÉS DE LA ITERACIÓN 0: PRIMERA ACCIÓN 2-1
OBJETIVO 2-3 EN LA ITERACIÓN 0.
EVOLUCIÓN CON TIEMPO DE LA POLÍTICA INFERIOR 1-2
ACCIONES DESPUÉS DE LA ITERACIÓN 0: ACCIÓN UNIDA 1-3
TIEMPO
t=0
t=1
t=k-1
t=k
2-1
ACCIONES DESPUÉS DE LA ITERACIÓN 0: SEGUNDA ACCIÓN 2-2
POR EJEMPLO, SISTEMA DE COMUNICACIÓN 1-4

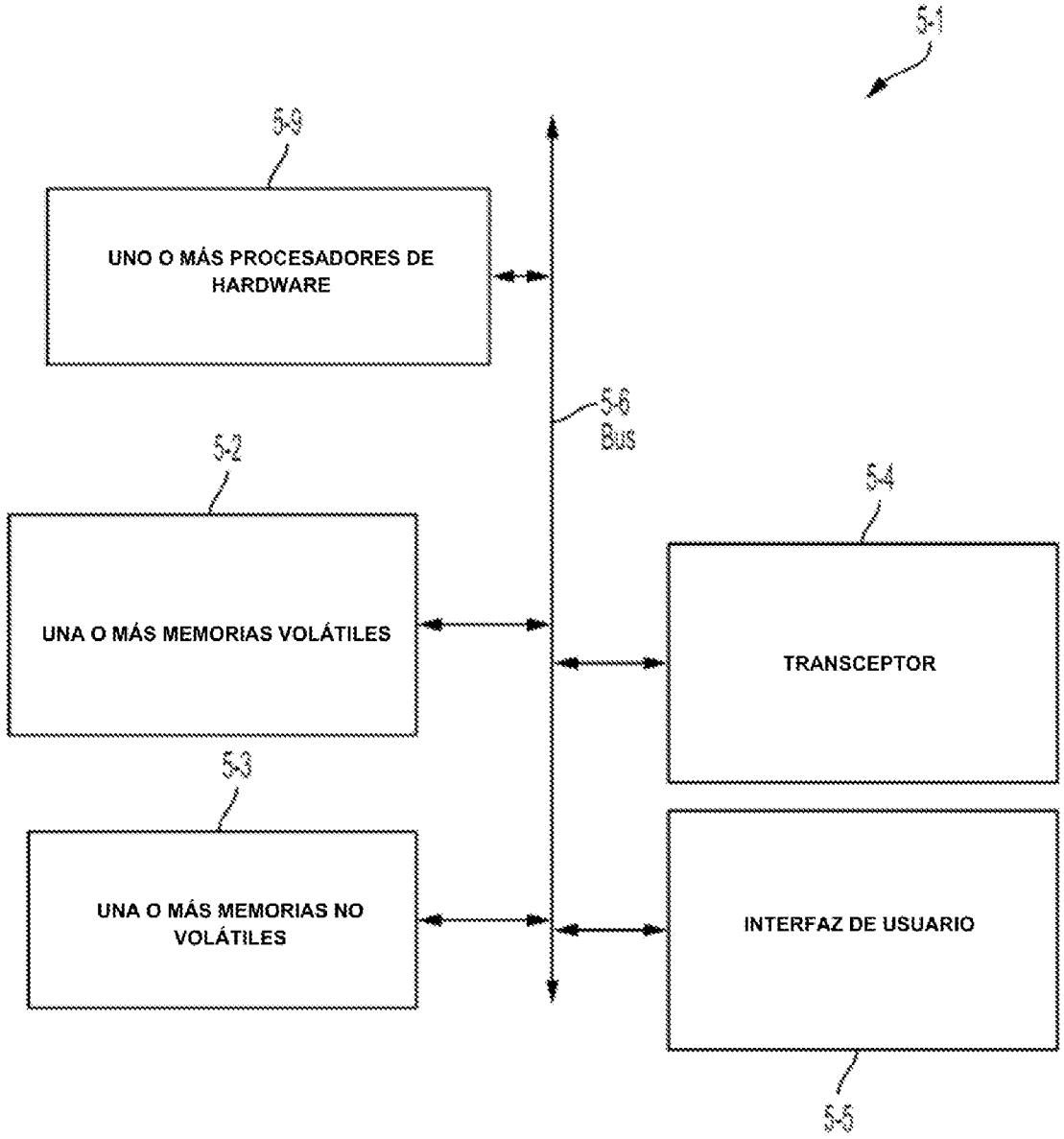
[Fig. 3]


3-1
SISTEMA DE COMUNICACIÓN 1-4
TIEMPO
SERVIDOR DE DETERMINACIÓN DE POLÍTICA 1-5
SERVIDOR DE RED 1-9
CELDAS 1-93
EUS DE ACAMPADA 1-91
EUS ACTIVOS 1-92
3-10
CONJUNTO DE PARÁMETROS DEL SISTEMA 1-6
3-21 ACCIÓN UNIDA 1-3
DETERMINAR CONJUNTAMENTE 1-10
3-11
MENSAJE DE ACCIÓN UNIDA 3-6
3-12
3-13
PRIMERA ACCIÓN 2-1
3-14
SEGUNDA ACCIÓN 2-2
ACTUALIZAR PARÁMETROS DE RESELECCIÓN 3-50
ACTUALIZAR PARÁMETROS DE TRASPASO 3-60
PORCIÓN 3-2 DE EUS DE ACAMPADA1-91 REALIZAN SELECCIÓN 3-3 EN BASE A LOS PARÁMETROS DE RESELECCIÓN 3-50
3-15
3-17
CONJUNTO DE PARÁMETROS DEL SISTEMA 1-6
3-16
PORCIÓN 3-4 DE EUS ACTIVOS 1-92 REALIZAN T 3-5 EN BASE A LOS PARÁMETROS DE TRASPASO 3-60
DETERMINAR CONJUNTAMENTE 1-10
3-18
MENSAJE DE ACCIÓN UNIDA 3-6
3-19
3-20
SE MEJORA LA UTILIZACIÓN DE LOS RECURSOS DE RADIO DEL SISTEMA DE COMUNICACIÓN 1-4

[Fig. 4]


4-1
FLUJO DE ALGORITMO
EL CONJUNTO DE PARÁMETROS DEL SISTEMA 1-6 PUEDE INCLUIR ESTOS CONJUNTOS:
{EUS 1-91 ACAMPANDO EN CUALES CELDAS},
{EUS 1-92 TRANSMITIENDO/RECIBIENDO EN CUALES CELDAS},
{DATOS DE RENDIMIENTO}.
ESTADO DE ALGORITMO 1
DETERMINAR CONJUNTAMENTE (1-10).
ESTADO DE ALGORITMO 2
ACTUALIZAR PARÁMETROS DE RESELECCIÓN 3-50.
ACTUALIZAR PARÁMETROS DE TRASPASO 3-60.
ESTADO DE ALGORITMO 3
LA PRIMERA ACCIÓN 2-1 DE LA ACCIÓN UNIDA 1-3 (INFLUENCIANDO LOS PARÁMETROS DE TRASPASO 3-60) CAUSA QUE UNA PORCIÓN 3-4 DE EUS ACTIVOS 1-92 REALICEN TRASPASO 3-5.
LA SEGUNDA ACCIÓN 2-2 DE LA ACCIÓN UNIDA 1-3 (INFLUENCIANDO LOS PARÁMETROS DE RESELECCIÓN 3-50) CAUSA QUE UNA PORCIÓN 3-2 DE EUS DE ACAMPADA 1-91 REALICEN RESELECCIÓN 3-3.

[Fig. 5]


5-1
5-9
UNO O MÁS PROCESADORES DE HARDWARE
5-6
Bus
5-2
UNA O MÁS MEMORIAS VOLÁTILES
5-4
TRANSCEPTOR
5-3
UNA O MÁS MEMORIAS NO VOLÁTILES
INTERFAZ DE USUARIO
5-5

[Fig. 6A]

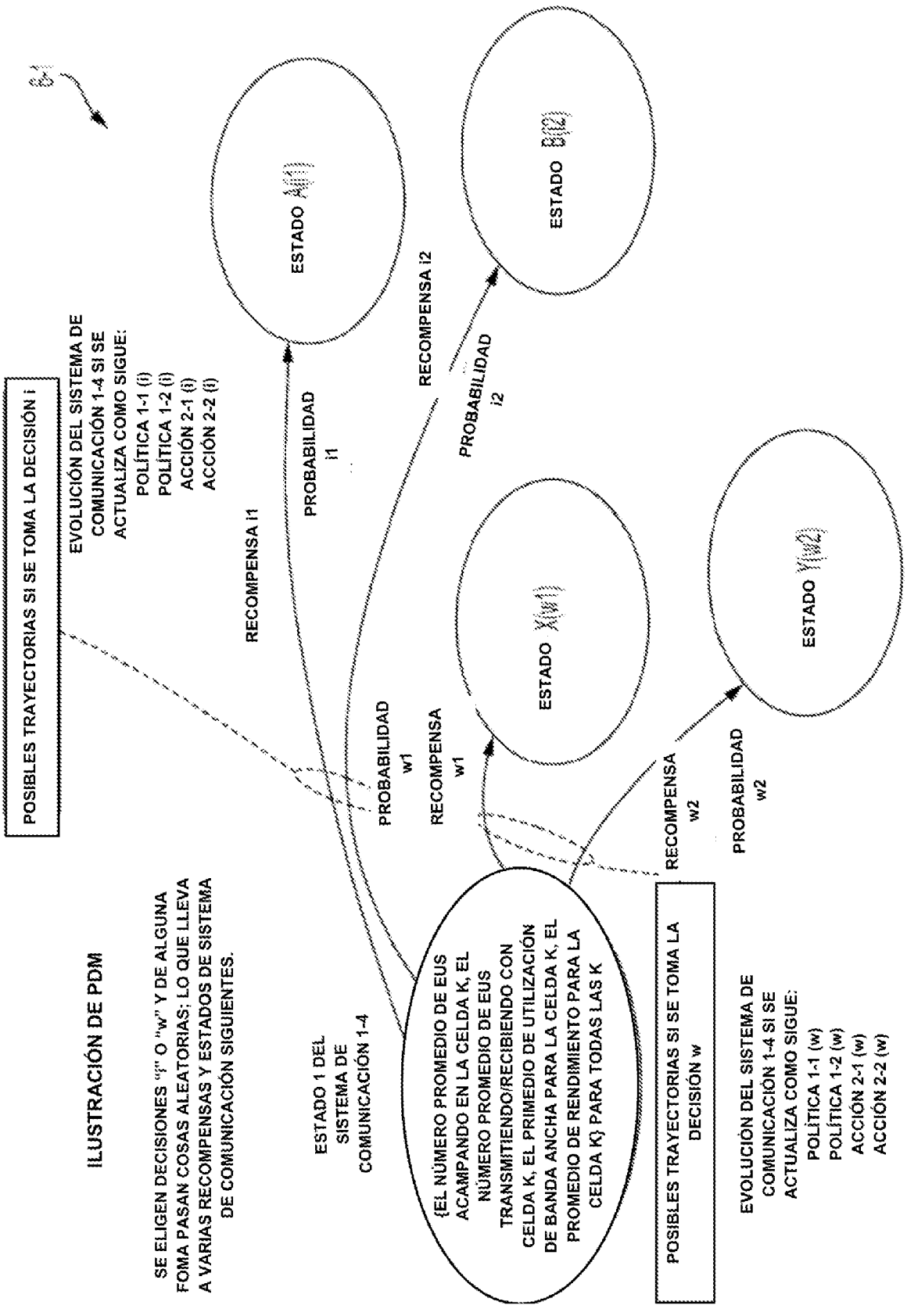


ILUSTRACIÓN DE PDM
SE ELIGEN DECISIONES "i" O "w" Y DE ALGUNA FOMA PASAN COSAS ALEATORIAS; LO QUE LLEVA A VARIAS RECOMPENSAS Y ESTADOS DE SISTEMA DE COMUNICACIÓN SIGUIENTES.
ESTADO 1 DEL SISTEMA DE COMUNICACIÓN 1-4
{EL NÚMERO PROMEDIO DE EUS ACAMPANDO EN LA CELDA K, EL NÚMERO PROMEDIO DE EUS TRANSMITIENDO/RECIBIENDO CON CELDA K, EL PRIMEDIO DE UTILIZACIÓN DE BANDA ANCHA PARA LA CELDA K, EL PROMEDIO DE RENDIMIENTO PARA LA CELDA K} PARA TODAS LAS K
POSIBLES TRAYECTORIAS SI SE TOMA LA DECISIÓN i
EVOLUCIÓN DEL SISTEMA DE COMUNICACIÓN 1-4 SI SE ACTUALIZA COMO SIGUE:
POLÍTICA 1-1 (i)
POLÍTICA 1-2 (i)
ACCIÓN 2-1 (i)
ACCIÓN 2-2 (i)
RECOMPENSA i1
PROBABILIDAD i1
ESTADO A(i1)
RECOMPENSA i2
PROBABILIDAD i2
ESTADO B(i2)
PROBABILIDAD w1
RECOMPENSA w1
ESTADO X(w1)
RECOMPENSA w2
PROBABILIDAD w2
ESTADO Y(w2)
POSIBLES TRAYECTORIAS SI SE TOMA LA DECISIÓN w
EVOLUCIÓN DEL SISTEMA DE COMUNICACIÓN 1-4 SI SE ACTUALIZA COMO SIGUE:
POLÍTICA 1-1 (w)
POLÍTICA 1-2 (w)
ACCIÓN 2-1 (w)
ACCIÓN 2-2 (w)

[Fig. 6B]

PROCEDIMIENTO 1 PROCEDIMIENTO DE APRENDIZAJE DE POLÍTICA JERÁRQUICA

1: INICIAR ALEATORIAMENTE a_0 Y g_0.

2: POR CADA ETAPA DE TIEMPO $t (t = 1, 2, ...)$ HACER

3: APLICAR ACCIÓN a_{t-1} AL ENTORNO

4: RECOLECTAR EL ESTADO ACTUAL S_t Y LA RECOMPENSA r_{t-1} DESDE EL ENTORNO.

5: CALCULAR LAS RECOMPENSAS PARA AMBOS NIVELES, ES DECIR,

$$r_{t-1}^{L} = R(g_{t-1}, S_t) \text{ Y } r_{t-1}^{H} = r_{t-1}.$$

6: COMPUTAR LAS FUNCIONES DE VENTAJA $Q_{\mu L}$ Y $Q_{\mu H}$.

7: ACTUALIZAR LOS PARÁMETROS DE LA POLÍTICA DE BAJO NIVEL μ_L MINIMIZANDO LA PÉRDIDA PRESENTADA EN Eqn. (11).

8: ACTUALIZAR LOS PARÁMETROS DE LA POLÍTICA DE ALTO NIVEL μ_H MINIMIZANDO LA PÉRDIDA PRESENTADA EN Eqn. (12).

9: ACTUALIZAR EL MODELO DE LSTM f_{LSTM} DEL GENERADOR DE SUBOBJETIVOS MINIMIZANDO LA PÉRDIDA PRESENTADA EN Eqn. (13).

10: GENERAR UNA ACCIÓN DE ALTO NIVEL a_t^H CON LA POLÍTICA $\mu_H(S_t)$.

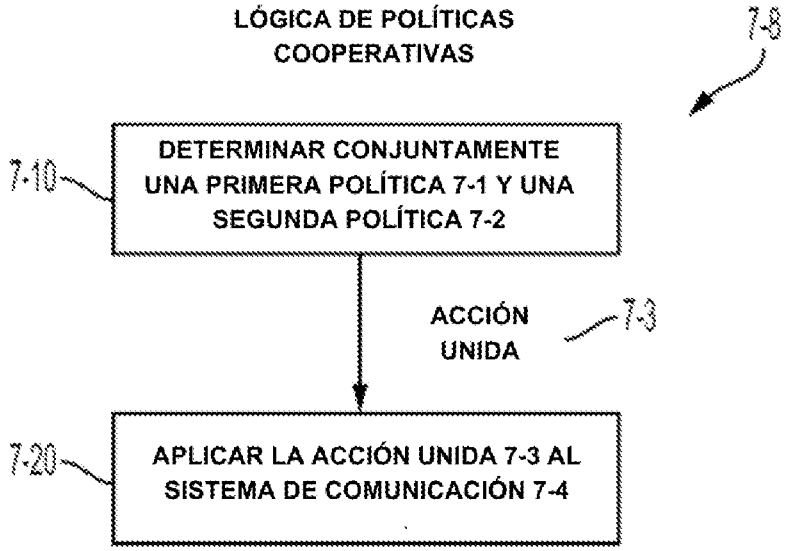
11: USAR EL GENERADOR DE SUBOBJETIVOS PARA PRODUCIR UN NUEVO SUBOBJETIVO $g_t = f_{LSTM}(S_t, a_t^H)$.

12: GENERAR UNA ACCIÓN DE BAJO NIVEL a_t^L CON LA POLÍTICA $\mu_L(S_t, g_t)$.

13: CONCADENAR ACCIONES DE AMBOS NIVELES PARA GENERAR LA ACCIÓN UNIDA $a_t = a_t^H \oplus a_t^L$.

14: TERMINAR

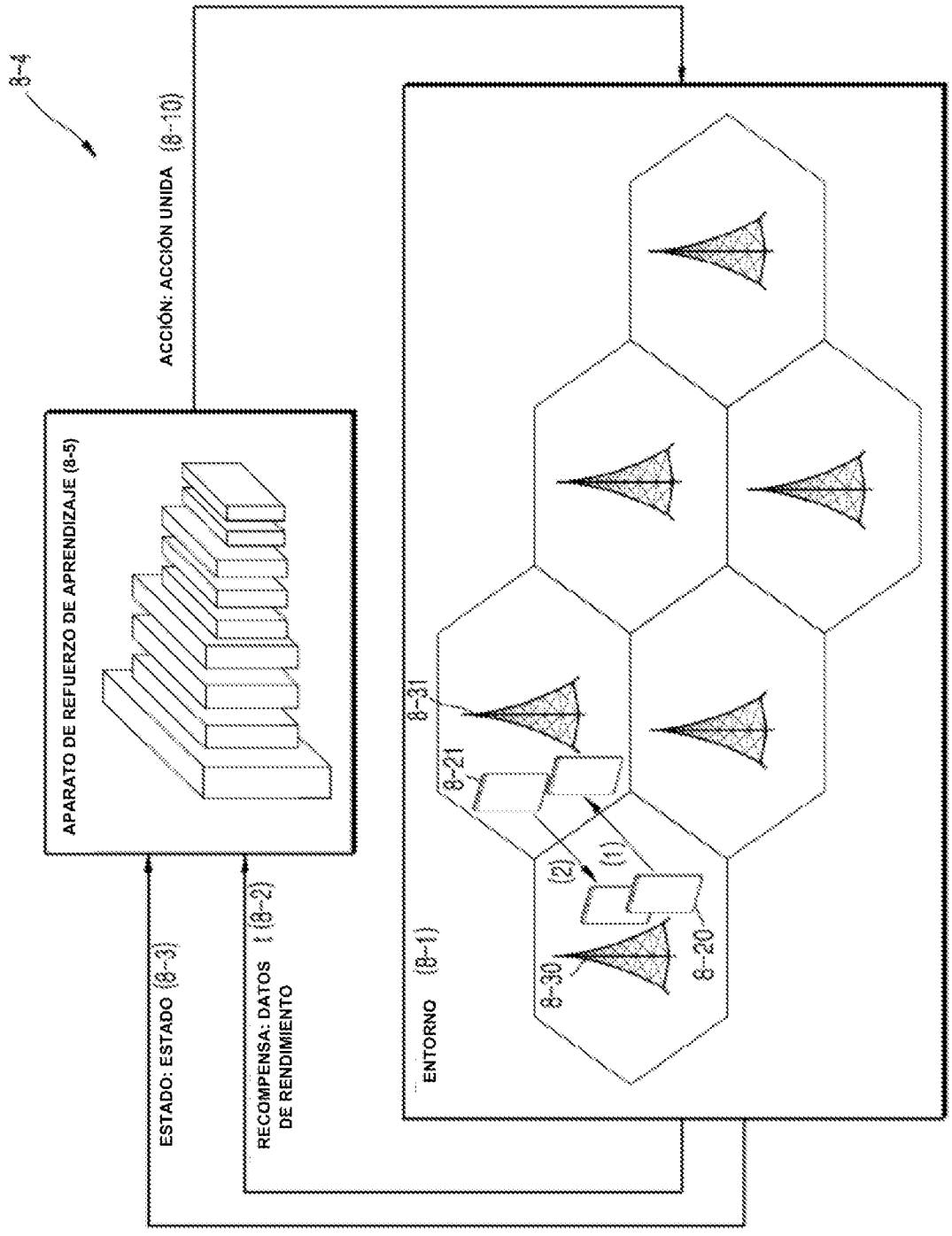
[Fig. 7A]


LÓGICA DE POLÍTICAS COOPERATIVAS
7-8
7-10
DETERMINAR CONJUNTAMENTE UNA PRIMERA POLÍTICA 7-1 Y UNA SEGUNDA POLÍTICA 7-2
ACCIÓN UNIDA
7-3
7-20
APLICAR LA ACCIÓN UNIDA 7-3 AL SISTEMA DE COMUNICACIÓN 7-4

[Fig. 7B]


7-19
POLÍTICAS COOPERATIVAS MEJORANDO EL SISTEMA DE COMUNICACIÓN
7-6
(1), (5), ...
COJUNTO DE PARÁMETROS DEL SISTEMA
7-5
SERVIDOR DE DETERMINACIÓN DE POLÍTICA
7-1
PRIMERA POLÍTICA
SEGUNDA POLÍTICA
7-2
7-3
(3)
ACCIÓN UNIDA
SISTEMA DE COMUNICACIÓN
7-4
(2)
(DETERMINAR CONJUNTAMENTE)
7-10
(4)
(APLICAR)
7-20

[Fig. 8]


8-4
ACCIÓN: ACCIÓN UNIDA (8-10)
APARATO DE REFUERZO DE APRENDIZAJE (8-5)
ESTADO: ESTADO (8-3)
RECOMPENSA: DATOS DE RENDIMIENTO (8-2)
ENTORNO (8-1)
8-31
8-21
(2)
(1)
8-30
8-20