(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
16 November 2017 (16.11.2017)

WIPO | PCT

(10) International Publication Number
WO 2017/196362 A1

(51) International Patent Classification:
H04W 72/04 (2009.01) *H04W 72/08* (2009.01)
H04W 72/06 (2009.01) *H04W 76/06* (2009.01)

(21) International Application Number:
PCT/US2016/032309

(22) International Filing Date:
13 May 2016 (13.05.2016)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **NOKIA TECHNOLOGIES OY** [FI/FI]; Karaportti 3, 02610 Espoo (FI).

(71) Applicant *(for LC only)*: **NOKIA USA INC.** [US/US]; 200 South Mathilda Avenue, Sunnyvale, California 94086 (US).

(72) Inventors: **TALUKDAR, Anup**; 1446 Eagle Ct, Glendale Heights, Illinois 60139 (US). **CUDAK, Mark**; 2800 Deerfield Lane, Rolling Meadows, Illinois 60008 (US).

(74) Agent: **VAN WAES, Nico**; Nokia Technologies OY, c/o Nokia USA Inc., 200 South Mathilda Avenue, Sunnyvale, California 94086 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:
— *of inventorship (Rule 4.17(iv))*

Published:
— *with international search report (Art. 21(3))*

(54) Title: MMWAVE 5G AIR-INTERFACE RADIO RESOURCE CONTROL STATES

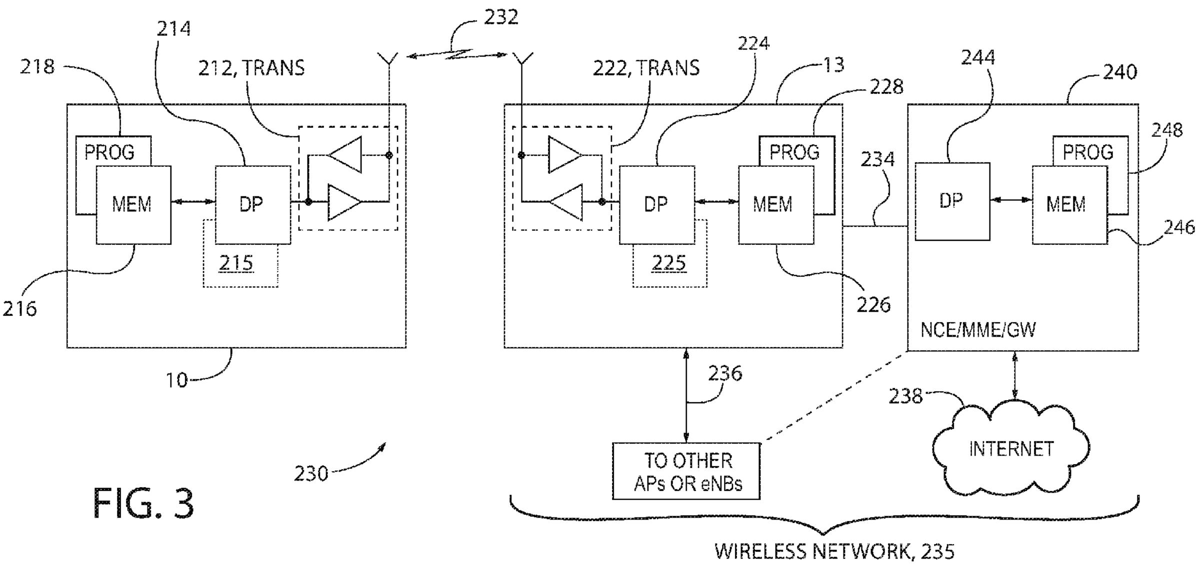

FIG. 3

(57) Abstract: A method is provided including transmitting, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device; receiving, from the cluster set manager, a radio resource allocation for the low-latency operation; and in response to receiving the radio resource allocation for the low-latency operation, triggering the user device to change from a connected state to a low-latency state, wherein the user device maintains connectivity with multiple access points in the cluster set while the user device is in the low-latency state.

WO 2017/196362 A1

mmWave 5G Air-interface Radio Resource Control States

BACKGROUND

Technical Field

[0001] The exemplary and non-limiting embodiments relate generally to wireless communications and, more specifically, to radio resource control states.

Brief Description of Prior Developments

[0002] Typically, existing wireless network include two broad radio resource control states: an idle state, and a connected state. The radio resource control state of a user device determines the functional capabilities and behaviors of the user device and the radio resource configuration for the user device.

SUMMARY

[0003] The following summary is merely intended to be exemplary. The summary is not intended to limit the scope of the claims.

[0004] In accordance with one aspect, an example method comprises transmitting, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device; receiving, from the cluster set manager, a radio resource allocation for the low-latency operation; and in response to receiving the radio resource allocation for the low-latency operation, triggering the user device to change from a connected state to a low-latency state, wherein the user device maintains connectivity with multiple access points in the cluster set while the user device is in the low-latency state.

[0005] In accordance with another aspect, an example embodiment is provided in an apparatus comprising at least one processor; and at least one non-transitory memory including computer program code, the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus to: transmit, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device; receive, from the cluster set manager, a radio resource allocation for the low-latency operation; and in response to receipt of the radio resource allocation for the low-latency operation, trigger the user device to change from a connected state to a low-latency state, wherein the user device maintains connectivity with multiple access points in the cluster set while the user device is in the low-latency state.

[0006] In accordance with another aspect, an example method comprises receiving, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device; allocating radio resources for the low-latency operation; changing a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point; and in response to receiving an indication that all low-latency operations of the user device have completed, changing the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device.

[0007] In accordance with another aspect, an example embodiment is provided in an apparatus comprising at least one processor; and at least one non-transitory memory including computer program code, the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus to: receive, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device; allocate radio resources for the low-latency operation; change a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point; and in response to receipt of an indication that all low-latency operations of the user device have completed, change the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device.

[0008] In accordance with another aspect, an example method comprises receiving an indication, by a cluster set manager of a user device, that a low-latency operation has been initiated; in response to receiving the indication, configuring and allocating radio frame resources in access points for the low-latency operation, wherein the access points are in a cluster set of the user device; notifying the user device of the radio resource allocations, wherein the radio frame resources are at least one of: a downlink control channel for each access point in the cluster set for the user device to monitor, and an uplink polling channel for each access point in the cluster set to be used by the user device for uplink access in the access points; changing a state of the user device maintained by the cluster set manager from a connected to a low-latency state; and on completion of the low-latency operation, changing the state of the user device to the connected state and releasing all resources allocated for low-latency operations if no other low-latency operation is running.

BRIEF DESCRIPTION OF THE DRAWINGS

[0009] The foregoing aspects and other features are explained in the following description, taken in connection with the accompanying drawings, wherein:

[0010] Fig. 1 is a diagram illustrating an example of an overall architecture of a E-UTRAN (evolved UMTS Terrestrial Radio Access) system (an air interface of 3GPP's Long Term Evolution (LTE) upgrade path for mobile networks);

[0011] Fig. 2 is a diagram illustrating a user device in a network of access points, where communication between the user device and one of the access points is blocked by an object;

[0012] Fig. 3 is a diagram illustrating some components of the wireless system shown in Figs. 1 and 2;

[0013] Fig 4 is a diagram illustrating an example mmWave frame structure;

[0014] Fig. 5 is a diagram illustrating an example of a cluster set of a user device and its cluster set manager;

[0015] Fig. 6 is a diagram illustrating an example Virtual Zero Latency state (RRC_VZL) in an mmWave system;

[0016] Fig. 7 is a diagram illustrating example sub-states of the Virtual Zero Latency state (RRC_VZL) in an mmWave system;

[0017] Fig. 8 is a diagram illustrating RRC states for a UD in an mmWave air-interface with the DRX mode;

[0018] Fig. 9 is a diagram illustrating an example method;

[0019] Fig. 10 is a diagram illustrating an example method; and

[0020] Fig. 11 is a diagram illustrating an example method.

DETAILED DESCRIPTION

[0021] The following abbreviations that may be found in the specification and/or the drawing figures are defined as follows:

3GPP	Third Generation Partnership Program
AP	Access Point
CSM	Cluster Set Manager
CSI	Channel State Information
DL	Downlink
DRX	Discontinuous Reception
DTX	Discontinuous Transmission
eNB	enhanced Node B (base station according to LTE terminology)
E-UTRAN	E- Universal Terrestrial Radio Access Network
LOS	Line Of Sight
LTE	Long-Term Evolution
mmWave	Millimeter wave
NACK	Negative acknowledgement
RACH	Random Access CHannel
RRC_CONNECTED	Connected state
RRC_VZL	Virtual Zero Latency state
RRC_IDLE	Idle state
SINR	Signal-to-Interference and Noise Ratio
TDD	Time division duplex
UD	User Device
UL	Uplink

VZL_DATA_XFER	Virtual Zero Latency Data Transfer
VZL_UL_ACCESS	Virtual Zero Latency Uplink Access

[0022] Fig. 1 shows an example of overall architecture of an E-UTRAN system. The E-UTRAN system includes eNBs, providing an E-UTRAN user plane (PDCP/RLC/MAC/PHY) and control plane (RRC) protocol terminations towards the UD (not shown in Fig. 1). The eNBs are interconnected with each other by means of an X2 interface. The eNBs are also connected by means of a S1 interface to an EPC (Enhanced Packet Core), more specifically to a MME (Mobility Management Entity) by means of a S1 MME interface and to a Serving Gateway (S-GW) by means of a S1 interface. The S1 interface supports a many-to-many relationship between MMEs/S-GW and eNBs. One or more of the eNB may form an access point (AP) for a MillimeterWave (mmWave) frequency bands, or the mmWave APs may be otherwise connected to the network shown in Fig. 1. Alternatively, the mmWave APs may form their own network, separate from the network shown in Fig. 1; perhaps connected to the network shown in Fig. 1 by a MME/S-GW for example.

[0023] Referring also to Fig. 2, a user device (UD) 10 is shown. In this example the UD 10 is a smartphone. However, in alternate examples the UD may be, for example, a tablet computer, a PDA, a smart watch, or any other suitable device configured to wireless communications including in a vehicle such as a car for example. The UD 10 is configured to be able to communicate with the APs 13, 14, 15 by mmWave frequency bands as illustrated by possible links 16, 17, 18 shown in Fig. 2.

[0024] Referring also to Fig. 3, in the wireless system 230 a wireless network 235 is adapted for communication over a wireless link 232 with an apparatus, such as a mobile communication device which may be referred to as a UD 10, via a network access node or access point (AP) 13. The network 235 may include a network control element (NCE) 240 that may include MME/S-GW functionality, and which provides connectivity with a network, such as a telephone network and/or a data communications network (e.g., the internet 238). In the example 5G mmWave system shown, the APs will be inter-connected among themselves. A few of the APs, designated as egress APs, will be connected to the NCE/MME/GW 240.

[0025] The UD 10 includes a controller, such as a computer or a data processor (DP) 214, a computer-readable memory medium embodied as a memory (MEM) 216 that stores a program of computer instructions (PROG) 218, and a suitable wireless interface, such as radio frequency (RF) transceiver 212, for bidirectional wireless communications with the AP 13 via one or more antennas.

[0026] The AP 13 also includes a controller, such as a computer or a data processor (DP) 224, a computer-readable memory medium embodied as a memory (MEM) 226 that stores a program of computer instructions (PROG) 228, and a suitable wireless interface, such as RF transceiver 222, for communication with the UD 10 via one or more antennas. The AP 13 is coupled via a data/control path 234 to the NCE 240. The path 234 may be implemented as an interface. The AP 13 may also be coupled to other APs and perhaps eNB(s) via data/control path 236, which may be implemented as an interface.

[0027] The NCE 240 includes a controller, such as a computer or a data processor (DP) 244, a computer-readable memory medium embodied as a memory (MEM) 246 that stores a program of computer instructions (PROG) 248.

[0028] At least one of the PROGs 218, 228 and 248 is assumed to include program instructions that, when executed by the associated DP, enable the device to operate in accordance with exemplary embodiments of this invention, as will be discussed below in greater detail. That is, various exemplary embodiments of this invention may be implemented at least in part by computer software executable by the DP 214 of the UD 10; by the DP 224 of the AP 13; and/or by the DP 244 of the NCE 240, or by hardware, or by a combination of software and hardware (and firmware).

[0029] For the purposes of describing various exemplary embodiments in accordance with this invention the UD 10 and the AP 13 may also include dedicated processors, for example RRC module 215 and a corresponding RRC module 225. RRC module 215 and RRC module 225 may be constructed so as to operate in accordance with various exemplary embodiments in accordance with this invention.

[0030] The computer readable MEMs 216, 226 and 246 may be of any type suitable to the local technical environment and may be implemented using any suitable data storage technology, such as semiconductor based memory devices, flash memory, magnetic memory devices and systems, optical memory devices and systems, fixed memory and removable memory. The DPs 214, 224 and 244 may be of any type suitable to the local technical environment, and may include one or more of general purpose computers, special purpose computers, microprocessors, digital signal processors (DSPs) and processors based on a multicore processor architecture, as non-limiting examples. The wireless interfaces (e.g., RF transceivers 212 and 222) may be of any type suitable to the local technical environment and may be implemented using any suitable communication technology such as individual transmitters, receivers, transceivers or a combination of such components.

[0031] The 5th Generation wireless networks are being designed to deliver peak data rates of the order of ~10Gbps and the target latency requirements have been set to the order of ~1msec in order to serve applications with ultra-low-latency performance requirements. Unlike traditional cellular bands, the availability of large blocks of contiguous spectrum of the order of 10 GHz or more in the mmWave band makes it a promising candidate for 5th generation (5G) cellular technology. The mmWave bands allow for multi-element antenna arrays composed of very small elements, on the order of IC chip scales for example, providing large antenna gain and sufficient power output through over-the-air power combining. This combination of large bandwidths and novel device architectures allows mmWave cellular to provide peak rates on the order of 10 Gbps and ample capacity to meet future demands.

[0032] However the propagation characteristics in the mmWave band are more challenging than traditional cellular. Diffraction at mmWave bands is effectively non-existent and propagation behaves similar to visible light. Transmission through most objects is diminished where foliage and other common obstacles can produce severe shadowing. The severe shadowing loss characteristics in the mmWave band implies that, the radio link between a user device (UD) and its serving AP will be disrupted if the line of sight (LOS) is blocked by obstacles, such as trees, pedestrians or trucks. An example of this is shown in Fig. 2, where the link 16 between the UD 10 and the AP 13 is blocked by the truck 20. Other types of LOS blocking may be caused by user motions such as hand or body rotations. In a campus courtyard or a tourist hotspot LOS blocking may be caused by crowds. In order to deliver reliable connectivity to a user in presence of obstacles, an mmWave access point network is built with enough redundancies of APs such that in the event of a LOS blocking, the network connection of the UD can be rapidly rerouted via another AP. For example, in Fig. 2 the APs 14 and 15 are other APs which can be used by the UD 10 when the line-of-sight (LOS) between the UD 10 and the AP 13 is blocked by the truck 20. The LOS between the UD 10 and the APs 14, 15 is not blocked and, thus, the possible links 17, 18 are potentially available.

[0033] Each UD in an mmWave network is served by a cluster of APs, called a cluster set**.** Members of the cluster set of a UD are selected based on the accessibility of the APs from the UD. In Fig. 2, APs 13, 14, 15 form the cluster set for the UD 10. Among the APs in the cluster set, one particular AP can be selected as the “serving AP” for the UD; through which the network communicates with the UD. In Fig. 2, AP 13 is the serving AP. The other APs 14, 15 in the cluster set are “stand-by APs”. The UD 10 attempts to maintain continuous connectivity with each member AP 13, 14, 15 of its cluster set by maintaining synchronization with the symbol and frame structure, downlink and uplink control channels, and also maintain beam synchronization by selecting best beams for DL and UL communication.

[0034] An example of an air-interface frame structure for an mmWave 5G system is shown in Fig. 4. In this example structure a 20 microsecond superframe 400 is subdivided into 40 subframes (subframes 0-39 for one superframe 400). Each subframe may have, for example, a duration of 500 microseconds. Each subframe is further divided into five (5) slots of 100 microsecond duration. A slot can be a synchronization slot 410, an uplink random access channel (RACH) slot 412, or a data slot 414. Slots 410, 412 and 414 are examples of these type of slots.

[0035] A synchronization slot 410 may be used for system acquisition and also for UD specific beam synchronization. In one example, the synchronization channel (via a synchronization slot) is transmitted every 20msec. The RACH slot can be used by a UD for uplink synchronization, to provide feedback on beam selection, and also to send an uplink resource request. A data slot may contain: downlink control channels, uplink control channels and data channels for downlink and uplink data transmissions. For high efficiency, UD-specific beamforming techniques may be used for all communications between an AP and a UD in a data slot. In addition, use of analog beamforming at the transmitter and receiver requires that all communication channels for different user and access point pairs are time division multiplexed.

[0036] Referring also to Fig. 5, an example is shown corresponding to the UD 10 and cluster set shown in Fig. 2. The cluster set of the UD is configured and managed by the Cluster Set Manager (CSM) 22. There is a logical instance of CSM 22 for each UD which is located in the network. The location of the CSM 22 may be close to the APs in the cluster set to enable low-latency communication with those APs and the UD. In Fig. 5 the cluster set contains three APs, AP_0 is the serving AP of the UD whereas AP_1 and AP_2 are stand-by APs. In this diagram, the stand-by AP 15 has the cluster set manager (CSM) 22 for the user device (UD) 10. The CSM 22 of the UD knows the identity of the current serving AP 13 of the UD.

[0037] According to an example implementation, certain schemes may be used for fast radio link blockage detection and rapid rerouting during downlink and uplink data transmission. In these schemes, for every downlink control message transmission to a UD, an AP allocates an uplink control channel, which may be referred to, for example, as a FastACK channel. Upon successful reception of the downlink control message from the AP, the UD sends an acknowledgement over the allocated FastACK channel. After transmitting a downlink control message to a UD if an AP does not receive any acknowledgement over a corresponding FastACK channel (which may be indicated by detection of a DTX), the AP may determine that the radio link is blocked. On detection of a radio link blockage, the AP notifies the CSM which, in turn, requests a stand-by AP of the UD to send a handover command to the UD. Thus, the UD needs to continuously monitor the downlink control channels of multiple APs in its cluster set to detect a radio link blockage and initiate a fast handover.

[0038] When a UD needs to transmit uplink data after a period of inactivity, it may send a request for uplink resources over an uplink access opportunity, which can be a RACH slot or a UD-specific dedicated uplink polling channel. The uplink access latency for a UD can be reduced by reducing the periodicity of the RACH or the UD-specific dedicated uplink polling channel. Uplink access mechanisms may be used to improve the robustness of uplink against radio link blockages where UL access opportunities are allocated in the APs of the cluster set in a staggered pattern. When data arrives at the empty uplink buffer, the UD may use the earliest next available uplink access opportunity in its cluster set. Since a UD may perform an uplink access via a stand-by AP, it may be necessary for the UD to monitor the downlink control channels of both its serving-AP and the stand-by AP. The use of analog beamforming scheme at the UD's RF receiver constrains the UD to TDM monitoring of the downlink control channels of different APs.

[0039] In typical wireless systems, such as LTE and LTE advanced, there are two broad radio resource control states: Idle (RRC_IDLE) and Connected (RRC_CONNECTED). The RRC state of the UD determines the functional capabilities and behaviors of a UD and the radio resource configuration for the UD. In the idle state, the UD is registered in the network, but it cannot perform any user-plane communication with the network. In order for the UD to perform any data communication, the UD must move into the connected state by establishing an RRC connection. When the UD is in RRC_CONNECTED state, radio resources are configured for the UD by the network and the UD actively monitors the downlink control channels of its serving base station. A user in RRC_CONNECTED state moves into the RRC_IDLE state if there is no data communication between the UD and the network for a certain timeout period.

[0040] Some applications requiring various types of latency targets can be served using various QoS classes of existing wireless systems, however they are not designed to deliver ultra-low-latency guarantees of the order of 1msec as proposed in 5G. Additionally, the existing wireless systems are typically deployed in frequency bands below 6GHz, in which the impacts of LOS blockages are not as severe as in the mmWave band. The radio link failure mitigation schemes used for these systems are inadequate to satisfy the low-latency requirements of a 5G wireless system for example.

[0041] Some examples of low-latency services and applications include human 'tactile interactions' where touch interfaces are used for control, particularly for remote work in the network cloud, may have delay requirement as low as 1 msec. Interactive remote multimedia applications, end-to-end latencies of less than 10msec are required, which may impose ~1msec latency requirement in the wireless segment. Augmented reality applications for entertainment and information retrieval may require both very low latencies and significant data rates.

[0042] Delivering ultra-low-latency performance in 5G wireless systems imposes an additional burden on the radio frame resources, such as:

- Fast radio link blockage detection by APs during downlink and uplink data transfer require allocation of a FastACK channel on the uplink for every downlink control message.
- UD monitoring of DL control channel of more than one AP in its cluster set in a TDM manner to detect radio link blockage and execute rapid rerouting protocol as well as to enable fast uplink data transfer. This TDM monitoring constraint of the UD, in turn, may make it necessary for an AP to allocate additional DL control channels.
- Low-latency uplink access requires allocation of UD-specific dedicated UL polling channel.
- Staggered allocation of uplink polling channels in the APs of the cluster set of a UD helps to make the uplink access robust against frequent radio link blockages, which requires allocation of uplink polling channels not only in the serving AP, but also the stand-by APs.

[0043] Some applications or services will not require the ultra-low-latency guarantees; and therefore, incurring the ultra-low-latency overheads described above for these applications or services is undesirable. Additionally, users who require the latency-critical services may not always be running sessions for those services or applications. Typically, a user device establishes connection with the network and then may request a latency-critical service. A session for the service begins and lasts for a finite duration and at the end of the session, the user may remain connected to the network. For these users, it is undesirable to incur the ultra-low-latency overheads all the time. Thus, it is expected that among all the users in the system, a fraction of them will be running latency-critical sessions simultaneously.

[0044] The additional DL and UL control channel resources necessary for delivering ultra-low-latency performance reduces the available radio frame resources for data channels. With features as described herein, the problem of how to efficiently manage these control channel overheads in an mmWave air-interface to minimize the adverse impacts on user data rates and system capacity may be addressed.

[0045] According to exemplary embodiments described herein, a new radio resource control (RRC) state is used to deliver latency-critical services. According to exemplary embodiments the following three states are provided in an mmWave system: idle state (e.g. RRC_IDLE); connected state (e.g. RRC_CONNECTED); and low-latency state (RRC_VZL). The low-latency state may be referred to as a 'Virtual-Zero Latency' state, and may be denoted as RRC_VZL. The functions of the various components of the system depend on the state of the UD. Accordingly, the RRC state of the UD is maintained in the components of the system (including the access points in the cluster set, the cluster set manager, and the user device).

[0046] Referring also to Fig. 6, a diagram is shown illustrating the RRC states, RRC state transitions, and corresponding functions of components in an exemplary system (such as, e.g., an mmWave 5G system) according to exemplary embodiments. Each of the RRC_IDLE 602, RRC_CONNECTED 604, and RRC_VZL states are discussed in more detail below with reference to Fig. 6.

Idle State (RRC_IDLE)

[0047] In the RRC_IDLE 602, a UD is registered in the network, but does not have any connection or service flow established over which user plane data can be communicated. The UD may perform cell selection and re-selection to determine a suitable cell to camp. The UD may periodically wake up, based on its DRX cycle, to monitor the paging channel and acquire system information. The UD may move to RRC_CONNECTED 604 when it establishes a RRC connection in response to a call received from the network or initiated by the user.

Connected State (RRC_CONNECTED)

[0048] According to exemplary embodiments, a UD in RRC_CONNECTED 604 is attached with a serving AP and has established a connection with the network. The UD may communicate user plane data with the network. The additional functions of the various components of the system for this RRC_CONNECTED 604 can be described as follows.

[0049] User device functions: The UD in RRC_CONNECTED state may perform at least one or more of the following:

- Monitor the synchronization channel of the UD's serving AP and maintain beam and DL/UL control channel synchronization with the serving AP. The UD may also acquire the best beams and DL/UL control channel configurations for the stand-by APs in its cluster set.
- Monitor the DL control channel of the UD's serving AP for radio frame resource allocation and various control and configuration messages.
- Perform uplink access over the RACH or dedicated UL polling channel of the serving AP.
- Periodically send channel state information (CSI) feedback to the UD's serving AP. The CSI information may include one or more best beams of the AP for downlink and uplink communication, and/or SINR for the best beams.
- Detect radio link blockage to the APs in its cluster set based on non-detection of the synchronization channel and/or non-detection of response to an uplink access request.
- In response to detection of radio link blockage to its serving AP, the UD may send a rerouting request over the RACH channel of one of the UD's stand-by APs in its cluster set. The UD may begin monitoring the DL control channel of the stand-by AP.

[0050] Serving access point functions: When the UD is in the RRC_CONNECTED state, the serving access point of the UD may perform at least one or more of the following:

- Maintain beam synchronization with the UD based on CSI feedback from UD.
- Allocate radio frame resources for downlink and uplink data transmission, as necessary, using the DL control channel.
- Detect radio link blockage to the UD based on at least one of: non-detection of HARQ ACK/NACK feedback during downlink data transmission, and non-detection of signals on uplink data channel during uplink data transmission.
- On detection of radio link blockage to the UD, the serving-AP may suspend further communication with the UD and may notify the blockage event to the CSM of the UD.

[0051] Stand-by access point functions: When the UD is in the RRC_CONNECTED state, a stand-by access point in the cluster set of the UD may perform at least:

- Receive a request for handover from the UD and forward the request to the CSM of the UD.
- On receiving a request, the stand-by AP may allocate a DL control channel and an UL polling channel for the UD and then send the allocation information to the CSM or the UD.

[0052] Cluster Set Manager (CSM) functions: The cluster set manager is UD specific and manages the cluster of APs of the UD. When the UD is in the RRC_CONNECTED state, the cluster set manager of the UD may perform at least one or more of the following:

- On receiving a notification of radio link blockage from the serving AP, suspend forwarding packets to the serving AP;
- On receiving a request for handover of the UD to a stand-by AP: configure the stand-by AP s the new serving AP and reroute data or notify anchor node to reroute data packets to the new serving AP.
- On receiving a request for initiation of a latency-critical service session for a UD, CSM may request the cluster APs of the UD to allocate DL control channels for the UD to monitor and UL polling channel for use by the UD for uplink access; on receiving these allocations from the cluster APs, the CSM sends these allocation information to the UD.

[0053] The role of the ‘anchor node’ is to hide the frequent handoffs of the UD from the network. All data packets for the UD are sent to the anchor node, which in turn forward them to the current serving AP of the UD. In some embodiments, the CSM itself may act as the anchor node, or, it may be some other element in the network.

[0054] State transitions: The UD in the RRC_CONNECTED state may transition to different state according to the following:

- The UD may move to the RRC_IDLE 602 state if all RRC connections of the UD are released. The RRC connections may be released after a period of inactivity or explicitly based on a connection release message from the network.
- The UD may move to the RRC_VZL 606 state, based on when a new latency-critical session is initiated. For example, a QoS value in the connection configuration parameter may indicate that a session is latency-critical.

Low-latency State (RRC_VZL)

[0055] According to exemplary embodiments, a UD in RRC_VZL 606, in addition to being attached to its serving AP, may also maintain connectivity with the stand-by APs in its cluster set. The additional functions of the various components of the system for the RRC_VZL 606 state can be described as follows.

[0056] User device functions: When the UD is in the RRC_VZL 606 state, the UD may perform at least one or more of the following:

- Monitor the synchronization channel of the UD’s serving AP and the UD’s stand-by APs and maintain beam and DL/UL control channel synchronization with both the serving AP and the stand-by APs.
- Monitor the DL control channel of the serving AP and one or more stand-by APs for radio frame resource allocation and various control and configuration messages.

- Perform uplink access over the RACH of the serving AP and the stand-by APs, or perform uplink access over the dedicated UL polling channels allocated for the UD in the APs in its cluster set.
- Periodically send CSI feedback for the UD's serving AP as well as the stand-by APs in the UD's cluster set. The CSI feedback for an AP may be sent to the serving AP, or may be sent to the corresponding AP. The CSI feedback may also be triggered by an event such as, the UD detecting a new best beam, or the SINR of the best beam is different from the previously reported SINR by a threshold value.
- Detect radio link blockage to the UD's serving AP based on at least one of: non-detection of synchronization channel, and non-detection of a response to an uplink access request by the UD.
- On detection of radio link blockage to the UD's serving AP, the UD may send a rerouting request over the RACH channel or uplink polling channel of one of the stand-by APs in its cluster set.
- On receiving a handover command from a stand-by AP, initiates handover to the stand-by AP, making the stand-by AP as the new serving AP.

[0057] Serving access point functions: In addition to the functions described above for the RRC_CONNECTED 604 state, the serving access point may perform at least one or more of the following when the UD is in the RRC_VZL 606 state:

- Forward the accessibility information, which includes the CSI feedbacks, for the stand-by APs (in addition to the serving AP) received from the UD to the CSM of the UD.
- Allocate one or more DL control channels; assign a DL control channel for the UD to monitor; allocate UL polling channel for the UD.
- Allocate a FastACK UL feedback channel for every DL control channel transmission to the UD.
- Detect radio link blockage to the UD based on DTX detection on the FastACK channel in response to the DL control message transmission.
- On detection of radio link blockage, the serving AP may send a fast handoff request for the UD to its CSM.
- On receiving a forwarded UL resource allocation request for the UD from the CSM or a stand-by AP, may allocate UL resources to the UD.

[0058] Stand-by access point functions: When the UD is in the RRC_VZL 606 state, a stand-by access point may perform at least one or more of the following:

- On receiving a request, the stand-by AP may allocate a DL control channel and an UL polling channel for the UD and then send the allocation information to the CSM or to the UD.

- The stand-by AP monitors for uplink transmissions from the UD over the UL polling channel it has allocated for the UD, which may be beamformed.
- Depending on the UL data transfer strategy, on receiving UL resource request, the stand-by AP may allocate UL resources for the UD and/or forward the request to the CSM or the serving AP.

[0059] Cluster Set Manager (CSM) functions: When the UD is in the RRC_VZL 606 state, a cluster set manager may perform at least one or more of the following:

- Receive the UD's accessibility information for the APs in the UD's cluster set, which includes CSI feedbacks for the APs; maintains beam and control channel synchronization between the UD and the cluster APs. For the synchronization, if necessary due to changes in accessibility information, may send request to APs in the cluster set for DL/UL control channel re-allocation and beam updates.
- On receiving a notification of radio link blockage or on receiving a handover request for the UD: selects a new serving AP from among the available stand-by APs and sends handover command to the selected standby AP and configure the stand-by AP as the new serving AP, and reroutes data packets for the UD, or notifies the anchor node to reroute the data packets for the UD, to the new serving AP.
- On receiving an UL resource allocation request for the UD from a stand-by AP, may forward the request to the serving-AP of the UD.

[0060] State transitions: When all latency-critical sessions end, the UD may move from the RRC_VZL 606 state to the RRC_CONNECTED 604 state. The DL and UL control channels allocated for the UD for its low-latency operation may be released.

[0061] According to some embodiments, the radio resource allocation for the low latency service is managed by the cluster set manager of the use device. The user device may send the request to its cluster set manager, or the user device may directly send the request to the access points in its cluster set. In other embodiments, the request originates from a network element other than the user device.

[0062] Referring also to Fig. 7, a diagram is shown illustrating example sub-states of the RRC_VZL 606 state according to exemplary embodiments. In these embodiments the efficiency of the system may be further improved by dividing the RRC_VZL 602 state into two sub-states, which may be denoted VZL_DATA_XFER 702 and VZL_UL_ACCESS 704. The use of these two sub-states may be particularly helpful, e.g., when a latency-critical session may require to deliver low-latency performance only for downlink and uplink data transfer bursts but not for uplink access after a short period of inactivity

[0063] According to exemplary embodiments, in the VZL_DATA_XFER 702 sub-state, radio frame resources are allocated in the cluster APs of the UD for supporting rapid rerouting during downlink and uplink data transfer bursts. A UD may move into VZL_DATA_XFER 702 from RRC_CONNECTED state when it begins a session of an application/service that requires low-latency guarantee during data transfer bursts, but the session does not require low-latency uplink access and can tolerate delays due to link blockages during uplink access. Thus, the system can avoid the UL polling channel allocation overheads at the APs in the cluster set for user devices in the VZL_DATA_XFER 702 state.

[0064] According to exemplary embodiments, in the VZL_UL_ACCESS 704 sub-state, UL polling channels are allocated for the UD in its cluster set to deliver low-latency and robust uplink access mechanism. Since, uplink access eventually progresses into, at least, a burst of uplink data or status transfer, radio frame resources are allocated for the UD for supporting rapid rerouting during the data transfer. The UD may move into VZL_UL_ACCESS 704 sub-state from the RRC_CONNECTED state or VZL_DATA_XFER 702 state when it begins a session in which uplink access is latency-critical.

[0065] Referring also to Fig. 8, a diagram is shown illustrating RRC states for a UD in a mmWave air-interface with DRX mode. In order to improve UD's battery life savings in the RRC_CONNECTED state, a discontinuous reception (DRX). In this mode of operation, after a certain period of inactivity in a *Continuous_Reception* 802 mode, the UD may move to a *Discontinuous Reception* (DRX) cycle. A DRX cycle consists of an 'ON duration', during which the UD may monitor the downlink control channel and a 'DRX period', during which the UD skips reception of downlink channels. According to exemplary embodiments, the RRC_CONNECTED 604 may be split into three sub-states, *Continuous_Reception* 802, *Short_DRX* 804 and *Long_DRX 806*. The properties of the *Continuous_Reception* state is same as those of the RRC_CONNECTED 604 state, whereas, Short*_DRX* and *Long_DRX* states correspond to the short and long DRX cycles. The *Long_DRX* 806 state has a longer DRX period for higher battery savings. The transitions between *Short_DRX* 804, *Long_DRX* 806 and *Continuous_Reception* 802 may be controlled either by timers or explicit controls from the network.

[0066] In the prior art all applications/services are served in the RRC-connected state of the UE; the required performances (throughput, latency) are delivered by the scheduler of the serving AP; radio link failures requires re-establishing RRC connections with new AP, during which performance guarantees are not available.

[0067] According to exemplary embodiments, the new RRC_VZL state significantly improves scalability of the control channel overheads required for serving latency-critical applications, which in turn improves the capacity of the network by allowing more users to be served in a realistic scenario where users are running applications with diverse performance requirements. Further, capturing the radio frame resource requirements for latency-critical services into the new states (RRC_VZL, or VZL_DATA_XFER or VZL_UL_ACCESS), implementation of the functional procedures at various layers of the protocol stack becomes easier.

[0068] For example, depending on the UD state:

- At the MAC layer: (1) the serving AP may or may not allocate the FastACK channel for a feedback from the UD, and (2) the APs may select the link blockage detection procedure (e.g. based on FastACK, based on HARQ ACK/NACK, etc.);
- In the RRC protocol: the CSM may determine its action in the event of a link blockage detection by the serving AP (e.g. whether to initiate a fast handover or not).

[0069] Without the use of the new RRC state for the UD, it would be necessary for the MAC and RRC protocol to have access to the higher layer service configuration parameters for each services running for the UD.

[0070] Referring also to Fig. 9, an example method may comprise transmitting, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device as indicated by block 900; receiving, from the cluster set manager, a radio resource allocation for the low-latency operation as indicated by block 902; and in response to receiving the radio resource allocation for the low-latency operation, triggering the user device to change from a connected state to a low-latency state, wherein the user device may maintain connectivity with multiple access points in the cluster set while the user device is in the low-latency state as indicated by block 904.

[0071] The user device may maintain connectivity with the multiple access points in the cluster set by at least one of: monitoring a synchronization channel of each of the multiple access points to maintain beam synchronization with each of the multiple access points, and monitoring a downlink control channel of each of the multiple access points based at least on the radio resource allocation. The radio resource allocation received by the user device may include at least one of: a downlink control channel allocated for the user device in each of the multiple access points in the cluster set; an uplink polling channel allocated for the user device in each of the multiple access points in the cluster set; and transmit and receive beams to be used by the user device with each of the multiple access points in the cluster set. The method may further comprise: receiving a downlink control message on at least one allocated downlink control channel while the user device is in the low-latency state; and transmitting an indication, on an uplink feedback channel allocated for the user device, that the downlink control message was successfully received and decoded. The method may further comprise: transmitting, when the user device is in the low-latency state, an uplink resource request or information during an uplink access opportunity of at least one of the multiple access points; wherein the uplink access opportunity may be at least one of: an uplink RACH and an uplink polling channel allocated for the user device. The method may further comprise: in response to completion of all low-latency operations of the user device, changing from the low-latency state to the connected state. The method may further comprise: sending channel state information (CSI) for each of the multiple access points, periodically or in response to an event-trigger; the CSI comprising at least one of: the best downlink and uplink beams at the access points and at the user device, and the SINR for the best beams.

[0072] An example embodiment may be provided in an apparatus comprising at least one processor; and at least one non-transitory memory including computer program code, the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus to: transmit, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device; receive, from the cluster set manager, a radio resource allocation for the low-latency operation; and in response to receipt of the radio resource allocation for the low-latency operation, trigger the user device to change from a connected state to a low-latency state, wherein the user device may maintain connectivity with multiple access points in the cluster set while the user device is in the low-latency state.

[0073] The user device may maintain the connectivity with the multiple access points in the cluster set by at least one of: monitoring a synchronization channel of each of the multiple access points to maintain beam synchronization with each of the multiple access points, and monitoring a downlink control channel of each of the multiple access points based at least on the radio resource allocation. The radio resource allocation received by the user device may be at least one of: a downlink control channel allocated for the user device in each of the multiple access points in the cluster set; an uplink polling channel allocated for the user device in each of the multiple access points in the cluster set; and transmit and receive beams to be used by the user device with each of the multiple access points in the cluster set. The at least one memory and the computer program code may be further configured to, with the at least one processor, cause the apparatus to: receive a downlink control message on at least one allocated downlink control channel while the user device is in the low-latency state; and transmit an indication, on an uplink feedback channel allocated for the user device, that the downlink control message was successfully received and decoded.

[0074] An example embodiment may be provided in a non-transitory program storage device, such as the memory 216 shown in Fig. 3 for example, readable by a machine, tangibly embodying a program of instructions executable by the machine for performing operations, the operations comprising: transmitting, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device; receiving, from the cluster set manager, a radio resource allocation for the low-latency operation; and in response to receiving the radio resource allocation for the low-latency operation, triggering the user device to change from a connected state to a low-latency state, wherein the user device may maintain connectivity with multiple access points in the cluster set while the user device is in the low-latency state.

[0075] Referring also to Fig. 10, an example method may comprise receiving, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device as indicated by block 1000; allocating radio resources for the low-latency operation as indicated by block 1002; changing a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point as indicated by block 1004; and in response to receiving an indication that all low-latency operations of the user device have completed; communicate using allocated radio resources for the low latency operation as indicated by block 1006; and changing the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device as indicated by block 1008.

[0076] Allocating the radio resources may include allocating at least one uplink polling channel for the user device based at least on a request from a cluster set manager of the user device, and wherein the method may further include monitoring the allocated uplink polling channels, when the user device is in the low-latency state, for at least one of: an uplink access request transmitted from the user device, and a control message transmitted from the user device. The access point may be a stand-by access point in the cluster set and wherein the method may further include, when the user device is in the low-latency state: receiving, based on the monitoring, an uplink access request from the user device on the allocated uplink polling channel; and allocating uplink resources for the user device based on the uplink access request, and/or forwarding the uplink access request to the serving access point. Allocating the radio resources may include: assigning an existing downlink control channel to be monitored by the user device; and if an existing downlink control channel cannot be assigned, allocating a new downlink control channel in a radio frame which can be assigned. The access point may be a serving access point of the user device and wherein the method further comprises, when the user device is in the low-latency state: transmitting, by the serving access point, a downlink control message to the user device and allocating a feedback channel for the user device; detecting discontinuous transmission on the feedback channel based on the downlink control message; determining radio link blockage between the serving access point and the user device; and transmitting, to a cluster set manager of the cluster set an indication of the radio link blockage. The access point may be a stand-by access point in the cluster set, and the method may further comprise, when the user device is in the low-latency state: receiving, from a cluster set manager of the cluster set, a handover command for the user device, and transmitting the handover command to the user device.

[0077] An example apparatus may be provided comprising at least one processor; and at least one non-transitory memory including computer program code, the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus to: receive, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device; allocate radio resources for the low-latency operation; change a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point; and in response to receipt of an indication that all low-latency operations of the user device have completed, change the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device.

[0078] Allocating the radio resources may include allocating at least one uplink polling channel for the user device based at least on a request from a cluster set manager of the cluster set, and wherein the at least one memory and the computer program code may be configured to, with the at least one processor, cause the apparatus further to monitor the allocated uplink polling channels, when the user device is in the low latency state, for at least one of: an uplink access request transmitted from the user device, and a control message transmitted from the user device. The access point may be a stand-by access point in the cluster set, and the at least one memory and the computer program code may be configured to, with the at least one processor, cause the apparatus further to, when the user device is in the low-latency state: receive, based on the monitoring, an uplink access request from the user device on the allocated uplink polling channel; and allocate uplink resources for the user device based on the uplink access request, and/or forward the uplink access request to the serving access point. The access point may be a serving access point of the user device and wherein the at least one memory and the computer program code may be configured to, with the at least one processor, cause the apparatus further to, when the user device is in the low-latency state: transmit, by the serving access point, a downlink control message to the user device and allocating a feedback channel for the user device; detect discontinuous transmission on the feedback channel based on the downlink control message; determine radio link blockage between the serving access point and the user device; and transmit, to a cluster set manager of the cluster set an indication of the radio link blockage. The access point may be a stand-by access point of the user device, and the at least one memory and the computer program code may be configured to, with the at least one processor, cause the apparatus further to, when the user device is in the low-latency state: receive, from a cluster set manager of the cluster set, a handover command for the user device, and transmit the handover command to the user device

[0079] An example embodiment may be provided in a non-transitory program storage device, such as memory 226 shown in Fig. 3 for example, readable by a machine, tangibly embodying a program of instructions executable by the machine for performing operations, the operations comprising: receiving, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device; allocating radio resources for the low-latency operation; changing a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point; and in response to receiving an indication that all low-latency operations of the user device have completed, changing the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device.

[0080] Referring also to Fig. 11, an example method may comprise receiving an indication, by a cluster set manager of a user device, that a low-latency operation has been initiated as indicated by block 1100; in response to receiving the indication, configuring and allocating radio frame resources in access points for the low-latency operation, wherein the access points are in a cluster set of the user device as indicated by block 1102; notifying the user device of the radio resource allocations, wherein the radio frame resources are at least one of: a downlink control channel for each access point in the cluster set for the user device to monitor, and an uplink polling channel for each access point in the cluster set to be used by the user device for uplink access in the access points as indicated by block 1104; changing a state of the user device maintained by the cluster set manager from a connected to a low-latency state as indicated by block 1106; and on completion of the low-latency operation, changing the state of the user device to the connected state and releasing all resources allocated for low-latency operations if no other low-latency operation is running as indicated by block 1108.

[0081] The method may include, when the user device is in the low-latency state: determining radio link blockage between a serving access point in the cluster set and the user device; and initiating a handover of the user device to a stand-by access point in the cluster set. The method may include, when the user device is in the low-latency state: receiving channel state information of access points in the cluster set, wherein the channel state information may be at least best beam information for the access points in the cluster set; and forwarding the channel state information to the respective access points in the cluster set.

[0082] Any combination of one or more computer readable medium(s) may be utilized as the memory. The computer readable medium may be a computer readable signal medium or a non-transitory computer readable storage medium. A non-transitory computer readable storage medium does not include propagating signals and may be, for example, but not limited to, an electronic, magnetic, optical, electromagnetic, infrared, or semiconductor system, apparatus, or device, or any suitable combination of the foregoing. More specific examples (a non-exhaustive list) of the computer readable storage medium would include the following: an electrical connection having one or more wires, a portable computer diskette, a hard disk, a random access memory (RAM), a read-only memory (ROM), an erasable programmable read-only memory (EPROM or Flash memory), an optical fiber, a portable compact disc read-only memory (CD-ROM), an optical storage device, a magnetic storage device, or any suitable combination of the foregoing.

[0083] If desired, the different functions discussed herein may be performed in a different order and/or concurrently with each other. Furthermore, if desired, one or more of the above-described functions may be optional or may be combined.

[0084] It should be understood that the foregoing description is only illustrative. Various alternatives and modifications can be devised by those skilled in the art. For example, features recited in the various dependent claims could be combined with each other in any suitable combination(s). In addition, features from different embodiments described above could be selectively combined into a new embodiment. Accordingly, the description is intended to embrace all such alternatives, modifications and variances which fall within the scope of the appended claims.

[0085] It is also noted herein that while the above describes example embodiments of the invention, these descriptions should not be viewed in a limiting sense. Rather, there are several variations and modifications which may be made without departing from the scope of the present invention as defined in the appended claims.

CLAIMS

What is claimed is:

1. A method comprising:

transmitting, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device;

receiving, from the cluster set manager, a radio resource allocation for the low-latency operation; and

in response to receiving the radio resource allocation for the low-latency operation, triggering the user device to change from a connected state to a low-latency state, wherein the user device maintains connectivity with multiple access points in the cluster set while the user device is in the low-latency state.

2. The method of claim 1, wherein the user device maintains connectivity with the multiple access points in the cluster set by at least one of:

monitoring a synchronization channel of each of the multiple access points to maintain beam synchronization with each of the multiple access points, and

monitoring a downlink control channel of each of the multiple access points based at least on the radio resource allocation.

3. The method of claim 1, wherein the radio resource allocation received by the user device includes at least one of:

a downlink control channel allocated for the user device in each of the multiple access points in the cluster set;

an uplink polling channel allocated for the user device in each of the multiple access points in the cluster set; and

transmit and receive beams to be used by the user device with each of the multiple access points in the cluster set.

4. The method of claim 1, the method further comprising:

receiving a downlink control message on at least one allocated downlink control channel while the user device is in the low-latency state; and

transmitting an indication, on an uplink feedback channel allocated for the user device, that the downlink control message was successfully received and decoded.

5. The method of claim 1, the method further comprising:

transmitting, when the user device is in the low-latency state, an uplink resource request or information during an uplink access opportunity of at least one of the multiple access points; wherein the uplink access opportunity is at least one of: an uplink RACH and an uplink polling channel allocated for the user device.

6. The method of claim 1, further comprising:

in response to completion of all low-latency operations of the user device, changing from the low-latency state to the connected state.

7. The method of claim 1, further comprising:

sending channel state information (CSI) for each of the multiple access points, periodically or in response to an event-trigger; the CSI comprising at least one of: the best downlink and uplink beams at the access points and at the user device, and the SINR for the best beams.

8. An apparatus comprising:

at least one processor; and

at least one non-transitory memory including computer program code, the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus to:

transmit, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device;

receive, from the cluster set manager, a radio resource allocation for the low-latency operation; and

in response to receipt of the radio resource allocation for the low-latency operation, trigger the user device to change from a connected state to a low-latency state, wherein the user device maintains connectivity with multiple access points in the cluster set while the user device is in the low-latency state.

9. The apparatus of claim 8, wherein the user device maintains the connectivity with the multiple access points in the cluster set by at least one of:

monitoring a synchronization channel of each of the multiple access points to maintain beam synchronization with each of the multiple access points, and

monitoring a downlink control channel of each of the multiple access points based at least on the radio resource allocation.

10. The apparatus of claim 8, wherein the radio resource allocation received by the user device includes at least one of:

a downlink control channel allocated for the user device in each of the multiple access points in the cluster set;

an uplink polling channel allocated for the user device in each of the multiple access points in the cluster set; and

transmit and receive beams to be used by the user device with each of the multiple access points in the cluster set.

11. The apparatus of claim 8, the at least one memory and the computer program code are further configured to, with the at least one processor, cause the apparatus to:

receive a downlink control message on at least one allocated downlink control channel while the user device is in the low-latency state; and

transmit an indication, on an uplink feedback channel allocated for the user device, that the downlink control message was successfully received and decoded.

12. A method comprising:

receiving, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device;

allocating radio resources for the low-latency operation;

changing a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point; and

in response to receiving an indication that all low-latency operations of the user device have completed, changing the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device.

13. The method of claim 12, wherein allocating the radio resources comprises allocating at least one uplink polling channel for the user device based at least on a request from a cluster set manager of the user device, and wherein the method further comprises monitoring the allocated uplink polling channels, when the user device is in the low-latency state, for at least one of:

an uplink access request transmitted from the user device, and

a control message transmitted from the user device.

14. The method of claim 13, wherein the access point is a stand-by access point in the cluster set and wherein the method further comprises, when the user device is in the low-latency state:

receiving, based on the monitoring, an uplink access request from the user device on the allocated uplink polling channel; and

allocating uplink resources for the user device based on the uplink access request, and/or forwarding the uplink access request to the serving access point.

15. The method of claim 12, wherein allocating the radio resources comprises:

assigning an existing downlink control channel to be monitored by the user device; and

if an existing downlink control channel cannot be assigned, allocating a new downlink control channel in a radio frame which can be assigned.

16. The method of claim 12, wherein the access point is a serving access point of the user device and wherein the method further comprises, when the user device is in the low-latency state:

transmitting, by the serving access point, a downlink control message to the user device and allocating a feedback channel for the user device;

detecting discontinuous transmission on the feedback channel based on the downlink control message;

determining radio link blockage between the serving access point and the user device; and

transmitting, to a cluster set manager of the cluster set an indication of the radio link blockage.

17. The method of claim 12, wherein the access point is a stand-by access point in the cluster set, and wherein the method further comprises, when the user device is in the low-latency state:

receiving, from a cluster set manager of the cluster set, a handover command for the user device, and

transmitting the handover command to the user device.

18. An apparatus comprising:

at least one processor; and

at least one non-transitory memory including computer program code, the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus to:

receive, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device;

allocate radio resources for the low-latency operation;

change a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point; and in response to receipt of an indication that all low-latency operations of the user device have completed, change the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device.

19. The apparatus of claim 18, wherein allocating the radio resources comprises allocating at least one uplink polling channel for the user device based at least on a request from a cluster set manager of the cluster set, and wherein the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus further to monitor the allocated uplink polling channels, when the user device is in the low latency state, for at least one of:

an uplink access request transmitted from the user device, and

a control message transmitted from the user device.

20. The apparatus of claim 19, wherein the access point is a stand-by access point in the cluster set, and wherein the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus further to, when the user device is in the low-latency state:

receive, based on the monitoring, an uplink access request from the user device on the allocated uplink polling channel; and

allocate uplink resources for the user device based on the uplink access request, and/or forward the uplink access request to the serving access point.

21. The apparatus of claim 18, wherein the access point is a serving access point of the user device and wherein the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus further to, when the user device is in the low-latency state:

transmit, by the serving access point, a downlink control message to the user device and allocating a feedback channel for the user device;

detect discontinuous transmission on the feedback channel based on the downlink control message;

determine radio link blockage between the serving access point and the user device; and

transmit, to a cluster set manager of the cluster set an indication of the radio link blockage.

22. The apparatus of claim 18, wherein the access point is a stand-by access point of the user device, and wherein the at least one memory and the computer program code configured to, with the at least one processor, cause the apparatus further to, when the user device is in the low-latency state:

receive, from a cluster set manager of the cluster set, a handover command for the user device, and

transmit the handover command to the user device.

23. A method comprising:

receiving an indication, by a cluster set manager of a user device, that a low-latency operation has been initiated;

in response to receiving the indication, configuring and allocating radio frame resources in access points for the low-latency operation, wherein the access points are in a cluster set of the user device;

notifying the user device of the radio resource allocations, wherein the radio frame resources are at least one of: a downlink control channel for each access point in the cluster set for the user device to monitor, and an uplink polling channel for each access point in the cluster set to be used by the user device for uplink access in the access points;

changing a state of the user device maintained by the cluster set manager from a connected to a low-latency state; and

on completion of the low-latency operation, changing the state of the user device to the connected state and releasing all resources allocated for low-latency operations if no other low-latency operation is running.

24. The method of claim 23, the method further comprising, when the user device is in the low-latency state:

determining radio link blockage between a serving access point in the cluster set and the user device; and

initiating a handover of the user device to a stand-by access point in the cluster set.

25. The method of claim 23, the method further comprising, when the user device is in the low-latency state:

receiving channel state information of access points in the cluster set, wherein the channel state information comprises at least best beam information for the access points in the cluster set; and

forwarding the channel state information to the respective access points in the cluster set.

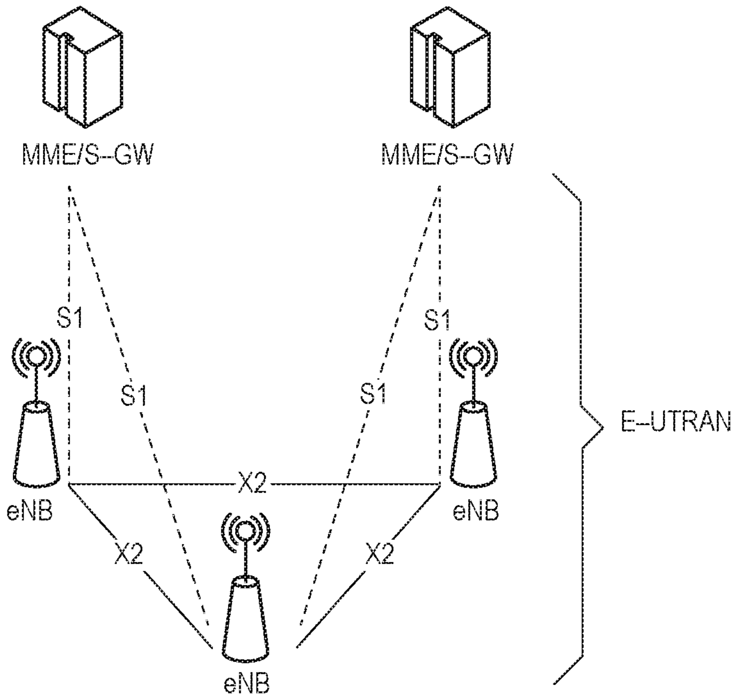
MME/S--GW
MME/S--GW
S1
S1
S1
S1
X2
X2
X2
eNB
eNB
eNB
E--UTRAN

FIG. 1

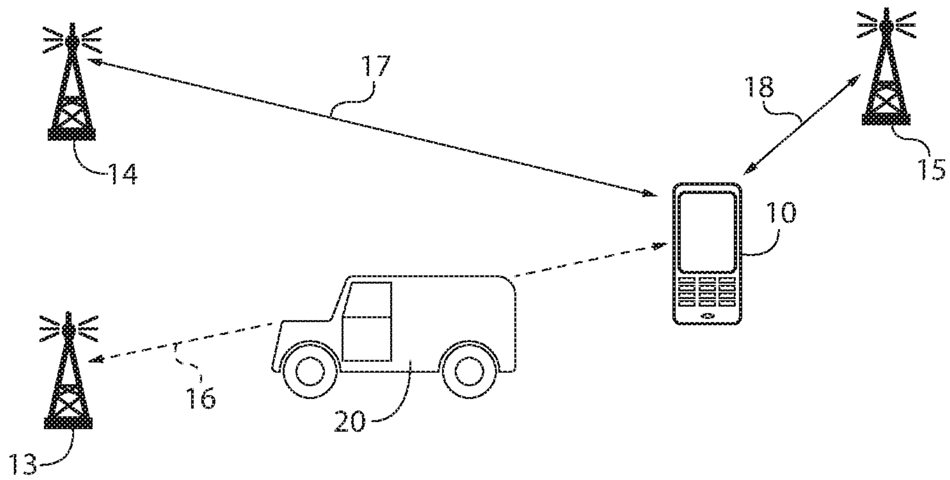
17
18
14
15
10
16
20
13

FIG. 2

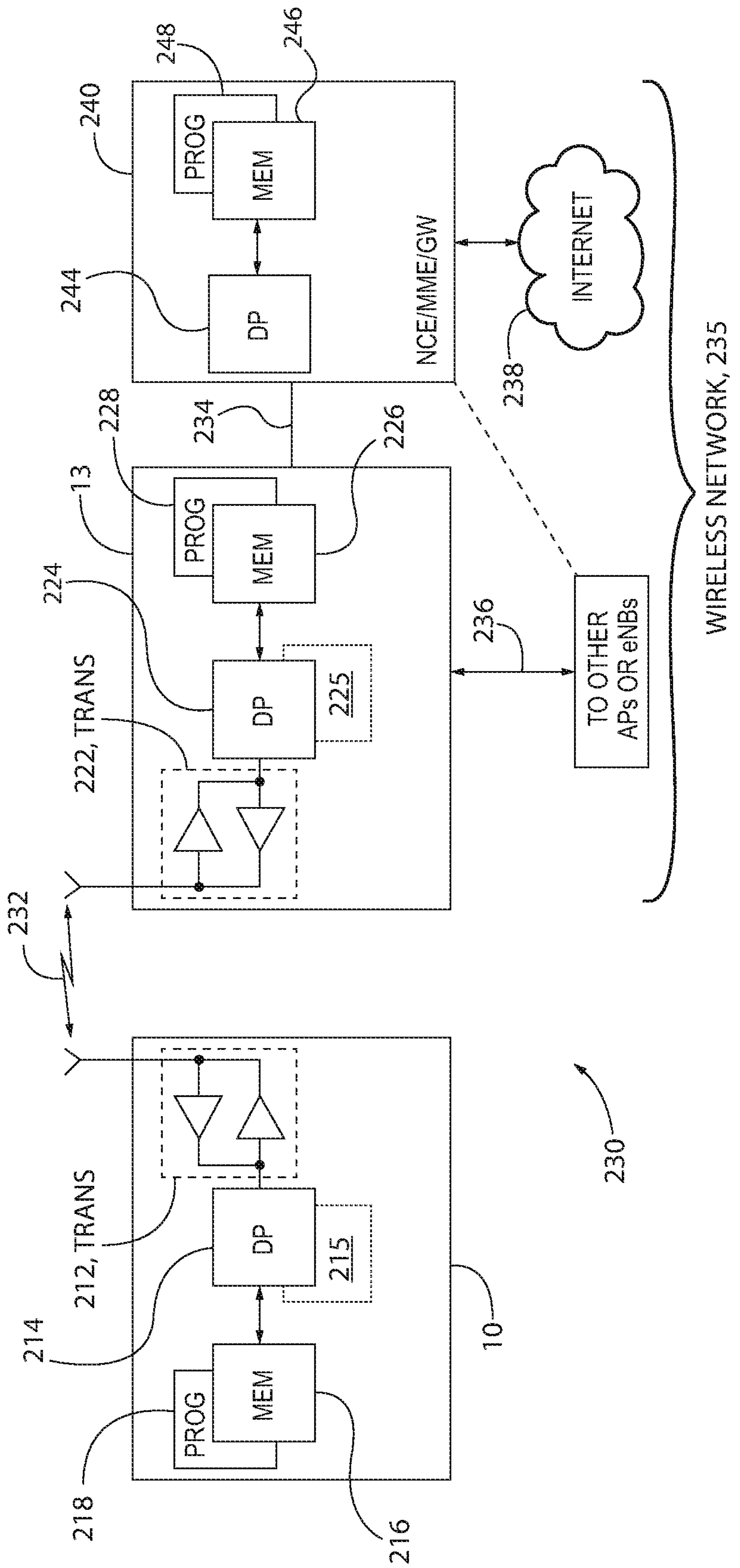
PROG
MEM
DP
248
246
240
244
NCE/MME/GW
INTERNET
238
234
228
226
13
224
225
222, TRANS
236
TO OTHER
APs OR eNBs
WIRELESS NETWORK, 235
232
212, TRANS
215
214
10
230
218
216

FIG. 3

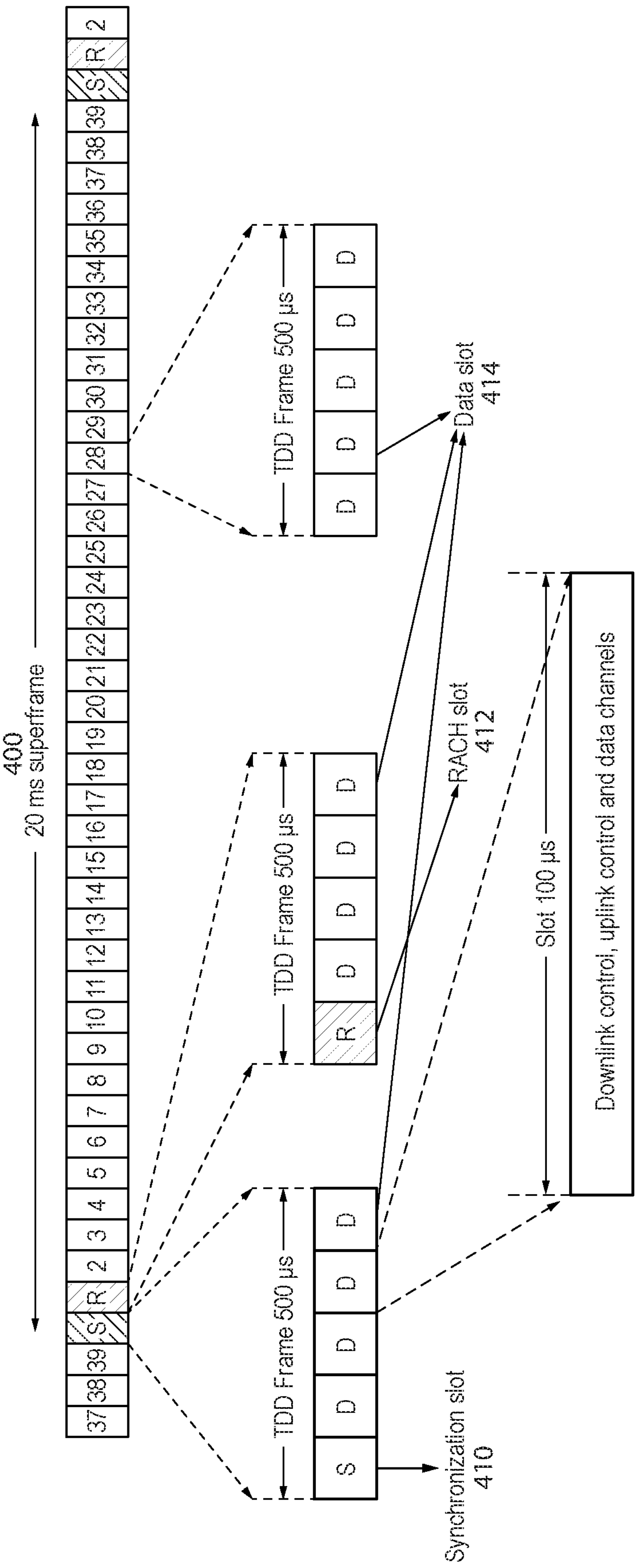
400
20 ms superframe
37
38
39
S
R
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
S
R
2
TDD Frame 500 μs
S
D
D
D
D
R
D
D
D
D
D
D
D
D
D
Synchronization slot
410
RACH slot
412
Data slot
414
Slot 100 μs
Downlink control, uplink control and data channels

FIG. 4

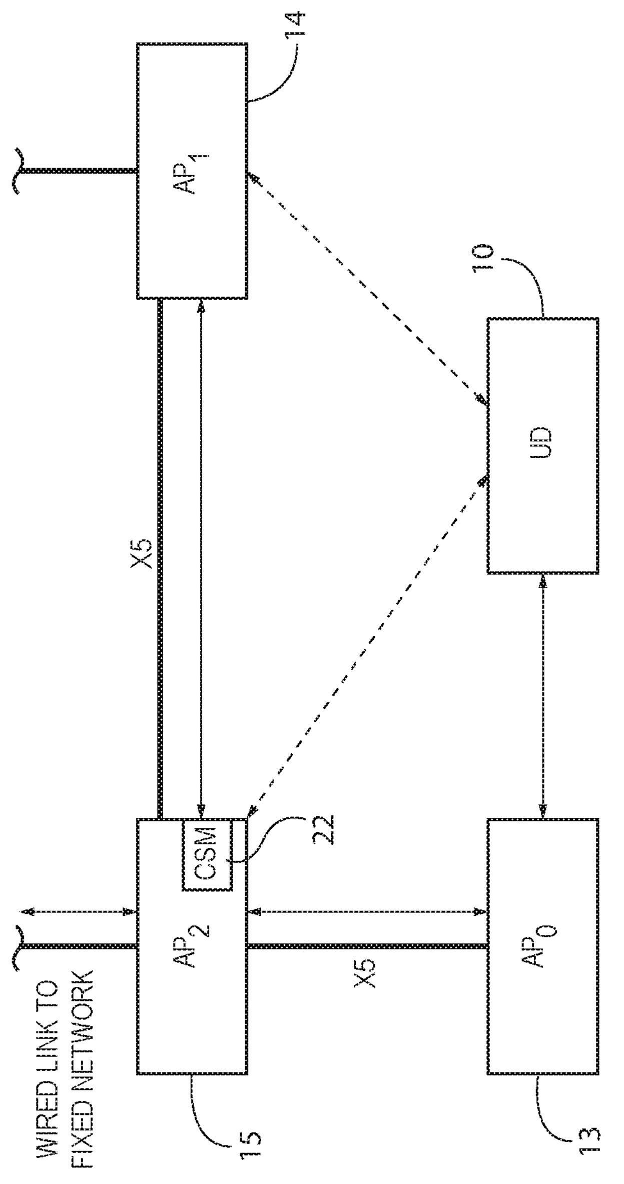
14
AP1
10
UD
X5
CSM
22
AP2
AP0
X5
WIRED LINK TO
FIXED NETWORK
15
13

FIG. 5

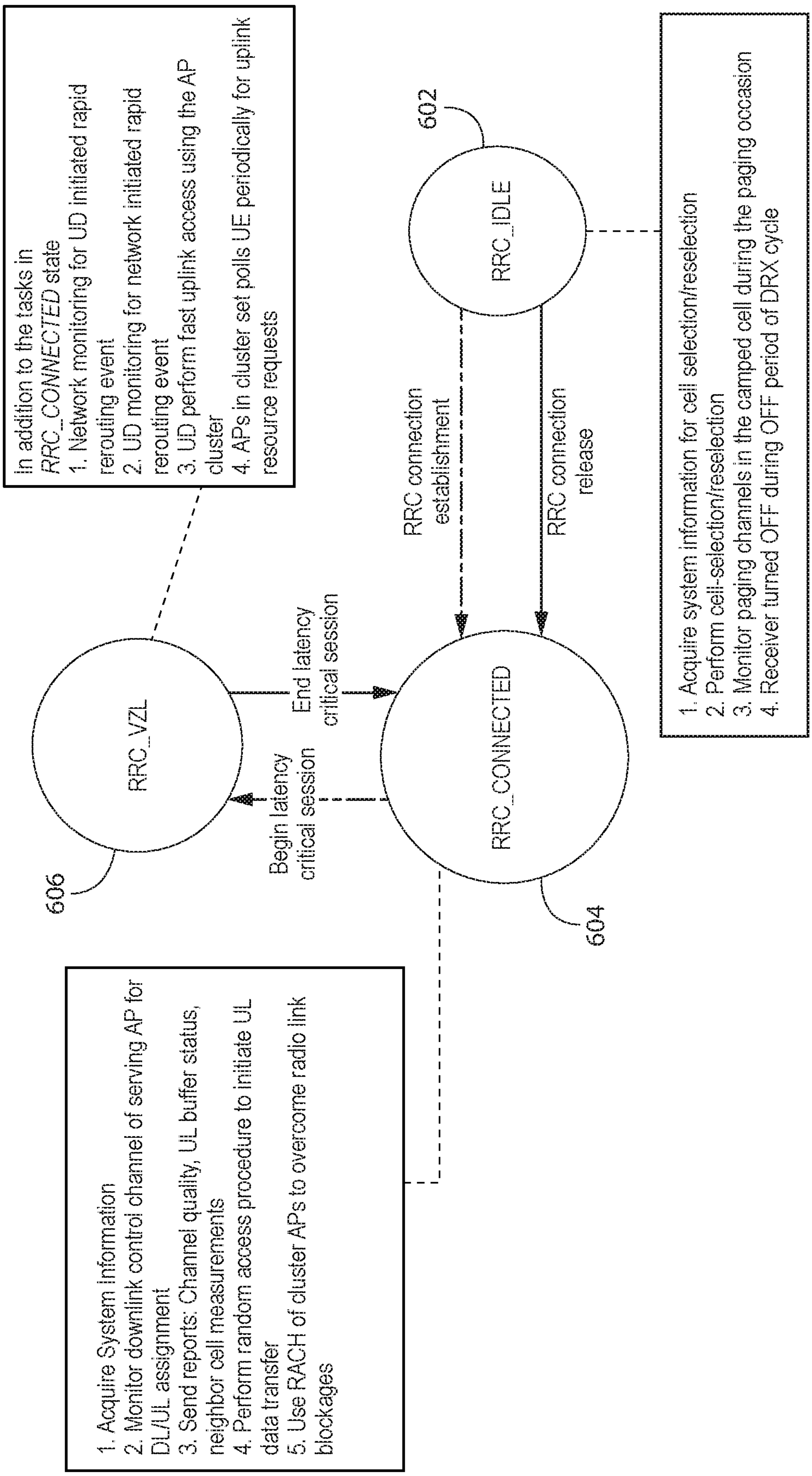
In addition to the tasks in
RRC_CONNECTED state
1. Network monitoring for UD initiated rapid
rerouting event
2. UD monitoring for network initiated rapid
rerouting event
3. UD perform fast uplink access using the AP
cluster
4. APs in cluster set polls UE periodically for uplink
resource requests
602
RRC_IDLE
RRC connection
establishment
RRC connection
release
1. Acquire system information for cell selection/reselection
2. Perform cell-selection/reselection
3. Monitor paging channels in the camped cell during the paging occasion
4. Receiver turned OFF during OFF period of DRX cycle
606
RRC_VZL
End latency
critical session
Begin latency
critical session
RRC_CONNECTED
604
1. Acquire System information
2. Monitor downlink control channel of serving AP for
DL/UL assignment
3. Send reports: Channel quality, UL buffer status,
neighbor cell measurements
4. Perform random access procedure to initiate UL
data transfer
5. Use RACH of cluster APs to overcome radio link
blockages

FIG. 6

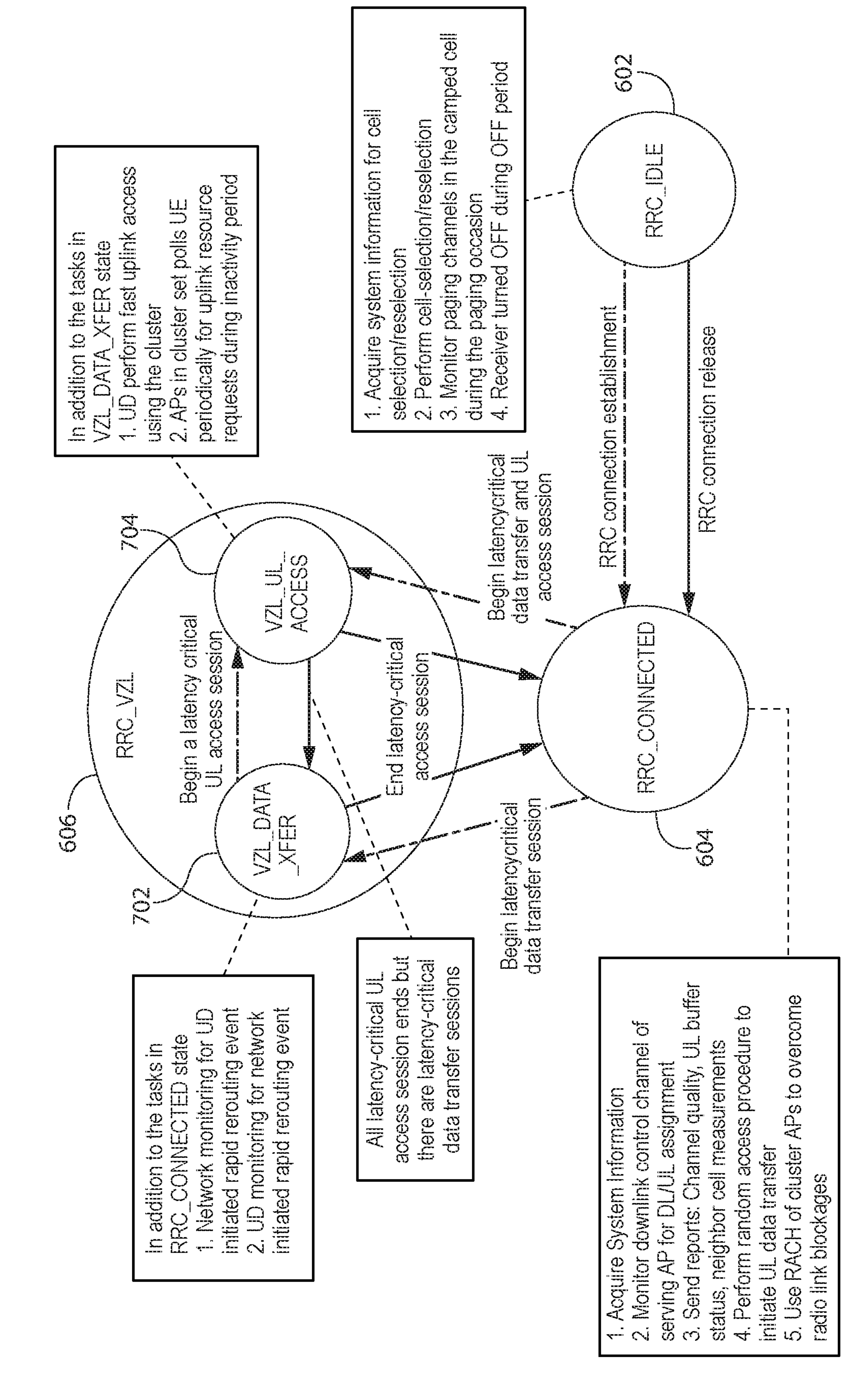
1. Acquire system information for cell selection/reselection
2. Perform cell-selection/reselection
3. Monitor paging channels in the camped cell during the paging occasion
4. Receiver turned OFF during OFF period
602
RRC_IDLE
In addition to the tasks in VZL_DATA_XFER state
1. UD perform fast uplink access using the cluster
2. APs in cluster set polls UE periodically for uplink resource requests during inactivity period
RRC connection establishment
RRC connection release
Begin latencycritical data transfer and UL access session
704
VZL_UL_ACCESS
RRC_VZL
Begin a latency critical UL access session
End latency-critical access session
606
VZL_DATA_XFER
702
RRC_CONNECTED
604
Begin latencycritical data transfer session
In addition to the tasks in RRC_CONNECTED state
1. Network monitoring for UD initiated rapid rerouting event
2. UD monitoring for network initiated rapid rerouting event
All latency-critical UL access session ends but there are latency-critical data transfer sessions
1. Acquire System Information
2. Monitor downlink control channel of serving AP for DL/UL assignment
3. Send reports: Channel quality, UL buffer status, neighbor cell measurements
4. Perform random access procedure to initiate UL data transfer
5. Use RACH of cluster APs to overcome radio link blockages

FIG. 7

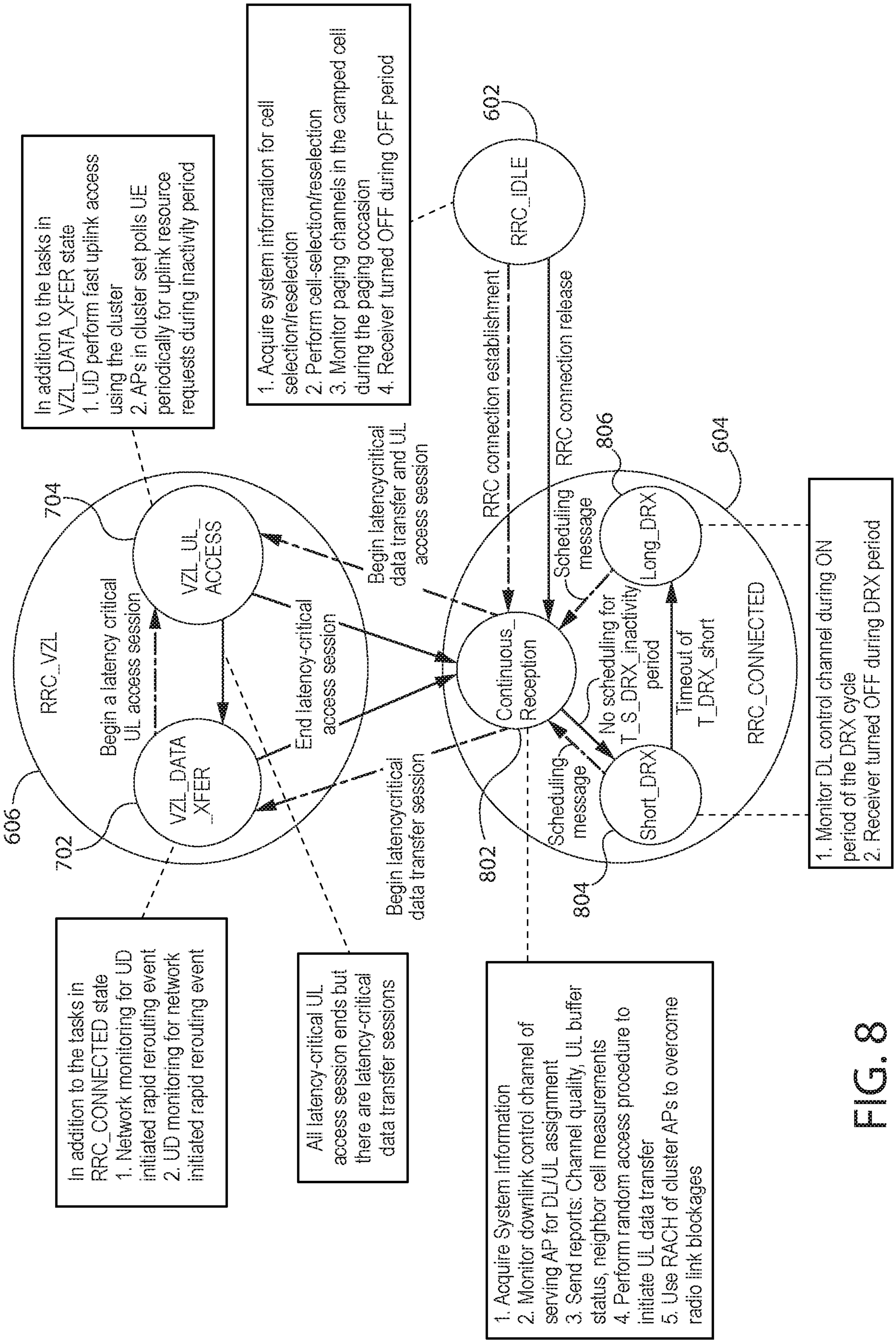
In addition to the tasks in VZL_DATA_XFER state
1. UD perform fast uplink access using the cluster
2. APs in cluster set polls UE periodically for uplink resource requests during inactivity period
1. Acquire system information for cell selection/reselection
2. Perform cell-selection/reselection
3. Monitor paging channels in the camped cell during the paging occasion
4. Receiver turned OFF during OFF period
602
RRC_IDLE
RRC connection establishment
RRC connection release
704
VZL_UL_ACCESS
RRC_VZL
Begin a latency critical UL access session
Begin latencycritical data transfer and UL access session
End latency-critical access session
806
Scheduling message
Long_DRX
604
Continuous_Reception
No scheduling for T_S_DRX_inactivity period
Timeout of T_DRX_short
RRC_CONNECTED
606
VZL_DATA_XFER
702
Begin latencycritical data transfer session
802
Scheduling message
Short_DRX
804
1. Monitor DL control channel during ON period of the DRX cycle
2. Receiver turned OFF during DRX period
In addition to the tasks in RRC_CONNECTED state
1. Network monitoring for UD initiated rapid rerouting event
2. UD monitoring for network initiated rapid rerouting event
All latency-critical UL access session ends but there are latency-critical data transfer sessions
1. Acquire System Information
2. Monitor downlink control channel of serving AP for DL/UL assignment
3. Send reports: Channel quality, UL buffer status, neighbor cell measurements
4. Perform random access procedure to initiate UL data transfer
5. Use RACH of cluster APs to overcome radio link blockages

FIG. 8

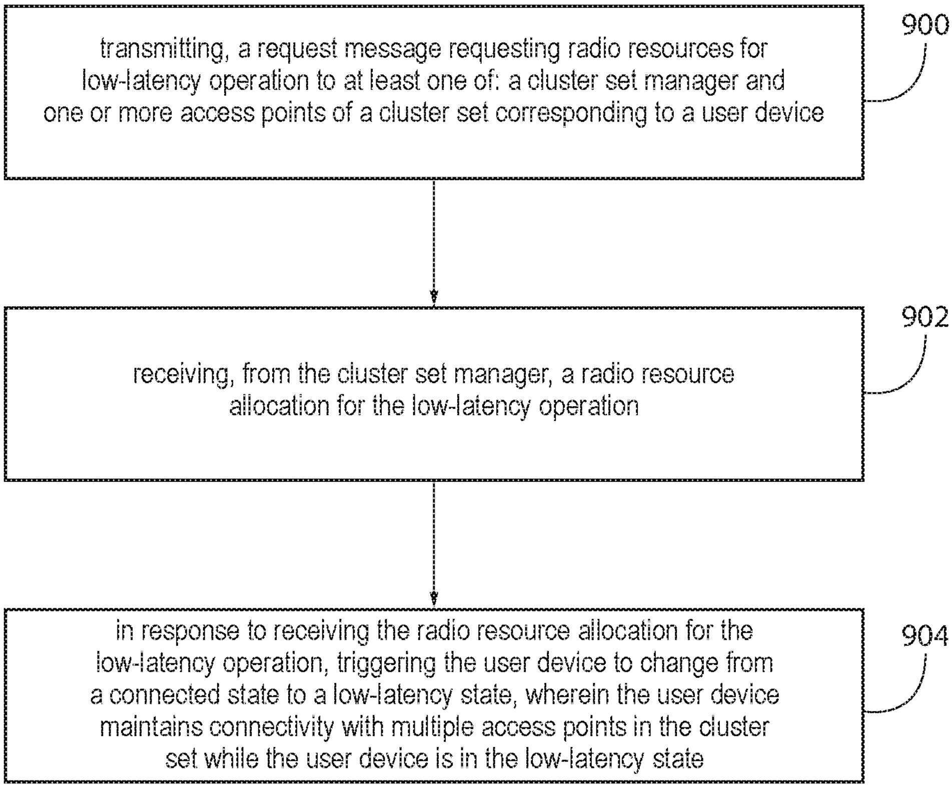
transmitting, a request message requesting radio resources for low-latency operation to at least one of: a cluster set manager and one or more access points of a cluster set corresponding to a user device
900
receiving, from the cluster set manager, a radio resource allocation for the low-latency operation
902
in response to receiving the radio resource allocation for the low-latency operation, triggering the user device to change from a connected state to a low-latency state, wherein the user device maintains connectivity with multiple access points in the cluster set while the user device is in the low-latency state
904

FIG. 9

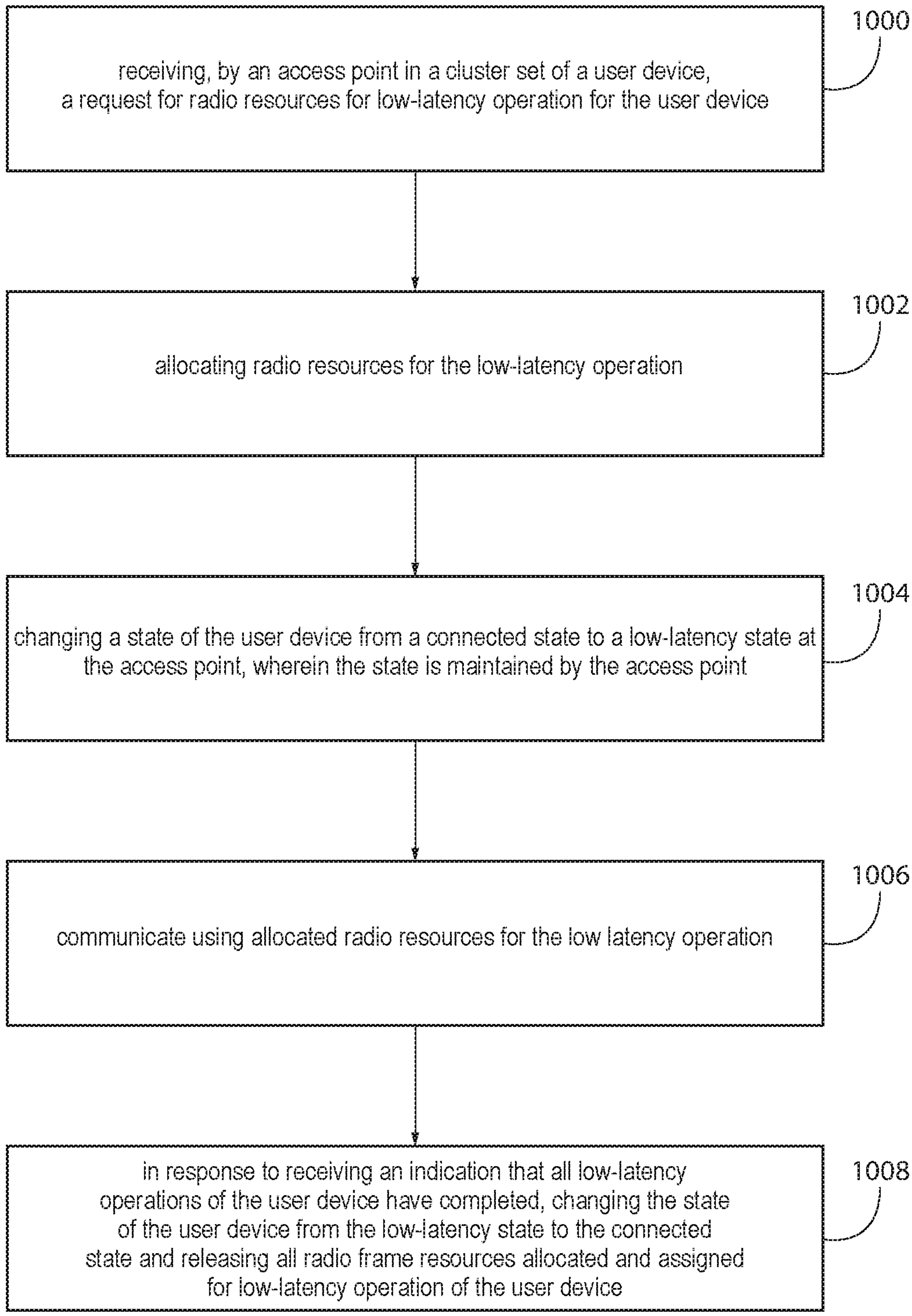
receiving, by an access point in a cluster set of a user device, a request for radio resources for low-latency operation for the user device
1000
allocating radio resources for the low-latency operation
1002
changing a state of the user device from a connected state to a low-latency state at the access point, wherein the state is maintained by the access point
1004
communicate using allocated radio resources for the low latency operation
1006
in response to receiving an indication that all low-latency operations of the user device have completed, changing the state of the user device from the low-latency state to the connected state and releasing all radio frame resources allocated and assigned for low-latency operation of the user device
1008

FIG. 10

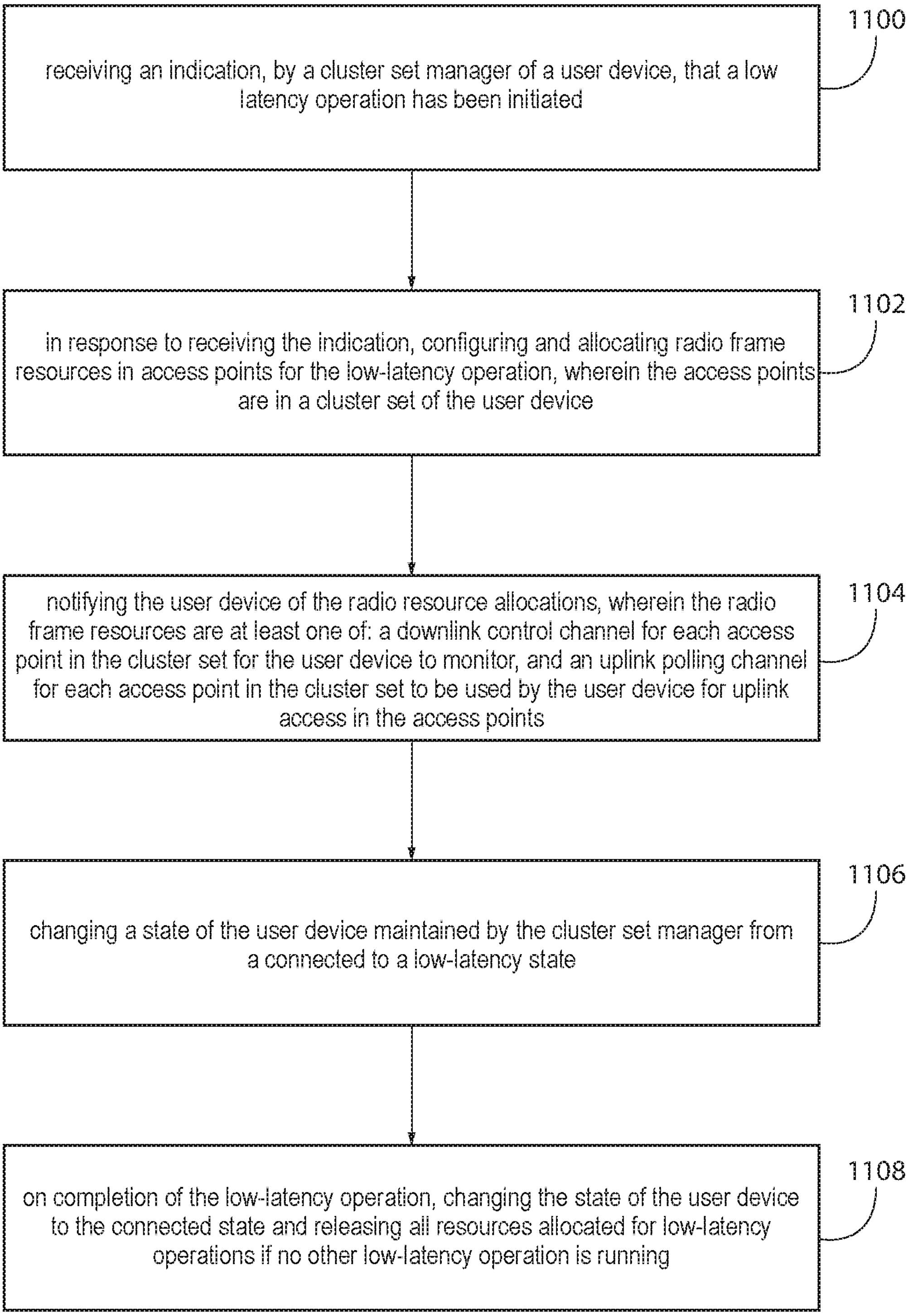
receiving an indication, by a cluster set manager of a user device, that a low latency operation has been initiated
1100
in response to receiving the indication, configuring and allocating radio frame resources in access points for the low-latency operation, wherein the access points are in a cluster set of the user device
1102
notifying the user device of the radio resource allocations, wherein the radio frame resources are at least one of: a downlink control channel for each access point in the cluster set for the user device to monitor, and an uplink polling channel for each access point in the cluster set to be used by the user device for uplink access in the access points
1104
changing a state of the user device maintained by the cluster set manager from a connected to a low-latency state
1106
on completion of the low-latency operation, changing the state of the user device to the connected state and releasing all resources allocated for low-latency operations if no other low-latency operation is running
1108

FIG. 10

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US16/32309

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - H04W 72/04, 72/06, 72/08, 76/06 (2016.01)
CPC - H04W 72/04, 72/0406, 72/0413, 72/0453
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
IPC(8) - H04W 72/04, 72/06, 72/08, 76/06 (2016.01)
CPC - H04W 72/04, 72/0406, 72/0413, 72/0453

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
PatSeer (US, EP, WO, JP, DE, GB, CN, FR, KR, ES, AU, IN, CA, RU, AT,CH,TH,BR,PH, SE, NO, DK, FI, BE, NL, LU, MX,INPADOC Data); Google Scholar/Google Patent/ IP.com; IEEE.org; allocate, resource, base station, access point, handover, low latency, cluster, 5G, channel, mmWave, channel state, connect, switch, operate, uplink, radio, request, transmit*, mm, millimeter, wave, downlink

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	US 2016/0066316 A1 (QUALCOM INC.) March 3, 2016, Figures 7-8, pararagraphs [0060], [0066], [0072], [0075],[0095]-[0095]	1-5, 8-11, 13-15, 17, 19-20, 22 --------------------------- 6-7, 12, 16, 18, 21, 23-25
Y	US 9,270,616 B1 (ARRIS GROUP, INC.) February 23, 2016, column 6, lines 33-38, column 5, lines 58-61	6-7, 12, 16, 18, 21, 23-25
Y	US 2014/0003269 A1 (GOLITSCHEK EDLER VON ELBWART, A. et al.) January 2, 2014, paragraphs [0106]-[0107], [0244].	7, 25
Y	WO 2015/106237 A1 (INTERDIGITAL PATENT HOLDINGS) July 16, 2015, paragraphs [0117], [0172],	16, 21, 24

☐ Further documents are listed in the continuation of Box C. ☐ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
18 July 2016 (18.07.2016)	11 AUG 2016
Name and mailing address of the ISA/ Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-8300	Authorized officer Shane Thomas PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774