



(51) International Patent Classification:
H04L 29/08 (2006.01)

(21) International Application Number:
PCT/US2017/060528

(22) International Filing Date:
08 November 2017 (08.11.2017)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
62/419,874 09 November 2016 (09.11.2016) US

(71) Applicant: **INTERDIGITAL PATENT HOLDINGS, INC.** [US/US]; 200 Bellevue Parkway, Suite 300, Wilmington, Delaware 19809 (US).

(72) Inventors: **PURKAYASTHA, Debashish**; 1200 East Autumn Court, Collegeville, Pennsylvania 19426 (US). **DE FOY, Xavier**; 3007 Rue Edmond, Kirkland, Québec H9H 5K5 (CA). **GAZDA, Robert G.**; 28 Highgate Circle, Spring City, Pennsylvania 19475 (US). **DJISSITCHI, Steve Alphonse Siani**; 8570 Avenue Champagneur APP. 3, Montréal, Québec H3N 2L2 (CA).

(74) Agent: **STECK, Jeffrey Alan**; Invention Mine LLC, 216 South Jefferson Street, Suite 102, Chicago, Illinois 60661 (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, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JO, JP, KE, KG, KH, KN, KP, KR, KW, 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).

(54) Title: SYSTEMS AND METHODS TO CREATE SLICES AT A CELL EDGE TO PROVIDE COMPUTING SERVICES

(57) Abstract: Systems and methods are described for creating slices at a cell edge to provide computing services. One embodiment takes the form of a method comprising: an edge cloud slice (ECS) selection function operating to dynamically map network statistics of a plurality of small footprint devices (SFDs) at a far edge; the ECS selection function receiving an application request from a user device; the ECS selection function generating a ECS definition comprising a subset of the SFDs from the plurality of SFDs; the ECS selection function transmitting the ECS definition to a platform manager, and responsive to the platform manager receiving the ECS definition, causing instantiation of the requested application on the subset of the SFDs based on the ECS definition.

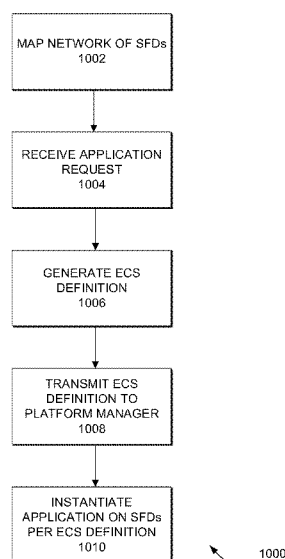


FIG. 10

Published:

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

SYSTEMS AND METHODS TO CREATE SLICES AT A CELL EDGE TO PROVIDE COMPUTING SERVICES

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] The present application is a non-provisional filing of, and claims benefit under 35 U.S.C. §119(e) from, U.S. Provisional Patent Application Serial No. 62/419,874 entitled "SYSTEMS AND METHODS TO CREATE SLICES AT A CELL EDGE TO PROVIDE COMPUTING SERVICES," filed November 9, 2016, which is hereby incorporated by reference in its entirety.

BACKGROUND

[0002] With the advent of revolutionary technological developments in the field of electronics, Internet-of-Things (IoT) landscape is rapidly evolving. The increased use of IoT is creating an environment of everyday object being communicatively coupled to each other. These connected devices include sensors, appliances, cars, real-time devices, etc. Modifications and changes to traditional wireless networks may improve operations of connected devices. One use of the 4G network was to support multimedia streaming and infotainment. Some of these modifications occur in the development of 5G wireless technology.

[0003] Considering this evolution to connected devices, a traditional cellular system, such as 4G for communication, and a typical data center based cloud infrastructure for applications would not be optimal.

[0004] 5G wireless networks are currently under development with the primary objective of establishing a unified connectivity framework that extends the capabilities of Human Type Communication (HTC) and thereby allows the interconnection of Machine Type Communication (MTC) from machines such as vehicles, robots, small IoT sensors and actuators, and other industrial equipment. This unified framework is expected to enable future industry-driven applications by supporting HTC and industry-grade MTC traffic of mixed priorities. Although there is uncertainty about what the final 5G framework will be, changes are expected to address latency, proximity services, and context awareness.

SUMMARY

[0005] Described herein are systems and methods to create slices at a cell edge to provide computing services. In some embodiments, slices are created via a far edge cloud, called Edge Cloud Slice (ECS) herein. The ECS may be either dedicated for an application or shared among applications. An exemplary

ECS includes a set of small footprint devices (SFDs) available for hosting application instances. A far edge cloud management platform, based on live measurements and policies from the application provider and cloud operator, will trigger the creation, expansion, shrinking, and deletion of an ECS.

[0006] One embodiment takes the form of a method, the method comprising: an edge cloud slice (ECS) selection function mapping network statistics of a plurality of small footprint devices (SFDs) at a far edge; the ECS selection function receiving an application request from a user device; the ECS selection function generating a ECS definition comprising a subset of the SFDs from the plurality of SFDs; the ECS selection function transmitting the ECS definition to a platform manager, and responsive to the platform manager receiving the ECS definition, causing instantiation of the requested application on the subset of the SFDs based on the ECS definition.

[0007] A method in an exemplary embodiment includes maintaining a dynamic map of far-edge network nodes, wherein the map stores information on the location, computing capacity, and available storage of each of the nodes. A resource request is received, wherein the resource request identifies at least an application and a location, and in response to the resource request, a set of computing resources for the application is identified. A group of network nodes is selected based at least on the location and set of computing requirements, and the selected group of network nodes is caused to instantiate the identified application.

BRIEF DESCRIPTION OF THE DRAWINGS

[0008] FIG. 1A is a system diagram illustrating an example communications system in which one or more disclosed embodiments may be implemented;

[0009] FIG. 1B is a system diagram illustrating an example wireless transmit/receive unit (WTRU) that may be used within the communications system illustrated in FIG. 1A according to an embodiment;

[0010] FIG. 1C is a system diagram illustrating an example radio access network (RAN) and an example core network (CN) that may be used within the communications system illustrated in FIG. 1A according to an embodiment;

[0011] FIG. 1D is a system diagram illustrating a further example RAN and a further example CN that may be used within the communications system illustrated in FIG. 1A according to an embodiment;

[0012] FIG. 2A depicts a network comprising a far edge cloud, in accordance with an embodiment.

[0013] FIG. 2B depicts a fog computing system, in accordance with an embodiment.

[0014] FIG. 3 depicts a far edge cloud as an extension to ETSI MEC, in accordance with an embodiment.

[0015] FIG. 4 depicts an example network comprising a network topology manager and a cloud network map manager, in accordance with an embodiment.

[0016] FIG. 5 depicts an ETSI MEC architecture comprising an SFD selection function, in accordance with an embodiment.

[0017] FIG. 6A depicts a system in an initial state of a first use case, in accordance with an embodiment.

[0018] FIG. 6B depicts the system of FIG. 6A in a final state of the first use case, in accordance with an embodiment.

[0019] FIG. 7 depicts an edge cloud slice creation procedure, in accordance with some embodiments.

[0020] FIG. 8 depicts a Docker system architecture, in accordance with an embodiment.

[0021] FIG. 9 depicts an example architecture used to create an ECS using Docker, in accordance with an embodiment.

[0022] FIG. 10 depicts an example method, in accordance with an embodiment.

EXAMPLE NETWORKS FOR IMPLEMENTATION OF THE EMBODIMENTS

[0023] FIG. 1A is a diagram illustrating an example communications system 100 in which one or more disclosed embodiments may be implemented. The communications system 100 may be a multiple access system that provides content, such as voice, data, video, messaging, broadcast, etc., to multiple wireless users. The communications system 100 may enable multiple wireless users to access such content through the sharing of system resources, including wireless bandwidth. For example, the communications systems 100 may employ one or more channel access methods, such as code division multiple access (CDMA), time division multiple access (TDMA), frequency division multiple access (FDMA), orthogonal FDMA (OFDMA), single-carrier FDMA (SC-FDMA), zero-tail unique-word DFT-Spread OFDM (ZT UW DTS-s OFDM), unique word OFDM (UW-OFDM), resource block-filtered OFDM, filter bank multicarrier (FBMC), and the like.

[0024] As shown in FIG. 1A, the communications system 100 may include wireless transmit/receive units (WTRUs) 102a, 102b, 102c, 102d, a RAN 104/113, a CN 106/115, a public switched telephone network (PSTN) 108, the Internet 110, and other networks 112, though it will be appreciated that the disclosed embodiments contemplate any number of WTRUs, base stations, networks, and/or network elements. Each of the WTRUs 102a, 102b, 102c, 102d may be any type of device configured to operate and/or communicate in a wireless environment. By way of example, the WTRUs 102a, 102b, 102c, 102d, any of which may be referred to as a “station” and/or a “STA”, may be configured to transmit and/or receive wireless signals and may include a user equipment (UE), a mobile station, a fixed or mobile subscriber unit, a subscription-based unit, a pager, a cellular telephone, a personal digital assistant (PDA), a smartphone, a

laptop, a netbook, a personal computer, a wireless sensor, a hotspot or Mi-Fi device, an Internet of Things (IoT) device, a watch or other wearable, a head-mounted display (HMD), a vehicle, a drone, a medical device and applications (e.g., remote surgery), an industrial device and applications (e.g., a robot and/or other wireless devices operating in an industrial and/or an automated processing chain contexts), a consumer electronics device, a device operating on commercial and/or industrial wireless networks, and the like. Any of the WTRUs 102a, 102b, 102c and 102d may be interchangeably referred to as a UE.

[0025] The communications systems 100 may also include a base station 114a and/or a base station 114b. Each of the base stations 114a, 114b may be any type of device configured to wirelessly interface with at least one of the WTRUs 102a, 102b, 102c, 102d to facilitate access to one or more communication networks, such as the CN 106/115, the Internet 110, and/or the other networks 112. By way of example, the base stations 114a, 114b may be a base transceiver station (BTS), a Node-B, an eNode B, a Home Node B, a Home eNode B, a gNB, a NR NodeB, a site controller, an access point (AP), a wireless router, and the like. While the base stations 114a, 114b are each depicted as a single element, it will be appreciated that the base stations 114a, 114b may include any number of interconnected base stations and/or network elements.

[0026] The base station 114a may be part of the RAN 104/113, which may also include other base stations and/or network elements (not shown), such as a base station controller (BSC), a radio network controller (RNC), relay nodes, etc. The base station 114a and/or the base station 114b may be configured to transmit and/or receive wireless signals on one or more carrier frequencies, which may be referred to as a cell (not shown). These frequencies may be in licensed spectrum, unlicensed spectrum, or a combination of licensed and unlicensed spectrum. A cell may provide coverage for a wireless service to a specific geographical area that may be relatively fixed or that may change over time. The cell may further be divided into cell sectors. For example, the cell associated with the base station 114a may be divided into three sectors. Thus, in one embodiment, the base station 114a may include three transceivers, i.e., one for each sector of the cell. In an embodiment, the base station 114a may employ multiple-input multiple output (MIMO) technology and may utilize multiple transceivers for each sector of the cell. For example, beamforming may be used to transmit and/or receive signals in desired spatial directions.

[0027] The base stations 114a, 114b may communicate with one or more of the WTRUs 102a, 102b, 102c, 102d over an air interface 116, which may be any suitable wireless communication link (e.g., radio frequency (RF), microwave, centimeter wave, micrometer wave, infrared (IR), ultraviolet (UV), visible light, etc.). The air interface 116 may be established using any suitable radio access technology (RAT).

[0028] More specifically, as noted above, the communications system 100 may be a multiple access system and may employ one or more channel access schemes, such as CDMA, TDMA, FDMA, OFDMA,

SC-FDMA, and the like. For example, the base station 114a in the RAN 104/113 and the WTRUs 102a, 102b, 102c may implement a radio technology such as Universal Mobile Telecommunications System (UMTS) Terrestrial Radio Access (UTRA), which may establish the air interface 115/116/117 using wideband CDMA (WCDMA). WCDMA may include communication protocols such as High-Speed Packet Access (HSPA) and/or Evolved HSPA (HSPA+). HSPA may include High-Speed Downlink (DL) Packet Access (HSDPA) and/or High-Speed UL Packet Access (HSUPA).

[0029] In an embodiment, the base station 114a and the WTRUs 102a, 102b, 102c may implement a radio technology such as Evolved UMTS Terrestrial Radio Access (E-UTRA), which may establish the air interface 116 using Long Term Evolution (LTE) and/or LTE-Advanced (LTE-A) and/or LTE-Advanced Pro (LTE-A Pro).

[0030] In an embodiment, the base station 114a and the WTRUs 102a, 102b, 102c may implement a radio technology such as NR Radio Access, which may establish the air interface 116 using New Radio (NR).

[0031] In an embodiment, the base station 114a and the WTRUs 102a, 102b, 102c may implement multiple radio access technologies. For example, the base station 114a and the WTRUs 102a, 102b, 102c may implement LTE radio access and NR radio access together, for instance using dual connectivity (DC) principles. Thus, the air interface utilized by WTRUs 102a, 102b, 102c may be characterized by multiple types of radio access technologies and/or transmissions sent to/from multiple types of base stations (e.g., a eNB and a gNB).

[0032] In other embodiments, the base station 114a and the WTRUs 102a, 102b, 102c may implement radio technologies such as IEEE 802.11 (i.e., Wireless Fidelity (WiFi)), IEEE 802.16 (i.e., Worldwide Interoperability for Microwave Access (WiMAX)), CDMA2000, CDMA2000 1X, CDMA2000 EV-DO, Interim Standard 2000 (IS-2000), Interim Standard 95 (IS-95), Interim Standard 856 (IS-856), Global System for Mobile communications (GSM), Enhanced Data rates for GSM Evolution (EDGE), GSM EDGE (GERAN), and the like.

[0033] The base station 114b in FIG. 1A may be a wireless router, Home Node B, Home eNode B, or access point, for example, and may utilize any suitable RAT for facilitating wireless connectivity in a localized area, such as a place of business, a home, a vehicle, a campus, an industrial facility, an air corridor (e.g., for use by drones), a roadway, and the like. In one embodiment, the base station 114b and the WTRUs 102c, 102d may implement a radio technology such as IEEE 802.11 to establish a wireless local area network (WLAN). In an embodiment, the base station 114b and the WTRUs 102c, 102d may implement a radio technology such as IEEE 802.15 to establish a wireless personal area network (WPAN). In yet another embodiment, the base station 114b and the WTRUs 102c, 102d may utilize a cellular-based

RAT (e.g., WCDMA, CDMA2000, GSM, LTE, LTE-A, LTE-A Pro, NR etc.) to establish a picocell or femtocell. As shown in FIG. 1A, the base station 114b may have a direct connection to the Internet 110. Thus, the base station 114b may not be required to access the Internet 110 via the CN 106/115.

[0034] The RAN 104/113 may be in communication with the CN 106/115, which may be any type of network configured to provide voice, data, applications, and/or voice over internet protocol (VoIP) services to one or more of the WTRUs 102a, 102b, 102c, 102d. The data may have varying quality of service (QoS) requirements, such as differing throughput requirements, latency requirements, error tolerance requirements, reliability requirements, data throughput requirements, mobility requirements, and the like. The CN 106/115 may provide call control, billing services, mobile location-based services, pre-paid calling, Internet connectivity, video distribution, etc., and/or perform high-level security functions, such as user authentication. Although not shown in FIG. 1A, it will be appreciated that the RAN 104/113 and/or the CN 106/115 may be in direct or indirect communication with other RANs that employ the same RAT as the RAN 104/113 or a different RAT. For example, in addition to being connected to the RAN 104/113, which may be utilizing a NR radio technology, the CN 106/115 may also be in communication with another RAN (not shown) employing a GSM, UMTS, CDMA 2000, WiMAX, E-UTRA, or WiFi radio technology.

[0035] The CN 106/115 may also serve as a gateway for the WTRUs 102a, 102b, 102c, 102d to access the PSTN 108, the Internet 110, and/or the other networks 112. The PSTN 108 may include circuit-switched telephone networks that provide plain old telephone service (POTS). The Internet 110 may include a global system of interconnected computer networks and devices that use common communication protocols, such as the transmission control protocol (TCP), user datagram protocol (UDP) and/or the internet protocol (IP) in the TCP/IP internet protocol suite. The networks 112 may include wired and/or wireless communications networks owned and/or operated by other service providers. For example, the networks 112 may include another CN connected to one or more RANs, which may employ the same RAT as the RAN 104/113 or a different RAT.

[0036] Some or all of the WTRUs 102a, 102b, 102c, 102d in the communications system 100 may include multi-mode capabilities (e.g., the WTRUs 102a, 102b, 102c, 102d may include multiple transceivers for communicating with different wireless networks over different wireless links). For example, the WTRU 102c shown in FIG. 1A may be configured to communicate with the base station 114a, which may employ a cellular-based radio technology, and with the base station 114b, which may employ an IEEE 802 radio technology.

[0037] FIG. 1B is a system diagram illustrating an example WTRU 102. As shown in FIG. 1B, the WTRU 102 may include a processor 118, a transceiver 120, a transmit/receive element 122, a speaker/microphone 124, a keypad 126, a display/touchpad 128, non-removable memory 130, removable

memory 132, a power source 134, a global positioning system (GPS) chipset 136, and/or other peripherals 138, among others. It will be appreciated that the WTRU 102 may include any sub-combination of the foregoing elements while remaining consistent with an embodiment.

[0038] The processor 118 may be a general purpose processor, a special purpose processor, a conventional processor, a digital signal processor (DSP), a plurality of microprocessors, one or more microprocessors in association with a DSP core, a controller, a microcontroller, Application Specific Integrated Circuits (ASICs), Field Programmable Gate Arrays (FPGAs) circuits, any other type of integrated circuit (IC), a state machine, and the like. The processor 118 may perform signal coding, data processing, power control, input/output processing, and/or any other functionality that enables the WTRU 102 to operate in a wireless environment. The processor 118 may be coupled to the transceiver 120, which may be coupled to the transmit/receive element 122. While FIG. 1B depicts the processor 118 and the transceiver 120 as separate components, it will be appreciated that the processor 118 and the transceiver 120 may be integrated together in an electronic package or chip.

[0039] The transmit/receive element 122 may be configured to transmit signals to, or receive signals from, a base station (e.g., the base station 114a) over the air interface 116. For example, in one embodiment, the transmit/receive element 122 may be an antenna configured to transmit and/or receive RF signals. In an embodiment, the transmit/receive element 122 may be an emitter/detector configured to transmit and/or receive IR, UV, or visible light signals, for example. In yet another embodiment, the transmit/receive element 122 may be configured to transmit and/or receive both RF and light signals. It will be appreciated that the transmit/receive element 122 may be configured to transmit and/or receive any combination of wireless signals.

[0040] Although the transmit/receive element 122 is depicted in FIG. 1B as a single element, the WTRU 102 may include any number of transmit/receive elements 122. More specifically, the WTRU 102 may employ MIMO technology. Thus, in one embodiment, the WTRU 102 may include two or more transmit/receive elements 122 (e.g., multiple antennas) for transmitting and receiving wireless signals over the air interface 116.

[0041] The transceiver 120 may be configured to modulate the signals that are to be transmitted by the transmit/receive element 122 and to demodulate the signals that are received by the transmit/receive element 122. As noted above, the WTRU 102 may have multi-mode capabilities. Thus, the transceiver 120 may include multiple transceivers for enabling the WTRU 102 to communicate via multiple RATs, such as NR and IEEE 802.11, for example.

[0042] The processor 118 of the WTRU 102 may be coupled to, and may receive user input data from, the speaker/microphone 124, the keypad 126, and/or the display/touchpad 128 (e.g., a liquid crystal display

(LCD) display unit or organic light-emitting diode (OLED) display unit). The processor 118 may also output user data to the speaker/microphone 124, the keypad 126, and/or the display/touchpad 128. In addition, the processor 118 may access information from, and store data in, any type of suitable memory, such as the non-removable memory 130 and/or the removable memory 132. The non-removable memory 130 may include random-access memory (RAM), read-only memory (ROM), a hard disk, or any other type of memory storage device. The removable memory 132 may include a subscriber identity module (SIM) card, a memory stick, a secure digital (SD) memory card, and the like. In other embodiments, the processor 118 may access information from, and store data in, memory that is not physically located on the WTRU 102, such as on a server or a home computer (not shown).

[0043] The processor 118 may receive power from the power source 134, and may be configured to distribute and/or control the power to the other components in the WTRU 102. The power source 134 may be any suitable device for powering the WTRU 102. For example, the power source 134 may include one or more dry cell batteries (e.g., nickel-cadmium (NiCd), nickel-zinc (NiZn), nickel metal hydride (NiMH), lithium-ion (Li-ion), etc.), solar cells, fuel cells, and the like.

[0044] The processor 118 may also be coupled to the GPS chipset 136, which may be configured to provide location information (e.g., longitude and latitude) regarding the current location of the WTRU 102. In addition to, or in lieu of, the information from the GPS chipset 136, the WTRU 102 may receive location information over the air interface 116 from a base station (e.g., base stations 114a, 114b) and/or determine its location based on the timing of the signals being received from two or more nearby base stations. It will be appreciated that the WTRU 102 may acquire location information by way of any suitable location-determination method while remaining consistent with an embodiment.

[0045] The processor 118 may further be coupled to other peripherals 138, which may include one or more software and/or hardware modules that provide additional features, functionality and/or wired or wireless connectivity. For example, the peripherals 138 may include an accelerometer, an e-compass, a satellite transceiver, a digital camera (for photographs and/or video), a universal serial bus (USB) port, a vibration device, a television transceiver, a hands free headset, a Bluetooth® module, a frequency modulated (FM) radio unit, a digital music player, a media player, a video game player module, an Internet browser, a Virtual Reality and/or Augmented Reality (VR/AR) device, an activity tracker, and the like. The peripherals 138 may include one or more sensors, the sensors may be one or more of a gyroscope, an accelerometer, a hall effect sensor, a magnetometer, an orientation sensor, a proximity sensor, a temperature sensor, a time sensor; a geolocation sensor; an altimeter, a light sensor, a touch sensor, a magnetometer, a barometer, a gesture sensor, a biometric sensor, and/or a humidity sensor.

[0046] The WTRU 102 may include a full duplex radio for which transmission and reception of some or all of the signals (e.g., associated with particular subframes for both the UL (e.g., for transmission) and downlink (e.g., for reception) may be concurrent and/or simultaneous. The full duplex radio may include an interference management unit to reduce and or substantially eliminate self-interference via either hardware (e.g., a choke) or signal processing via a processor (e.g., a separate processor (not shown) or via processor 118). In an embodiment, the WTRU 102 may include a half-duplex radio for which transmission and reception of some or all of the signals (e.g., associated with particular subframes for either the UL (e.g., for transmission) or the downlink (e.g., for reception)).

[0047] FIG. 1C is a system diagram illustrating the RAN 104 and the CN 106 according to an embodiment. As noted above, the RAN 104 may employ an E-UTRA radio technology to communicate with the WTRUs 102a, 102b, 102c over the air interface 116. The RAN 104 may also be in communication with the CN 106.

[0048] The RAN 104 may include eNode-Bs 160a, 160b, 160c, though it will be appreciated that the RAN 104 may include any number of eNode-Bs while remaining consistent with an embodiment. The eNode-Bs 160a, 160b, 160c may each include one or more transceivers for communicating with the WTRUs 102a, 102b, 102c over the air interface 116. In one embodiment, the eNode-Bs 160a, 160b, 160c may implement MIMO technology. Thus, the eNode-B 160a, for example, may use multiple antennas to transmit wireless signals to, and/or receive wireless signals from, the WTRU 102a.

[0049] Each of the eNode-Bs 160a, 160b, 160c may be associated with a particular cell (not shown) and may be configured to handle radio resource management decisions, handover decisions, scheduling of users in the UL and/or DL, and the like. As shown in FIG. 1C, the eNode-Bs 160a, 160b, 160c may communicate with one another over an X2 interface.

[0050] The CN 106 shown in FIG. 1C may include a mobility management entity (MME) 162, a serving gateway (SGW) 164, and a packet data network (PDN) gateway (or PGW) 166. While each of the foregoing elements are depicted as part of the CN 106, it will be appreciated that any of these elements may be owned and/or operated by an entity other than the CN operator.

[0051] The MME 162 may be connected to each of the eNode-Bs 162a, 162b, 162c in the RAN 104 via an S1 interface and may serve as a control node. For example, the MME 162 may be responsible for authenticating users of the WTRUs 102a, 102b, 102c, bearer activation/deactivation, selecting a particular serving gateway during an initial attach of the WTRUs 102a, 102b, 102c, and the like. The MME 162 may provide a control plane function for switching between the RAN 104 and other RANs (not shown) that employ other radio technologies, such as GSM and/or WCDMA.

[0052] The SGW 164 may be connected to each of the eNode Bs 160a, 160b, 160c in the RAN 104 via the S1 interface. The SGW 164 may generally route and forward user data packets to/from the WTRUs 102a, 102b, 102c. The SGW 164 may perform other functions, such as anchoring user planes during inter-eNode B handovers, triggering paging when DL data is available for the WTRUs 102a, 102b, 102c, managing and storing contexts of the WTRUs 102a, 102b, 102c, and the like.

[0053] The SGW 164 may be connected to the PGW 166, which may provide the WTRUs 102a, 102b, 102c with access to packet-switched networks, such as the Internet 110, to facilitate communications between the WTRUs 102a, 102b, 102c and IP-enabled devices.

[0054] The CN 106 may facilitate communications with other networks. For example, the CN 106 may provide the WTRUs 102a, 102b, 102c with access to circuit-switched networks, such as the PSTN 108, to facilitate communications between the WTRUs 102a, 102b, 102c and traditional land-line communications devices. For example, the CN 106 may include, or may communicate with, an IP gateway (e.g., an IP multimedia subsystem (IMS) server) that serves as an interface between the CN 106 and the PSTN 108. In addition, the CN 106 may provide the WTRUs 102a, 102b, 102c with access to the other networks 112, which may include other wired and/or wireless networks that are owned and/or operated by other service providers.

[0055] Although the WTRU is described in FIGS. 1A-1D as a wireless terminal, it is contemplated that in certain representative embodiments that such a terminal may use (e.g., temporarily or permanently) wired communication interfaces with the communication network.

[0056] In representative embodiments, the other network 112 may be a WLAN.

[0057] A WLAN in Infrastructure Basic Service Set (BSS) mode may have an Access Point (AP) for the BSS and one or more stations (STAs) associated with the AP. The AP may have an access or an interface to a Distribution System (DS) or another type of wired/wireless network that carries traffic in to and/or out of the BSS. Traffic to STAs that originates from outside the BSS may arrive through the AP and may be delivered to the STAs. Traffic originating from STAs to destinations outside the BSS may be sent to the AP to be delivered to respective destinations. Traffic between STAs within the BSS may be sent through the AP, for example, where the source STA may send traffic to the AP and the AP may deliver the traffic to the destination STA. The traffic between STAs within a BSS may be considered and/or referred to as peer-to-peer traffic. The peer-to-peer traffic may be sent between (e.g., directly between) the source and destination STAs with a direct link setup (DLS). In certain representative embodiments, the DLS may use an 802.11e DLS or an 802.11z tunneled DLS (TDLS). A WLAN using an Independent BSS (IBSS) mode may not have an AP, and the STAs (e.g., all of the STAs) within or using the IBSS may communicate

directly with each other. The IBSS mode of communication may sometimes be referred to herein as an “ad-hoc” mode of communication.

[0058] When using the 802.11ac infrastructure mode of operation or a similar mode of operations, the AP may transmit a beacon on a fixed channel, such as a primary channel. The primary channel may be a fixed width (e.g., 20 MHz wide bandwidth) or a dynamically set width via signaling. The primary channel may be the operating channel of the BSS and may be used by the STAs to establish a connection with the AP. In certain representative embodiments, Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) may be implemented, for example in 802.11 systems. For CSMA/CA, the STAs (e.g., every STA), including the AP, may sense the primary channel. If the primary channel is sensed/detected and/or determined to be busy by a particular STA, the particular STA may back off. One STA (e.g., only one station) may transmit at any given time in a given BSS.

[0059] High Throughput (HT) STAs may use a 40 MHz wide channel for communication, for example, via a combination of the primary 20 MHz channel with an adjacent or nonadjacent 20 MHz channel to form a 40 MHz wide channel.

[0060] Very High Throughput (VHT) STAs may support 20MHz, 40 MHz, 80 MHz, and/or 160 MHz wide channels. The 40 MHz, and/or 80 MHz, channels may be formed by combining contiguous 20 MHz channels. A 160 MHz channel may be formed by combining 8 contiguous 20 MHz channels, or by combining two non-contiguous 80 MHz channels, which may be referred to as an 80+80 configuration. For the 80+80 configuration, the data, after channel encoding, may be passed through a segment parser that may divide the data into two streams. Inverse Fast Fourier Transform (IFFT) processing, and time domain processing, may be done on each stream separately. The streams may be mapped on to the two 80 MHz channels, and the data may be transmitted by a transmitting STA. At the receiver of the receiving STA, the above described operation for the 80+80 configuration may be reversed, and the combined data may be sent to the Medium Access Control (MAC).

[0061] Sub 1 GHz modes of operation are supported by 802.11af and 802.11ah. The channel operating bandwidths, and carriers, are reduced in 802.11af and 802.11ah relative to those used in 802.11n, and 802.11ac. 802.11af supports 5 MHz, 10 MHz and 20 MHz bandwidths in the TV White Space (TVWS) spectrum, and 802.11ah supports 1 MHz, 2 MHz, 4 MHz, 8 MHz, and 16 MHz bandwidths using non-TVWS spectrum. According to a representative embodiment, 802.11ah may support Meter Type Control/Machine-Type Communications, such as MTC devices in a macro coverage area. MTC devices may have certain capabilities, for example, limited capabilities including support for (e.g., only support for) certain and/or limited bandwidths. The MTC devices may include a battery with a battery life above a threshold (e.g., to maintain a very long battery life).

[0062] WLAN systems, which may support multiple channels, and channel bandwidths, such as 802.11n, 802.11ac, 802.11af, and 802.11ah, include a channel which may be designated as the primary channel. The primary channel may have a bandwidth equal to the largest common operating bandwidth supported by all STAs in the BSS. The bandwidth of the primary channel may be set and/or limited by a STA, from among all STAs in operating in a BSS, which supports the smallest bandwidth operating mode. In the example of 802.11ah, the primary channel may be 1 MHz wide for STAs (e.g., MTC type devices) that support (e.g., only support) a 1 MHz mode, even if the AP, and other STAs in the BSS support 2 MHz, 4 MHz, 8 MHz, 16 MHz, and/or other channel bandwidth operating modes. Carrier sensing and/or Network Allocation Vector (NAV) settings may depend on the status of the primary channel. If the primary channel is busy, for example, due to a STA (which supports only a 1 MHz operating mode), transmitting to the AP, the entire available frequency bands may be considered busy even though a majority of the frequency bands remains idle and may be available.

[0063] In the United States, the available frequency bands, which may be used by 802.11ah, are from 902 MHz to 928 MHz. In Korea, the available frequency bands are from 917.5 MHz to 923.5 MHz. In Japan, the available frequency bands are from 916.5 MHz to 927.5 MHz. The total bandwidth available for 802.11ah is 6 MHz to 26 MHz depending on the country code.

[0064] FIG. 1D is a system diagram illustrating the RAN 113 and the CN 115 according to an embodiment. As noted above, the RAN 113 may employ an NR radio technology to communicate with the WTRUs 102a, 102b, 102c over the air interface 116. The RAN 113 may also be in communication with the CN 115.

[0065] The RAN 113 may include gNBs 180a, 180b, 180c, though it will be appreciated that the RAN 113 may include any number of gNBs while remaining consistent with an embodiment. The gNBs 180a, 180b, 180c may each include one or more transceivers for communicating with the WTRUs 102a, 102b, 102c over the air interface 116. In one embodiment, the gNBs 180a, 180b, 180c may implement MIMO technology. For example, gNBs 180a, 180b may utilize beamforming to transmit signals to and/or receive signals from the gNBs 180a, 180b, 180c. Thus, the gNB 180a, for example, may use multiple antennas to transmit wireless signals to, and/or receive wireless signals from, the WTRU 102a. In an embodiment, the gNBs 180a, 180b, 180c may implement carrier aggregation technology. For example, the gNB 180a may transmit multiple component carriers to the WTRU 102a (not shown). A subset of these component carriers may be on unlicensed spectrum while the remaining component carriers may be on licensed spectrum. In an embodiment, the gNBs 180a, 180b, 180c may implement Coordinated Multi-Point (CoMP) technology. For example, WTRU 102a may receive coordinated transmissions from gNB 180a and gNB 180b (and/or gNB 180c).

[0066] The WTRUs 102a, 102b, 102c may communicate with gNBs 180a, 180b, 180c using transmissions associated with a scalable numerology. For example, the OFDM symbol spacing and/or OFDM subcarrier spacing may vary for different transmissions, different cells, and/or different portions of the wireless transmission spectrum. The WTRUs 102a, 102b, 102c may communicate with gNBs 180a, 180b, 180c using subframe or transmission time intervals (TTIs) of various or scalable lengths (e.g., containing varying number of OFDM symbols and/or lasting varying lengths of absolute time).

[0067] The gNBs 180a, 180b, 180c may be configured to communicate with the WTRUs 102a, 102b, 102c in a standalone configuration and/or a non-standalone configuration. In the standalone configuration, WTRUs 102a, 102b, 102c may communicate with gNBs 180a, 180b, 180c without also accessing other RANs (e.g., such as eNode-Bs 160a, 160b, 160c). In the standalone configuration, WTRUs 102a, 102b, 102c may utilize one or more of gNBs 180a, 180b, 180c as a mobility anchor point. In the standalone configuration, WTRUs 102a, 102b, 102c may communicate with gNBs 180a, 180b, 180c using signals in an unlicensed band. In a non-standalone configuration WTRUs 102a, 102b, 102c may communicate with/connect to gNBs 180a, 180b, 180c while also communicating with/connecting to another RAN such as eNode-Bs 160a, 160b, 160c. For example, WTRUs 102a, 102b, 102c may implement DC principles to communicate with one or more gNBs 180a, 180b, 180c and one or more eNode-Bs 160a, 160b, 160c substantially simultaneously. In the non-standalone configuration, eNode-Bs 160a, 160b, 160c may serve as a mobility anchor for WTRUs 102a, 102b, 102c and gNBs 180a, 180b, 180c may provide additional coverage and/or throughput for servicing WTRUs 102a, 102b, 102c.

[0068] Each of the gNBs 180a, 180b, 180c may be associated with a particular cell (not shown) and may be configured to handle radio resource management decisions, handover decisions, scheduling of users in the UL and/or DL, support of network slicing, dual connectivity, interworking between NR and E-UTRA, routing of user plane data towards User Plane Function (UPF) 184a, 184b, routing of control plane information towards Access and Mobility Management Function (AMF) 182a, 182b and the like. As shown in FIG. 1D, the gNBs 180a, 180b, 180c may communicate with one another over an Xn interface.

[0069] The CN 115 shown in FIG. 1D may include at least one AMF 182a, 182b, at least one UPF 184a, 184b, at least one Session Management Function (SMF) 183a, 183b, and possibly a Data Network (DN) 185a, 185b. While each of the foregoing elements are depicted as part of the CN 115, it will be appreciated that any of these elements may be owned and/or operated by an entity other than the CN operator.

[0070] The AMF 182a, 182b may be connected to one or more of the gNBs 180a, 180b, 180c in the RAN 113 via an N2 interface and may serve as a control node. For example, the AMF 182a, 182b may be responsible for authenticating users of the WTRUs 102a, 102b, 102c, support for network slicing (e.g.,

handling of different PDU sessions with different requirements), selecting a particular SMF 183a, 183b, management of the registration area, termination of NAS signaling, mobility management, and the like. Network slicing may be used by the AMF 182a, 182b in order to customize CN support for WTRUs 102a, 102b, 102c based on the types of services being utilized WTRUs 102a, 102b, 102c. For example, different network slices may be established for different use cases such as services relying on ultra-reliable low latency (URLLC) access, services relying on enhanced massive mobile broadband (eMBB) access, services for machine type communication (MTC) access, and/or the like. The AMF 162 may provide a control plane function for switching between the RAN 113 and other RANs (not shown) that employ other radio technologies, such as LTE, LTE-A, LTE-A Pro, and/or non-3GPP access technologies such as WiFi.

[0071] The SMF 183a, 183b may be connected to an AMF 182a, 182b in the CN 115 via an N11 interface. The SMF 183a, 183b may also be connected to a UPF 184a, 184b in the CN 115 via an N4 interface. The SMF 183a, 183b may select and control the UPF 184a, 184b and configure the routing of traffic through the UPF 184a, 184b. The SMF 183a, 183b may perform other functions, such as managing and allocating UE IP address, managing PDU sessions, controlling policy enforcement and QoS, providing downlink data notifications, and the like. A PDU session type may be IP-based, non-IP based, Ethernet-based, and the like.

[0072] The UPF 184a, 184b may be connected to one or more of the gNBs 180a, 180b, 180c in the RAN 113 via an N3 interface, which may provide the WTRUs 102a, 102b, 102c with access to packet-switched networks, such as the Internet 110, to facilitate communications between the WTRUs 102a, 102b, 102c and IP-enabled devices. The UPF 184a, 184b may perform other functions, such as routing and forwarding packets, enforcing user plane policies, supporting multi-homed PDU sessions, handling user plane QoS, buffering downlink packets, providing mobility anchoring, and the like.

[0073] The CN 115 may facilitate communications with other networks. For example, the CN 115 may include, or may communicate with, an IP gateway (e.g., an IP multimedia subsystem (IMS) server) that serves as an interface between the CN 115 and the PSTN 108. In addition, the CN 115 may provide the WTRUs 102a, 102b, 102c with access to the other networks 112, which may include other wired and/or wireless networks that are owned and/or operated by other service providers. In one embodiment, the WTRUs 102a, 102b, 102c may be connected to a local Data Network (DN) 185a, 185b through the UPF 184a, 184b via the N3 interface to the UPF 184a, 184b and an N6 interface between the UPF 184a, 184b and the DN 185a, 185b.

[0074] In view of Figures 1A-1D, and the corresponding description of Figures 1A-1D, one or more, or all, of the functions described herein with regard to one or more of: WTRU 102a-d, Base Station 114a-b, eNode-B 160a-c, MME 162, SGW 164, PGW 166, gNB 180a-c, AMF 182a-b, UPF 184a-b, SMF 183a-b,

DN 185a-b, and/or any other device(s) described herein, may be performed by one or more emulation devices (not shown). The emulation devices may be one or more devices configured to emulate one or more, or all, of the functions described herein. For example, the emulation devices may be used to test other devices and/or to simulate network and/or WTRU functions.

[0075] The emulation devices may be designed to implement one or more tests of other devices in a lab environment and/or in an operator network environment. For example, the one or more emulation devices may perform the one or more, or all, functions while being fully or partially implemented and/or deployed as part of a wired and/or wireless communication network in order to test other devices within the communication network. The one or more emulation devices may perform the one or more, or all, functions while being temporarily implemented/deployed as part of a wired and/or wireless communication network. The emulation device may be directly coupled to another device for purposes of testing and/or may performing testing using over-the-air wireless communications.

[0076] The one or more emulation devices may perform the one or more, including all, functions while not being implemented/deployed as part of a wired and/or wireless communication network. For example, the emulation devices may be utilized in a testing scenario in a testing laboratory and/or a non-deployed (e.g., testing) wired and/or wireless communication network in order to implement testing of one or more components. The one or more emulation devices may be test equipment. Direct RF coupling and/or wireless communications via RF circuitry (e.g., which may include one or more antennas) may be used by the emulation devices to transmit and/or receive data.

DETAILED DESCRIPTION

[0077] Described herein are systems and methods to create slices at a cell edge to provide computing services. In some embodiments, the slices are created in the context of wireless networks and may be used to provide low latency, proximity services, and context awareness.

[0078] To facilitate low latency, breakthroughs in medium access and advanced waveform technologies combined with novel coding and modulation schemes are expected to provide 5G networks with transmission latencies of less than 1ms.

[0079] To facilitate proximity services, 5G systems enable devices to communicate directly with other devices in the proximity in a Device-to-Device (D2D) fashion through a direct local link.

[0080] To facilitate context awareness, the network under 5G is expected to be context aware. For any given device, the network is not only expected to be continuously aware of its individual location and features but is also expected to possess information regarding its surroundings and environment.

[0081] With the demands of higher bandwidth and lower latency applications, new concepts are being sought after to fulfill the challenging requirements of future mobile communications. Mobile Edge Computing (MEC) is an emerging technology which enables service and content providers to offer their applications and services on the edge of the network, rather than utilizing the core network. In other words, under MEC, application and service deployment is enabled through a cloud-like environment at the edge of the mobile network. This concept does not only reduce the latency, but also avoids congesting the backbone network by restricting traffic to the geographical location of the subscribers. Within the European Telecommunications Standards Institute (ETSI) MEC architecture, MEC Cloud resources are deployed in Mobile Operator managed data centers that are co-located with macro cell sites.

[0082] Small Cells, HeNB, Wi-Fi Access Points, Femtocells, and HetNet Gateways may be an integral part of 5G network (along with Macrocell). These categories of edge nodes may have surplus and unused computing resources, storage, and the like. These nodes may form a smaller cloud, a Far Edge Cloud, at the very edge and act as an extension to the MEC platform deployed at the Macrocell or the Distant Cloud.

[0083] A network may provide contextual information in addition to low latency communications. An application platform may exploit the contextual information provided by the network to provision an ad hoc real-time collaboration in a given geographical area. In general, Small cell, HeNB, AP, STB, and the like are categorized as Small Footprint Devices (SFD). In some embodiments, the SFDs may be any computing device located at the far edge of the network.

[0084] FIG. 2A depicts a network comprising a far edge cloud, in accordance with an embodiment. FIG. 2A depicts the network 200. The network 200 includes a user device (depicted as a smart phone), a far edge comprising having Small Cells and Access Points forming the far edge cloud, a macro cell and wireless cellular network where an ETSI defined MEC system may be deployed, and the internet where the distant cloud is deployed.

[0085] In accordance with some embodiments, the Far Edge Cloud is a cloud formed out of Small cells, Wi-Fi AP, HeNB, Set top boxes, HetNet Gateways, In-home Media Gateways, and the like. The Far Edge Cloud is the cloud formed at the far edge of the network outside of managed data centers, beyond what is being defined by ETSI MEC. The Far Edge Cloud may provide services independently or in collaboration with the MEC/Distant cloud. The resources available to the far edge cloud are generally limited in terms of computing power, storage, and network connectivity. However, being closest to the end user device may have the advantage of responding with lowest latency.

Fog Computing.

[0086] FIG. 2B depicts a fog computing system, in accordance with an embodiment. In particular, FIG. 2B depicts the system 250 that includes a cloud level at the top, a fog level in the middle, and a device

level at the bottom. The higher levels are associated with core functions and the lower levels are associated with edge functions. The cloud level is located at a more centralized location than the fog level, and the devices are located at a broader set of locations than the fog level.

[0087] In a fog computing system, services can be hosted at end devices such as set-top-boxes, access points, HeNB, and the like. The infrastructure of this fog computing allows applications to run as close as possible to sensed actionable and massive data, coming out of people, processes and things. Both the cloud level and the fog level provide data, computation, storage and application services to end-users. However, the fog level can be distinguished from the cloud level by its proximity to end-users, the dense geographical distribution, and its support for mobility.

[0088] As the fog level computing is implemented at the edge of the network, it provides low latency, location awareness, and improves quality-of-services (QoS) for streaming and real time applications. Examples include industrial automation, transportation, and networks of sensors and actuators. Moreover, this infrastructure supports heterogeneity, as the fog level devices include end-user devices, access points, edge routers and switches. The fog paradigm is well positioned for real-time big data analytics, supports densely distributed data collection points, and provides advantages in entertainment, advertising, personal computing and other applications.

Grid Computing.

[0089] Grid Computing employs a collection of computer resources from multiple locations to reach a common goal. A Grid operates as a distributed system with non-interactive workloads that involve many files. Coordinating applications on Grids comprises coordinating the flow of information across distributed computing resources and grid workflow systems that have been developed as a specialized form of a workflow management system and designed specifically to compose and execute a series of computational or data manipulation steps, or a workflow, in the Grid context.

[0090] Grid computing supports sharing of resources with a virtual organization. It supports flexible, secure, coordinated resource sharing among dynamic collections of individual and institutional compute resources and may be referred to as Virtual Organizations.

[0091] Features of virtual organizations may include one or more of the following:

- Coordinated resource sharing and problem solving in dynamic, multi-institutional virtual organizations.
- Highly flexible sharing relationships.
- Support Client-server, peer-to-peer and brokered interaction mechanism.
- Complex and high levels of control over how shared resources are used, including fine-grained access control, delegation, and application of local and global policies.

- Sharing of varied resources, ranging from programs, files, and data to computers, sensors, and networks.
- Diverse usage modes, ranging from single user to multi-user and from performance sensitive to cost-sensitive.
- Embracing issues of quality of service, scheduling, co-allocation and accounting.

Network Functions Virtualization Management and Organization (NFV MANO).

[0092] In NFV MANO, functions that are used in NFV orchestration may include one or more of the following:

- Service coordination and instantiation. The orchestration software communicates with the underlying NFV platform to instantiate a service, which means it creates the virtual instance of a service on the platform.
- Service chaining. Enables a service to be cloned and multiplied to scale for either a single customer or many customers.
- Scaling services: When more services are added, finding and managing sufficient resources to deliver the service.
- Service monitoring: Tracks the performance of the platform and resources to make sure they are adequate to provide for good service.

[0093] Resource orchestration is used to ensure there are adequate computing, storage, and network resources available to provide a network service. To meet that objective, the NFV Orchestrator (NFVO) can coordinate either with the virtualized infrastructure manager (VIM) or directly with NFV infrastructure (NFVI) resources, depending on the requirements. It can coordinate, authorize, release, and engage NFVI resources independently of any specific VIM. It also provides governance of virtual network function (VNF) instances sharing resources of the NFVI.

[0094] In some cases, it may be more effective to deploy NFV-based solutions across different points of presence (POPs) or within one POP but across multiple resources. Without NFV, this would not be feasible. But with NFV MANO, service providers can build in this capability using an NFVO, which has the ability to engage the VIMs directly through their northbound APIs instead of engaging with the NFVI resources directly. This eliminates the physical boundaries that may normally hinder such deployments.

[0095] To provide service orchestration, the NFV orchestrator creates end-to-end service among different VNFs, which may be managed by different VNFM with which the NFVO coordinates.

Network Map.

[0096] In some embodiments, network mapping is used to track the physical connectivity of a network. Using network mapping, an exemplary system discovers the devices on the network and their connectivity. Some techniques used for network mapping include approaches based on simple network management protocol (SNMP), active probing, and route analytics.

[0097] An exemplary SNMP based approach retrieves data from Router and Switch management information bases (MIBs) to build the network map. An exemplary active probing approach uses a series of traceroute-like probe packets to build the network map. An exemplary route analytics approach uses information from the routing protocols to build the network map.

[0098] Application-Layer Traffic Optimization (ALTO) services, e.g. as described in RFC 7285, provide network information (e.g., basic network location structure and preferences of network paths) with the goal of modifying network resource consumption patterns while maintaining or improving application performance. The basic information of ALTO is based on abstract maps of a network. It provides the knowledge of the underlying network topologies with network map and path cost structures.

[0099] ALTO service indicates preferences amongst network locations in the form of path costs. Path costs are generic costs and can be internally computed by a network provider per its own policy. For a given ALTO network map, an ALTO cost map defines path costs pairwise amongst the set of source and destination network locations defined by the Process Identifiers (PIDs) contained in the network map. Each path cost is the end-to-end cost when a unit of traffic goes from the source to the destination.

Virtualization.

[0100] Hardware virtualization or platform virtualization, refers to the creation of a virtual machine that acts like a standalone computer with an operating system. Software executed on these virtual machines is separated from the underlying hardware resources.

[0101] Software virtualization may be performed using one or more of the following techniques:

- Operating system-level virtualization: hosting of multiple virtualized environments within a single OS instance.
- Application virtualization and workspace virtualization: the hosting of individual applications in an environment separated from the underlying OS.
- Service virtualization: emulating the behavior of dependent (e.g., third-party, evolving, or not implemented) system components that are used to exercise an application under test (AUT) for development or testing purposes.

[0102] Memory virtualization may comprise aggregating random-access memory (RAM) resources from networked systems into a single memory pool. Virtual memory is giving an application program the impression that it has contiguous working memory, isolating it from the underlying physical memory implementation

[0103] Storage virtualization may comprise abstracting logical storage from physical storage.

[0104] Network virtualization may comprise the creation of a virtualized network addressing space within or across network subnets. Virtual private network (VPN), a network protocol that replaces the actual wire or another physical media in a network with an abstract layer, allowing a network to be created over the Internet.

Virtual Machines and Linux Containers.

[0105] A virtual machine (VM) is an emulation of a computer system. Virtual machines operate based on the computer architecture and functions of a real or hypothetical computer, and their implementations may involve specialized hardware, software, or a combination of both.

[0106] In some embodiments, Linux Containers (LXC) are used to provide operating system-level virtualization through a virtual environment that has its own process and network space, instead of creating a full-fledged virtual machine. LXC relies on the Linux kernel cgroups functionality that was released in version 2.6.24. It also relies on other kinds of namespace isolation functionality, which were developed and integrated into the mainline Linux kernel.

Overview of exemplary embodiments.

[0107] FIG. 3 depicts a far edge cloud as an extension to ETSI MEC according to an embodiment. FIG. 3 depicts the system 300 that includes a distant cloud 302, an ETSI orchestrator 304, a Far Edge Cloud (FEC) Orchestrator 306, a plurality of interconnected servers 308A-C, a plurality of interconnected devices 310A-F, an access point 312, a partition 314, and an area of improved computing resources 316.

[0108] In one embodiment, many small foot print computers, such as the devices 310A-F, are available around and within network point of attachment. These small computers may vary widely with respect to proximity, connectivity, and capability (e.g. the level of available computational power and storage). While providing edge computing services, one of the devices 310A-F may not provide desired performance. In FIG. 3, a user device is connected through the access point 312. To provide a computing service, the ideal computing resources are those shown within the dotted box 316. In an exemplary embodiment, the orchestrator is configured to load the service or application in devices 310E and 310D. These two devices thereby form a “cloud slice” and collaborate between them.

[0109] Virtualization platforms allow a level of control where application/service can be loaded. Data centers allow hosting applications on either collocated servers or remote servers, such as the servers 308A-C, for reasons such as load balancing, fault tolerance, resource utilization and the like. Data center computing resources are large and connected through a high capacity network. They are also fewer in number compared to computing resources at the far edge. The data center orchestrator (such as ETSI orchestrator 304) has the information related to the available servers 308A-C, such as where they are located and how much computing resources are available. Data centers are generally associated with locations covering large areas, such as at a regional or national level. The orchestrator in a data center is provisioned to run an application at Region A or Region B and may not change frequently.

[0110] Inside the data center, the orchestrator has information on the available servers and may move an application from one to another. Therefore, orchestration in a data center may not be dynamic, as the data center orchestrator does not need to make decisions regarding orchestration based on the exact location of servers. Since applications may not be as sensitive to latency, the precise location of resources may not be very important. But in the case of far edge, there may be many smaller resources with limited computing capacity. In exemplary embodiments, applications are hosted on the far edge to provide the lowest latency. In such embodiments, the far edge orchestrator (such as the SMC orchestrator 306) may have information at a very granular level regarding the locations of the resources, available computing resources, and network capacity. The environment is dynamic because users are moving, resources may disappear and reappear, and network conditions may change. Thus, at the far edge, an exemplary orchestrator may operate in a dynamic and autonomous mode, adjusting based on real-time conditions. The orchestrator collects information at a very granular level, such as available resource and network conditions at precise locations. Based on that information, the orchestrator determines where an application may be hosted so that it can provide the lowest latency to an application.

[0111] In NFV MANO, the orchestrator makes sure there are adequate computing, storage, and network resources available to provide a network service. NFV-based solutions are typically deployed across different points of presence (POPs) or within one POP but across multiple resources. This is similar to data center like deployment at POPs. The orchestrator manages smaller number of data centers, which may have a large computing capability and a high bandwidth network connection. The orchestrator may not need to know the location of the resources at a very granular level.

[0112] Compared to orchestration in a data center and NFV MANO framework, an orchestrator in the far edge may have one or more of the following characteristics: (1) capability of tracking resources at a very precise level of location granularity; (2) dynamic evaluation of resource availability; (3) capability of having detailed information on available resources at particular locations; (4) capability of having information on

available network capacity at particular locations; and (5) capability of using this information to identify the exact location where applications and services can be hosted.

[0113] In exemplary embodiments, Edge Cloud Slices (ECSs) may be used to enable different classes of applications to share the network and the available cloud resources of a far edge cloud. These ECSs may be used to host a single or many different applications: e.g. all real-time video streaming applications may be allocated a slice by the network operator, making it possible for other types of applications to get an appropriate amount of computing power in a different slice. Slices may be adjusted over time to adapt to actual usage.

[0114] In exemplary embodiments using ECSs for Far Edge Deployment, ECSs may be created dynamically based on user location, traffic load, and the like. Such ECSs may be dynamically added, expanded, shrunk, or deleted. SFDs may be selected for inclusion in ECSs. Instances on the ECSs may be loaded to provide services to users, wherein instances from single applications or multiple applications are loaded into same ECS, and the decision to use the same ECS may be made dynamically. ECSs may be managed automatically without any operator or administrator intervention.

[0115] In some embodiments, a subset of SFDs from a set of SFDs is chosen to form an ECS. In some embodiments, a portion of computing resources from each of a plurality of SFDs is chosen to form an ECS. For example, in a case with two small footprint devices, SFD1 and SFD2, a first ECS, ECS1, may be created using 50% of the resources from SFD1 and 50% of the resources from SFD2, and a second ECS, ECS2, may be created using the remaining resources in SFD1 together with the remaining resources in SFD2. In some embodiments, the need for an ECS reconfiguration is determined. In some embodiments, either one or both of the Cloud Management System or Orchestrator is informed of the selected SFDs. In some embodiments, the existing ECS is managed and may be determined whether it may be reused to load an application. In some embodiments, these steps are executed without manual intervention.

[0116] In some embodiments, an application instance is a specific realization of a program, or part of an application or service. A single application instance may provide a desired user function. Alternatively, a collection of one or more application instances (same or different types) may be combined to provide a desired user function. An application instance may support a single user or multiple users.

[0117] In some embodiments, small footprint devices (SFDs) include a physical machine, such as, AP, SMALL CELL, STB, HeNB, Gateway, and the like. Many such SFD may form a large cluster or far edge cloud.

[0118] In exemplary embodiments, Edge Cloud Slices (ECSs), sometimes referred to simply as a “slice” herein, may include a set of SFDs, or resources that are part of SFDs, which are made available to hold instances of multiple or singular applications. An orchestrator may then be used to distribute instances of

these applications among a slice. Setting up an ECS may include setting up other resources such as virtual networks for the use of applications.

[0119] In some embodiments, ECSs are organized such that each slice handles a particular application. In some embodiments, ECSs are organized such that each slice handles a particular group of applications. Applications may thereby share the available computing resources.

[0120] In some embodiments, an orchestrator identifies the correct ECS for a particular application and forwards the request to load application instance on the selected ECS.

[0121] In some embodiments, a platform manager executes the request from the orchestrator. For example, the platform manager creates the ECS and loads the application instance on the ECS.

[0122] FIG. 4 depicts an example network comprising a network topology manager and a cloud network map manager, in accordance with an embodiment. FIG. 4 depicts the network 400 that includes a network topology manager (NTM) 402, a cloud network map manager (CNMM) 404, an orchestrator 406, an application instance/container 408, a platform manager/master 410, devices 412A-F that may be SFDs, an access point 414, and a region 416.

[0123] The NTM 402 and the CNMM 404 may be included in a far edge deployment to support creation of and ECS to provide edge computing services.

[0124] In accordance with an embodiment, the CNMM 404 may perform the following functions. As small cells, AP, HeNB, and other SFDs are installed, autonomous reporting from each installed device about their location with respect to other access points, link cost to reach neighboring nodes, available computing and storage capacity, and the like is enabled. The CNMM 404 creates a network map with computing/storage capacity available within the SFDs and link cost of connectivity to each respective SFD. The SFDs periodically report their status in terms of load, resource availability, and the like. The CNMM 404 processes the received information and updates the cloud network map.

[0125] The NTM 402 receives application requirements from the orchestrator 406. The requirements may include an Application ID, user information (e.g., location, user ID), computing capacity requirements, storage requirements, latency requirements, throughput or network bandwidth requirements, and other required platform services.

[0126] The NTM 402 may also receive policy information from each application from the operator. The policy information may include priority of the application (e.g., if it is a paid application then it should be run on a host with the most available resources), restrictions associated with the application (e.g., maximum allowed storage, computing resources, permitted time of usage, restricted usages), and allowed features (e.g., session continuity, application relocation).

[0127] The NTM 402 processes the policy information and cloud network map information received from the CNMM 404. It then determines suitable SFDs, such as the devices 412A-C within the region 416, for an ECS creation. The NTM provides an ECS definition of the selected set of SFDs and portions of SFD resources (e.g., fixed amount of computing and storage on SFDs) to the orchestrator 406. The orchestrator 406 implements a new slice based on the received ECS definition and may further operate to perform functions such as adding computing resources to an existing slice, removing computing resources from a slice, or reusing an existing slice for new application instantiation.

[0128] The orchestrator 406 gets a list of SFDs from NTM 402. An entry in that list may include IP address, host name, storage and computing resources. If no storage and computing resource is included in an entry, that may be taken as a signal that all the resources in the host are available to be used by an application. The orchestrator 406 may pass that list to the platform manager 410 with additional information, such as an Application ID, which may be instantiated on a specific host. The platform manager 41 may operate using known platform management techniques to set up the SFDs to create a slice and instantiate applications. Any communication mechanism such as message bus, virtual network, and the like may be used to set up the connectivity within a slice.

[0129] In accordance with an embodiment, the CNMM 404 collects information from the SFDs and builds a representation that is made available through a cloud network map API.

[0130] In another embodiment, the NTM 402 uses the cloud network map and other input (e.g., policy/SLA from application provider and network operator) to determine new or updated ECS size and composition. It provides this information to the orchestrator 406, which sets up and/or modifies the ECS, and then the orchestrator 406 may create, delete, or move instances.

[0131] In another embodiment, external application managers may use the CNMM 404 to request their own slice from the edge cloud.

[0132] The CNMM and NTM functions, described herein, allow for selection of SFDs to create an ECS. These are two separate functions with clearly defined responsibilities. In some embodiments, an orchestration system uses both of these functions. In other embodiments, an orchestration system uses only one of these two functions. These functions together as may be referred to as SFD selection function. The SFD selection function may work as an independent service provider or may be integrated with an orchestrator function. Referring to ETSI MEC architecture or NFV MANO, the SFD selection function can be implemented as part of the Mobile Edge Orchestrator (MEO) or NFV Orchestrator (NFVO). It may be noted that the SFD selection function is not limited to be used with ETSI or NFV, it may be also be extended to other cloud orchestrators.

[0133] FIG. 5 depicts an ETSI MEC architecture comprising an SFD selection function, in accordance with an embodiment. FIG. 5 depicts the architecture 500 that includes a mobile edge orchestrator 502, a SFD selection function (NTM & CNMM) 504, an OSS 506, a user portal 508, a mobile edge platform manager 510, a mobile edge host having a mobile edge platform 512 and a virtualization infrastructure 514, and device interconnections, with connections supporting the SFD selection function connecting to the ETSI MEC architecture shown in dashed lines.

[0134] The SFD selection function 504 interfaces with mobile edge orchestrator 502 to receive an application instantiation request and provides a SFD description in response. Additionally, the SFD selection function 504 interfaces with the operation support system (OSS) 506 to obtain operator policy, rules, and the like. The SFD selection function 504 also interfaces with the mobile edge platform manager (MEPM) 510 to obtain frequent reports about the SFD. As an alternative, the mobile edge orchestrator 502 may already have information about the SFDs. If sufficient information, then this interface may be implemented between the mobile edge orchestrator 502 and the SFD selection function 504. These are logical interfaces, and they may be implemented between other functions/entities based on system architecture and function partition.

Example Use Case.

[0135] FIG. 6A depicts a system in an initial state of a first use case, in accordance with an embodiment. FIG. 6A depicts the system 600 that includes a SFD selection function 602, an orchestrator 604, a portal 606, a platform manager 608, devices 610A-F, an application provider (APP1) 612, an access point 614, and a mobile device 618.

[0136] In one example use case, a user (USER1) associated with the mobile device 618 moves to a small cell coverage area and connects to an application (APP1) in the internet. The application APP1 is allocated resources such as App1_Compute and App1_Storage. The application provider 612 determines that it may serve the user better if the applications can be run at the Cell Edge Platform (of devices 610A-F). In response to this determination, in step 651, the application provider 612 uses the Portal 606 to instantiate application APP1 for the user connected to a specific cell. The application provider 612 provides details about APP1 such as the sub applications it consists of, user location or Cell ID it is connected to.

[0137] In step 652, the portal 606 provides security information, a location where the application images are available, and the details about the application (e.g. sub components, location of user) to the orchestrator 604. The orchestrator 604 verifies the security information and checks the application's validity. In step 653, the orchestrator 604 provides the information to SFD selection function 602 and requests an ECS definition where the application and its sub components can be instantiated. The SFD

selection process 602 receives the request and provides a list of SFD (e.g., IP address of devices) and possibly the sub applications which may run on the SFD.

[0138] In step 654, the orchestrator 604 forwards the list of SFD and the Application/Sub application ID to platform manager 608. The platform manager 608 requests the platform service to instantiate applications at specific SFDs. Resources may be assigned such as computing resource App1_Compute at SFD 610E and storage resource App2_Storage at SFD 610D. The platform manager 608 sets up the connection rules and requests the platform service to connect these applications.

[0139] FIG. 6B depicts the system of FIG. 6A in a final state of the example use case, in accordance with an embodiment. Once Applications are instantiated, the user device 618 will be served locally by the application running at SFD 610D (App 1 storage portions) and SFD 610E (App 1 computing portions).

Cloud Network Map and Cost Map Creation.

[0140] In an exemplary embodiment, a cloud network map manager (CNMM) provides a network map service. This function may be implemented as an individual service in any edge cloud platform, which can be used by other applications.

[0141] The map service provides information including one or more of network map statistics and information, group definitions, cost maps providing costs between defined groups, and the map service may support queries. The network statistics may include computing and storage resources available on each type of SFD, latency of communications between SFDs, bandwidth available, and the like.

[0142] In some embodiments, a network map provides a full set of network location groupings defined by the map service and the endpoints contained within each grouping. In the case of Far Edge Cloud, the grouping may be done based on Point of Attachment (POA) or Anchor POA. A POA and SFDs in close proximity with it may form a group. The proximity may be defined in terms of hops, such as single hop, two hops, etc. These groups are labelled with an estimated value of available bandwidth, overall latency (processing plus roundtrip to/from the anchor POA). Each of these groups may be identified by an identifier, (e.g., assigned by the operator or calculated from characteristics of the POA as to be unique.)

[0143] In some embodiments, the group definitions include details about each SFD, such as computing capacity, available storage, and the like. An available processing capacity may be a term of standardized capacity slot, e.g. expressed as a set of {compute, storage, etc.} summarizing CPU, memory and local storage capacity, similar to "instance sizes" used in Cloud computing. (e.g., in Azure, A0 designates 1 core, 0.75GB RAM, 19GB disk and A4 designates 8 cores, 14GB RAM and 2,039GB disks.) In exemplary embodiments, edge cloud instance sizes are smaller than those used in traditional cloud computing.

[0144] In some embodiments, cost maps provide costs between defined groupings. A cost map indicates preferences amongst network locations in the form of path costs. Path costs are generic costs and can be internally computed by a network provider according to its own policy. A cost map defines path costs pairwise amongst the set of source and destination network group identified by an identifier. Each path cost is the end-to-end cost when a unit of traffic goes from the source to the destination group. Routing cost, which may be a part of path cost, conveys a generic measure for the cost of routing traffic from a source to a destination. A lower value indicates a higher preference for traffic to be sent from a source to a destination. (e.g., a query for routing cost may return a set of IP addresses with costs indicating a ranking of the IP addresses.)

[0145] In some embodiments, the map service provides an API or otherwise provides an interface for supporting queries. In a first example, a query includes a request that specifies a list of source network locations, (e.g., [Src_1, Src_2, ..., Src_m]), and a list of destination network locations, (e.g., [Dst_1, Dst_2, ..., Dst_n]). In response, the server in this example returns the path cost for each of the m*n communicating pairs (i.e., the pairs Src_1 → Dst_1, ..., Src_1 → Dst_n, ..., Src_m → Dst_1, ..., Src_m → Dst_n). The query may have a syntax such as the following:

```
GetPathCost { SourceList(1..N), DestinationList(1..N) };
```

[0146] In another example the map service API supports queries regarding available computing capacity (e.g., Compute Metric) between those network locations. The server will return computing capacity for each of the m*n communicating pairs (i.e., Src_1 → Dst_1, ..., Src_1 → Dst_n, ..., Src_m → Dst_1, ..., Src_m → Dst_n). For example, computing capacity from Src_1 to Dst_1 may be expressed as (A0:3, A1:1) to express that three “A0” instance size slots are available on the path and one “A1” slot is available on the path. The query may have a syntax such as the following:

```
GetComputeMetric { SourceList(1..N), DestinationList(1..N) };
```

[0147] A combination of “path cost” and “compute metric” may be used by other applications to make decisions such as choosing the node to start an application, move an application, and the like. The Network Topology Manager (NTM) function, described later, uses a CNMM function to determine suitable resources.

[0148] In accordance with an embodiment, a network map and a cost map may be created. During deployment, a system administrator may provision the CNMM with the small cell and SFD information, such as available computing and storage capacity, proximity with respect to POA. Alternatively, a standard reporting protocol may be invented among POA, SFD, and the CNMM to report proximity, computing capacity, live latency measurement, and the like. As POAs and SFDs are deployed, automatic reporting to CNMM is triggered.

Algorithm to choose SFD for ECS.

[0149] The ECS algorithm is provided with information such as AP or Small Cell ID with which the user is attached. The algorithm may obtain network operator policy such as, service level agreement between a third-party application and MNO. The SLA will determine how big or small of a cloud slice may be allocated to an application. The policy may include location specific policy such as in a specified geographic area if there is any restriction about storage, caching, and the like.

[0150] To choose the appropriate SFDs to be included in an ECS, the algorithm queries the MAP Service to obtain network information such as, "Path Cost", "Node topology", "Compute Metric", and the like. The algorithm seeks to select appropriate SFDs to reduce latency and power consumption, improve resource usage, or a combination of both. In one example, the algorithm may be described as a latency minimization problem. For each group returned by the MAP service, let *Routing Cost* = r , *Available Compute capacity* = c , *Available Storage Capacity* = s .

[0151] Latency is assumed to increase with r and may be defined as $f(r, c, s)$. The application has minimal requirements such as Required Compute capacity (cr) and Required Storage (sr). The algorithm consists in selecting the group with a minimal value of r while respecting $c > cr$ and $s > sr$.

ECS Management.

[0152] In an exemplary embodiment, once a slice is created on a request for an application, the ECS is assigned an identifier (ECS id). The ECS id is stored along with the associated SFDs, assigned computing resources and Application ID (1...N). Another database may be created for Application ID to resource requirement mapping.

[0153] In accordance with an embodiment, ECS management includes one or more of the following: Monitoring the validity / suitability of existing ECS; after receiving a request for application invocation, determining if an existing ECS can be used or a new slice needs to be created; and responsively updating the information model/database based on the information.

[0154] In some embodiments, monitoring the validity or suitability of an existing ECS comprises periodically querying the map service (at the CNMM) to determine the computing status and network status. Otherwise, the monitoring function may include subscribing for notification with the map service to be notified when certain thresholds (e.g., processor utilization percentage, available storage percentage) are reached. The information is collected for each ECS. The status information is compared with the application requirement. If the ECS is not able to support the application, then a determination may be made to expand the ECS by adding more computing resources (or storage resources, or both, based on the determination) which may involve adding a new SFD or new SFD resources in the ECS. The corresponding database may be updated.

Sharing of an ECS.

[0155] The monitoring function maintains updated status / resource availability for an existing ECS. In some embodiments, ECSs can be shared. For example, in some such embodiments, multiple ECSs may be created out of a group of nodes or SFDs, wherein ECSs may share computing resources. In some embodiments, multiple applications may share a single ECS.

[0156] If a new request for application instantiation is received, the management function may operate to determine whether an existing ECS can support the application. If so, then the ECS ID is forwarded to the orchestrator for instantiating the application. The same ECS can be shared by more than one application.

[0157] In another embodiment, the management function may determine that an existing ECS can support a new application by addition of computing resources. The additional resource may be a new SFD or partial compute resource from an SFD. Instead of creating a new ECS, it may be decided to use an existing ECS with additional resources from an SFD.

[0158] The sharing of ECSs by applications may be driven by policy from network operator. A group of applications with similar requirement in terms of computing power, storage, and latency may be instantiated in a first ECS. Applications with different requirement may be instantiated on second ECS. For example, video streaming and multimedia applications may be grouped into a single ECS to allow greater control and ease of management.

Edge Cloud Slice (ECS) Creation.

[0159] FIG. 7 depicts an edge cloud slice creation procedure, in accordance with some embodiments. An ECS is created between a far edge computing system (e.g., the orchestrator and platform manager) and ECS functions (e.g., NTM and CNMM). FIG. 7 depicts the procedure 700 that includes communications between a CNMM 702, an NTM 704, an orchestrator 706, and a platform manager 708.

[0160] The orchestrator 706 (alternatively the platform manager 708) may initiate the procedure with Network Topology Manager (NTM) 704 to setup the ECS. To create a new slice, the orchestrator 706 (alternatively the platform manager 708) may request Network Topology Manager (NTM) 704 with the Application ID, User ID, User Group, POAs used by the user or group of users. An example request syntax may be as follows.

`GetEcs{ ApplicationID, ApplicationDescriptor, AnchorPOA}.`

The CNMM 702 and NTM 704 operate in step 710 to determine a set of SFDs to provide the service and to provide a response to the orchestrator 706. The response may contain information such a `ECS_ID` , `ListOfResources {SFD1, SFD2, SFD3}` or `ListOfResources {SFD1:A0, SFD2:A4}` to indicate resource partition from an existing SFD.

[0161] The NTM 704 obtains network map and Cost map from CNMM 702. The NTM may request policy level information from a network operator. After processing all these information, the NTM may return an ECS_ID and a list of suitable SFDs, which are part of the edge cloud slice. If an existing ECS can be used to host the application, then only the ECS_ID is returned.

[0162] The orchestrator 706 may forward the list of the SFDs to the platform manager 708 or an infrastructure manager to instantiate the application on the ECS by using an API. An example syntax for instantiating an application may appear as follows.

InstantiateApplication {ECS_ID, ListOfSFDs, ApplicationID}.

Note that in this example, the absence of “ListOfSFDs” indicates the reuse of the ECS with the supplied ECS_ID.

[0163] The platform manager 708 may create or use an existing ECS to instantiate applications on the nodes, which are part of that ECS. Additional functions may include configuring a network path, updating traffic rules, setting up DNS rules, and the like.

[0164] Periodically, (e.g., a few hours of operation), if the monitoring function in NTM 704 makes a determination to modify an existing ECS (step 712), it sends a request to the orchestrator 706 to modify the ECS by evoking an API. An example syntax for modifying an API may be as follows.

ModifyEcs {ECS_ID, ListOfSFDsToAdd, ListOfSFDsToDelete}.

The monitoring may be based on variable such as application load, number of users, link cost, and the like. The NTM 704 may modify an ECS by adding or deleting SFDs, and providing the orchestrator 706 with a “modify ECS” request with the ECS ID and new list of SFDs. The orchestrator 706 may send this information to the platform manager 708. The platform manager 708 responsively updates the ECS in step 714 by adding or deleting indicated SFDs. The modification may require relocating the application, restoring the application state, redirecting user traffic, changing the DNS and traffic rules, and the like. It may also be desirable that the SFDs should not be required to reboot or restart, similar to hot swap.

[0165] The NTM 704 may operate to delete an ECS by sending an instruction with a syntax such as DeleteEcs{ECS_ID} to the orchestrator 706.

Container and Docker System Embodiments.

[0166] In some embodiments, the SFD selection function (e.g., the CNMM and NTM) is integrated into a readily available orchestration system with a Docker system, which orchestrates the containers. It is noted that these functions may work with other orchestration systems which may use VM or any other virtualization technique. A Docker system allows orchestration using containers. A Docker Swarm Master implements the orchestration functionality. A Docker Engine is the platform service, which may be installed

in the SFDs. Upon request from the Docker Swarm Master, a new ECS (e.g., network of SFDs) is created and applications are deployed.

[0167] FIG. 8 depicts a Docker system architecture, in accordance with an embodiment. In particular, FIG. 8 depicts the Docker system architecture 800 that includes a Docker client device in communication with a router in cluster A. Cluster A comprises a Docker Swarm Master with an instance of an application, and physical Docker nodes (nodes 1-3) each instances of applications provided by the services deployment in containers.

[0168] In a Docker System, a master node (e.g., a POP, Aggregation point) running “Docker Swarm Master” acts as the orchestrator and may also be referred to as a docker manager. In some embodiments, it may also run “Consul Docker Image”. Consul is a datacenter runtime that provides service discovery, configuration, and orchestration capabilities. The Docker Swarm Master or Docker Management instances may also be installed on other nodes, managing Physical nodes in a distributed fashion. The physical nodes managed by Docker Swarm Master is referred to as Docker Node.

[0169] The Docker client device 802, as shown in FIG. 8, acts as the user interface to the Docker System. An IT administrator may provision the Docker System using the Docker client device. The Docker client device provides the Docker Swarm Master 804 the list of physical nodes, which are included in the Cluster A 806. As disclosed herein, Cluster A acts as an Edge Cloud Slice - ECS. The Docker Swarm Master then creates a cluster 808 of Docker nodes a.k.a “Swarm” using those physical nodes. At this point, the physical Docker nodes are interconnected and form a cluster or a slice. Once the ECS is created, services can be deployed through the containers. When a new ECS is created, it is represented as a single node instance to applications and services trying to access the ECS.

[0170] FIG. 9 depicts an example architecture used to create an ECS using Docker, in accordance with an embodiment. In particular, FIG. 9 depicts the architecture 900 that includes an orchestrator 902, an SFD selection function 904 having an NTM 906 and a CNMM 908, Docker swarm master 910, an ECS 912 having a plurality of instances of applications, and client devices 914, 916, 918 associated with respective Users A, B, and C.

[0171] In the embodiment depicted in FIG. 9, the orchestration functions are split between two components. The orchestrator 902 interfaces with an external Customer Portal. The Docker swarm master 910 acts as the other part of the orchestrator and includes the platform manager function.

[0172] The SFD selection function includes at least two functions, the NTM and the CNMM. As shown, the NTM interfaces with orchestrator and Docker swarm master.

[0173] The orchestrator sends the “Get ECS” message to the NTM. The NTM uses “Create API” towards the Docker Swarm Master. The Create API acts as the “Response message” disclosed above with

FIG. 7. Alternatively, the NTM may use Docker API or Command Line Interface to implement the Create API. The Create API may use HTTP or other form of message-based communication. The NTM provides list of SFDs (Hostname) and Application Information (image, user) to the Docker swarm master for ECS creation.

[0174] The Docker swarm master manages task scheduling and allocates resources per container within a pool of hosts (SFDs). The Docker swarm master may add a new node/application instance to a specific machine in the cluster during run time.

[0175] Nodes can be added in an ECS with labels. Personalized labels may be used for more specific constraints. In some embodiments, affinity rules are defined, such as, instruct the system to add a particular image to a specific node, by using a set of constraints.

[0176] Note that in this embodiment, an ECS is formed by a full set of SFDs, such that a physical node is 100% in an ECS or not. Other embodiments may enable an SFD to be shared between ECS, e.g. providing 50% of its resources to one ECS, and 50% to another.

[0177] FIG. 10 depicts an example method, in accordance with an embodiment. In particular, FIG. 10 depicts the method 1000 that includes mapping a network of SFDs at 1002, receiving an application request at 1004, generating an ECS definition at 1006, and transmitting the ECS definition to a platform manager at 1008, and instantiating the application on SFDs per the ECS definition at 1010.

[0178] The method 1000 may be performed using the systems and methods disclosed herein. In some embodiments, an edge cloud slice (ECS) selection function maps network statistics (at 1002) of a plurality of small footprint devices (SFDs) at a far edge; the ECS selection function receives an application request (at 1004) from a user device; the ECS selection function generates a ECS definition (at 1006) comprising a subset of the SFDs from the plurality of SFDs; the ECS selection function transmits the ECS definition to a platform manager (at 1008), and responsive to the platform manager receiving the ECS definition, instantiating the requested application on the subset of the SFDs based on the ECS definition (at 1010).

Alternative embodiments.

[0179] One exemplary embodiment is a method of creating an edge cloud slice. In the method, a cloud network map manager maps network statistics of a plurality of small footprint devices (SFDs) at a far edge. A network topology manager receives an application request from a user device. An edge cloud slice (ECS) definition is generated comprising a subset of the SFDs from the plurality of SFDs. The ECS definition is transmitted to a platform manager, and the platform manager responsively causing instantiation of the requested application on the subset of the SFDs based on the ECS definition.

[0180] In another embodiment of a method of creating an edge cloud slice, an edge cloud slice (ECS) selection function maps network statistics (e.g. latency, available computing resources, available storage resources, available bandwidth, and hops) of a plurality of small footprint devices (SFDs) at a far edge. The ECS selection function receives an application request from a user device. The ECS selection function generates an ECS definition comprising a subset of the SFDs from the plurality of SFDs. The ECS selection function transmits the ECS definition to a platform manager, and in response to the platform manager receiving the ECS definition, the platform manager causes instantiation of the requested application on the subset of the SFDs based on the ECS definition. In some embodiments, the ECS selection function comprises a cloud network map manager (CNMM) and a network topology manager (NTM). The CNMM maps network statistics of a plurality of SFDs at the far edge, and the NTM receives an application request from a user device. At least one of the SFDs may be selected from the group consisting of: a home e-node B, an access point, a set top box, a small cell, and a gateway computing device.

[0181] In some embodiments, instantiation of the requested application includes instantiating a first portion of the application's computations on a first SFD and a second portion of the application's computations on a second SFD, the first and second SFDs being in communication with each other. In some embodiments, instantiation of the application includes instantiating a first portion of the application's storage on a first SFD and a second portion of the application's storage on a second SFD, the first and second SFD being in communication with each other.

[0182] Note that various hardware elements of one or more of the described embodiments are referred to as "modules" that carry out (i.e., perform, execute, and the like) various functions that are described herein in connection with the respective modules. As used herein, a module includes hardware (e.g., one or more processors, one or more microprocessors, one or more microcontrollers, one or more microchips, one or more application-specific integrated circuits (ASICs), one or more field programmable gate arrays (FPGAs), one or more memory devices) deemed suitable by those of skill in the relevant art for a given implementation. Each described module may also include instructions executable for carrying out the one or more functions described as being carried out by the respective module, and it is noted that those instructions could take the form of or include hardware (i.e., hardwired) instructions, firmware instructions, software instructions, and/or the like, and may be stored in any suitable non-transitory computer-readable medium or media, such as commonly referred to as RAM, ROM, etc.

[0183] Although features and elements are described above in particular combinations, one of ordinary skill in the art will appreciate that each feature or element can be used alone or in any combination with the other features and elements. In addition, the methods described herein may be implemented in a computer program, software, or firmware incorporated in a computer-readable medium for execution by a

computer or processor. Examples of computer-readable storage media include, but are not limited to, a read only memory (ROM), a random access memory (RAM), a register, cache memory, semiconductor memory devices, magnetic media such as internal hard disks and removable disks, magneto-optical media, and optical media such as CD-ROM disks, and digital versatile disks (DVDs). A processor in association with software may be used to implement a radio frequency transceiver for use in a WTRU, UE, terminal, base station, RNC, or any host computer.

CLAIMS**What is Claimed:**

1. A method comprising:
 - maintaining a dynamic map of far-edge network nodes, wherein the map stores information on the location, computing capacity, and available storage of each of the nodes;
 - receiving a resource request, wherein the resource request identifies at least an application and a location;
 - identifying a set of computing resources for the application;
 - selecting a group of network nodes based at least on the location and set of computing requirements; and
 - causing the selected group of network nodes to instantiate the identified application.
2. The method of claim 1, wherein a plurality of the far-edge network nodes are small footprint devices.
3. The method of claim 1, wherein a plurality of the far-edge network nodes are access points.
4. The method of claim 1, wherein a plurality of the far-edge network nodes are edge routers.
5. The method of claim 1, wherein maintaining the dynamic map comprises building the dynamic map using simple network management protocol (SNMP) probe packets.
6. The method of claim 1, wherein maintaining the dynamic map comprises receiving autonomous reporting from the far-edge network nodes.
7. The method of claim 1, wherein maintaining the dynamic map comprises grouping network nodes into a plurality of groups based on proximity to respective points of attachment.
8. The method of claim 7, wherein maintaining the dynamic map further comprises storing a path cost for each of a plurality of pairs of groups.
9. The method of claim 1, wherein the far-edge network nodes comprise nodes within one hop of a point of attachment.
10. The method of claim 1, wherein the far-edge network nodes comprise nodes within two hops of a point of attachment.

11. The method of claim 1, wherein the set of computing requirements comprises at least one requirement selected from the group consisting of: computing capacity requirements, storage requirements, latency requirements, throughput requirements, and network bandwidth requirements.
12. The method of claim 1, wherein the selection of the group of network nodes is further based on available storage of each of the nodes.
13. The method of claim 1, further comprising organizing a plurality of the far-edge network nodes into a plurality of edge cloud slices, wherein selecting a group of network nodes comprises selecting an edge cloud slice.
14. The method of claim 13, wherein causing the selected group of network nodes to instantiate the identified application comprises loading an instance of the identified application on the selected edge cloud slice.
15. The method of claim 1, wherein the identified application includes at least two sub-applications, and wherein causing the selected group of network nodes to instantiate the identified application comprises causing the sub-applications to be instantiated on different nodes in the selected group.

100

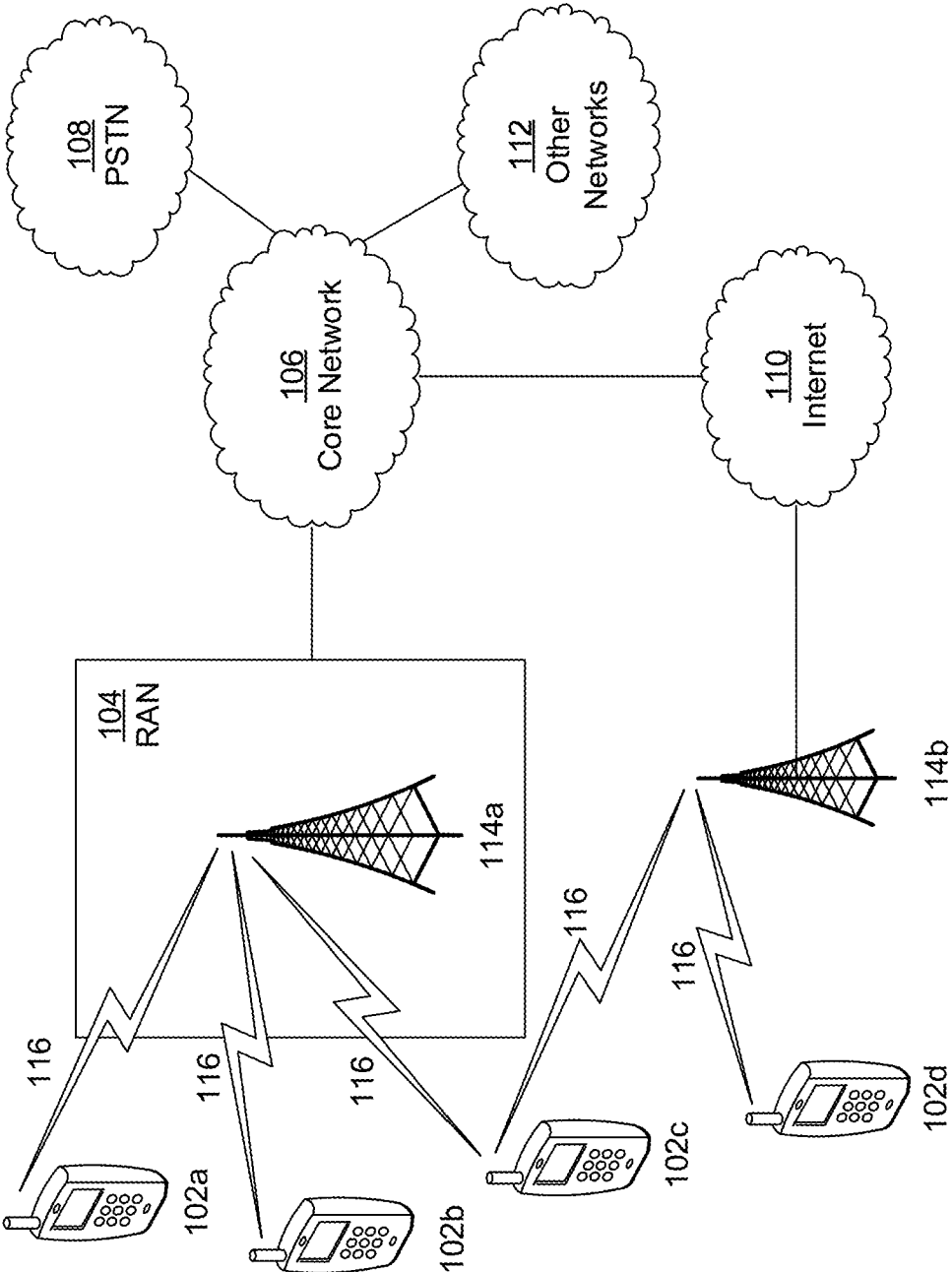
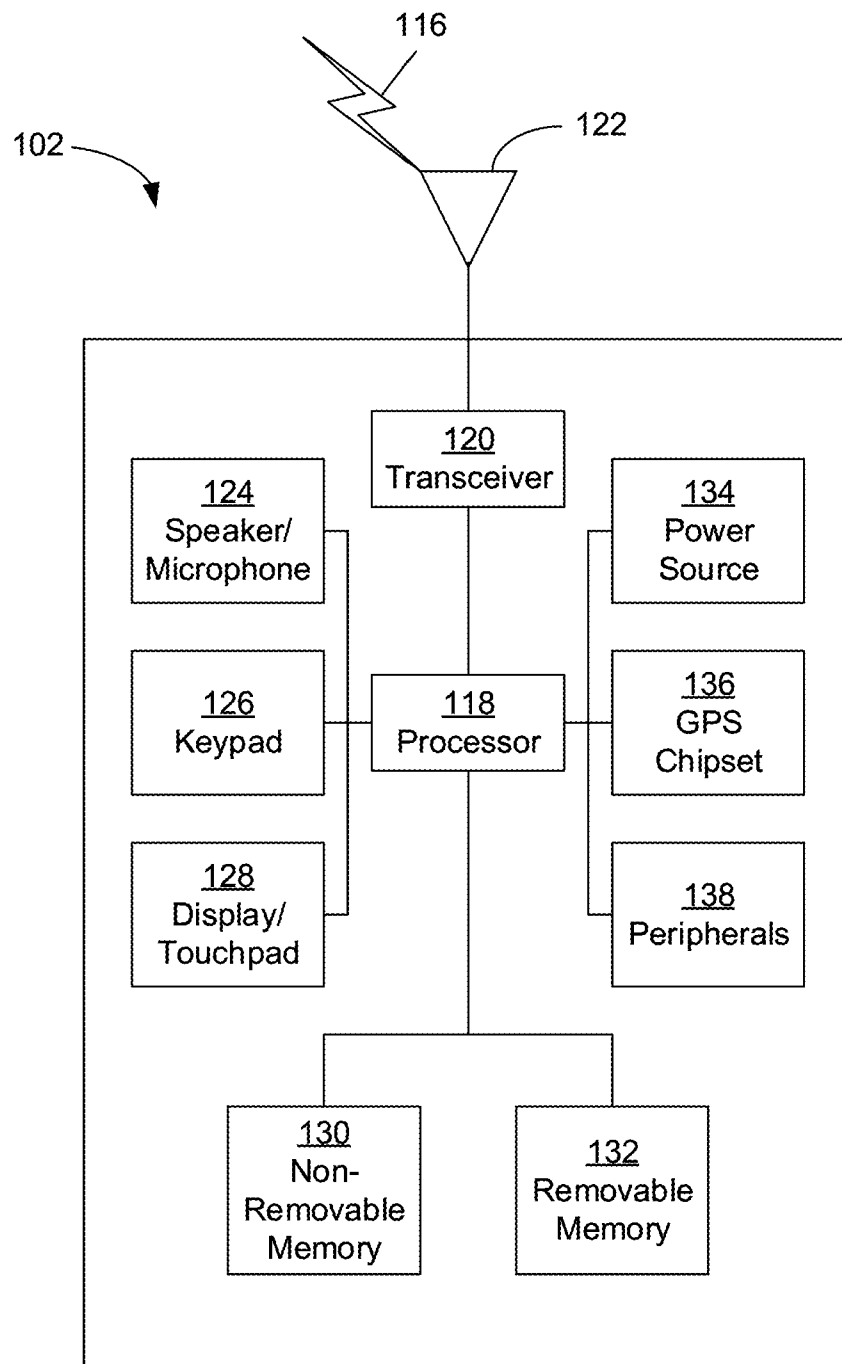


FIG. 1A

**FIG. 1B**

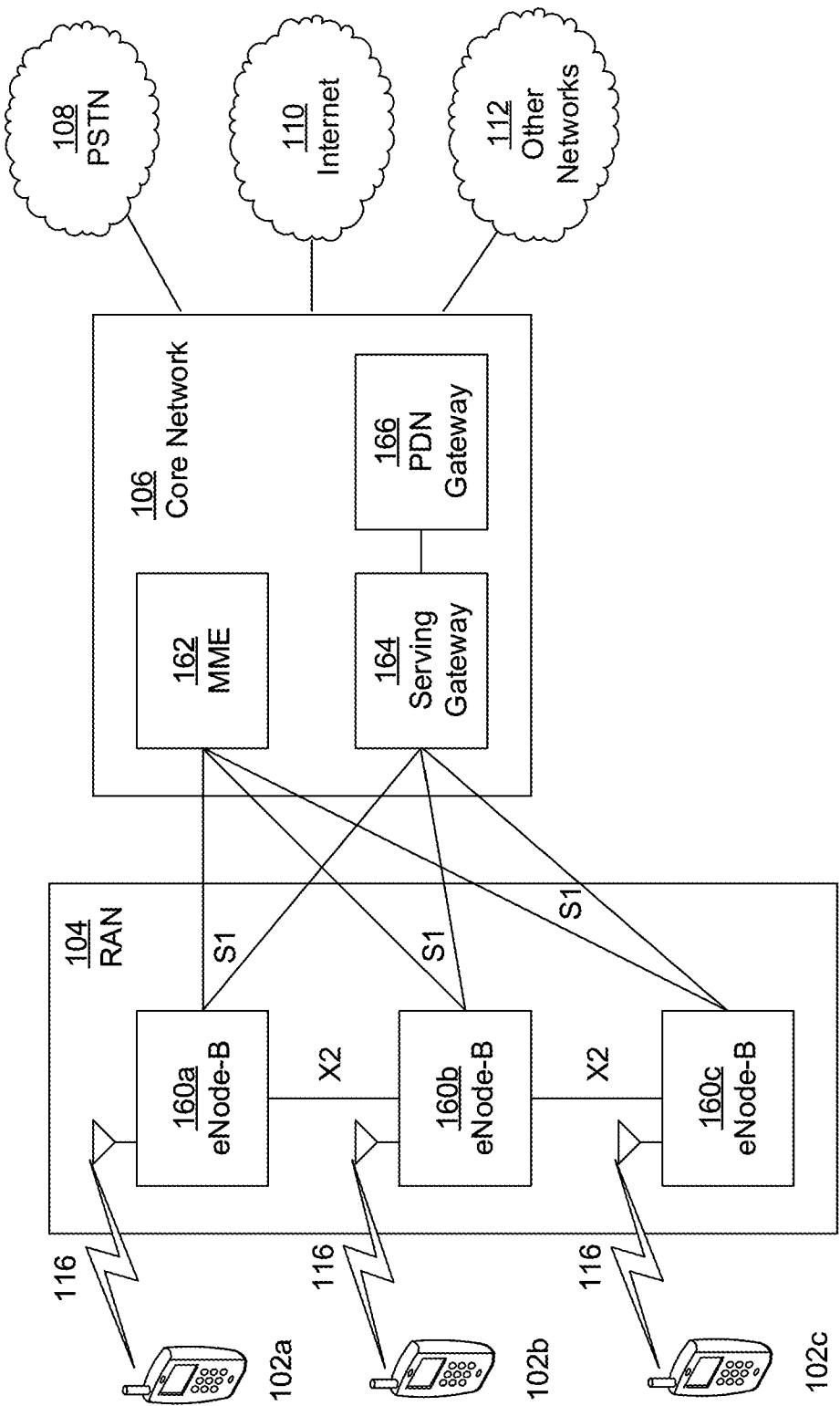


FIG. 1C

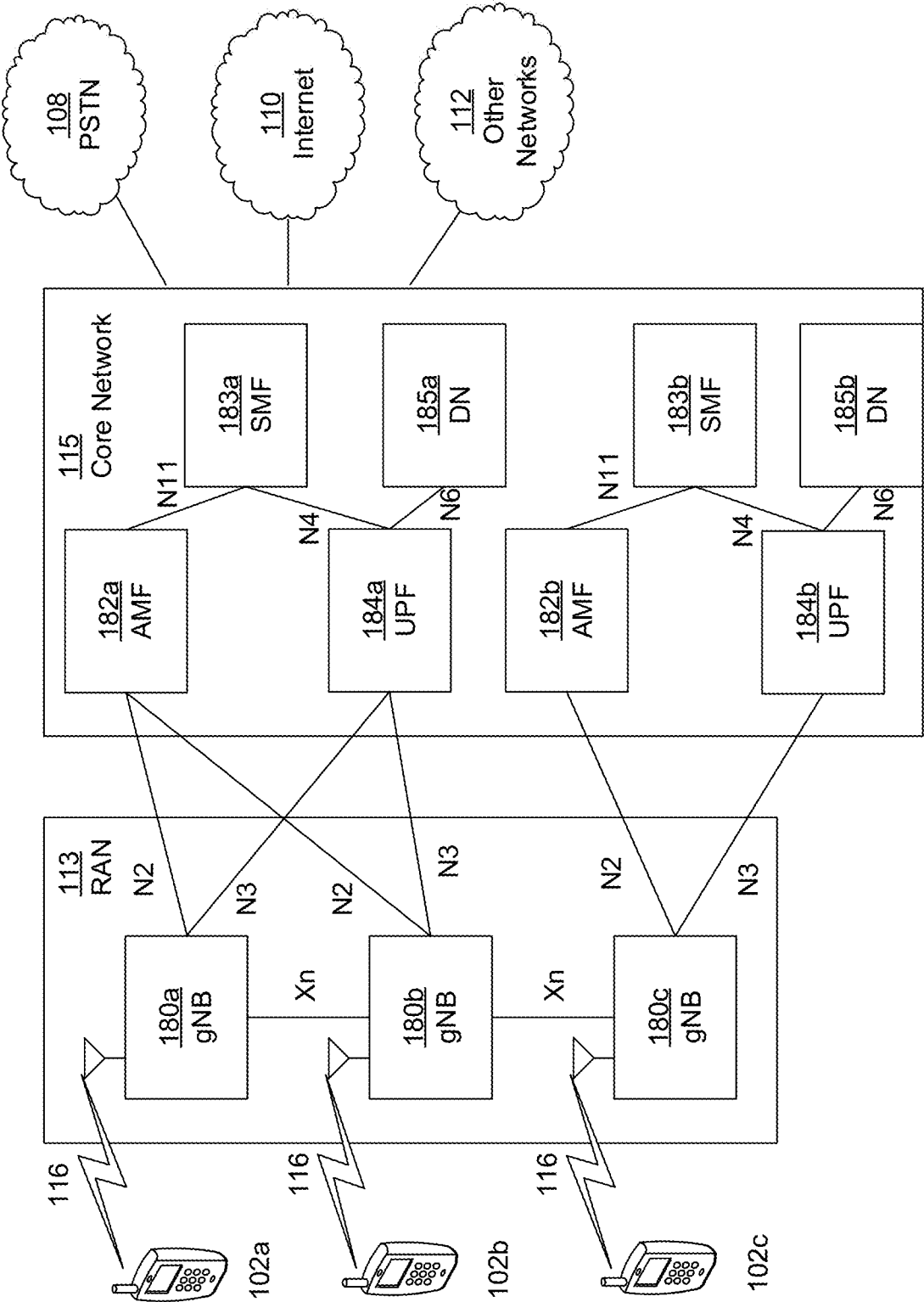


FIG. 1D

200

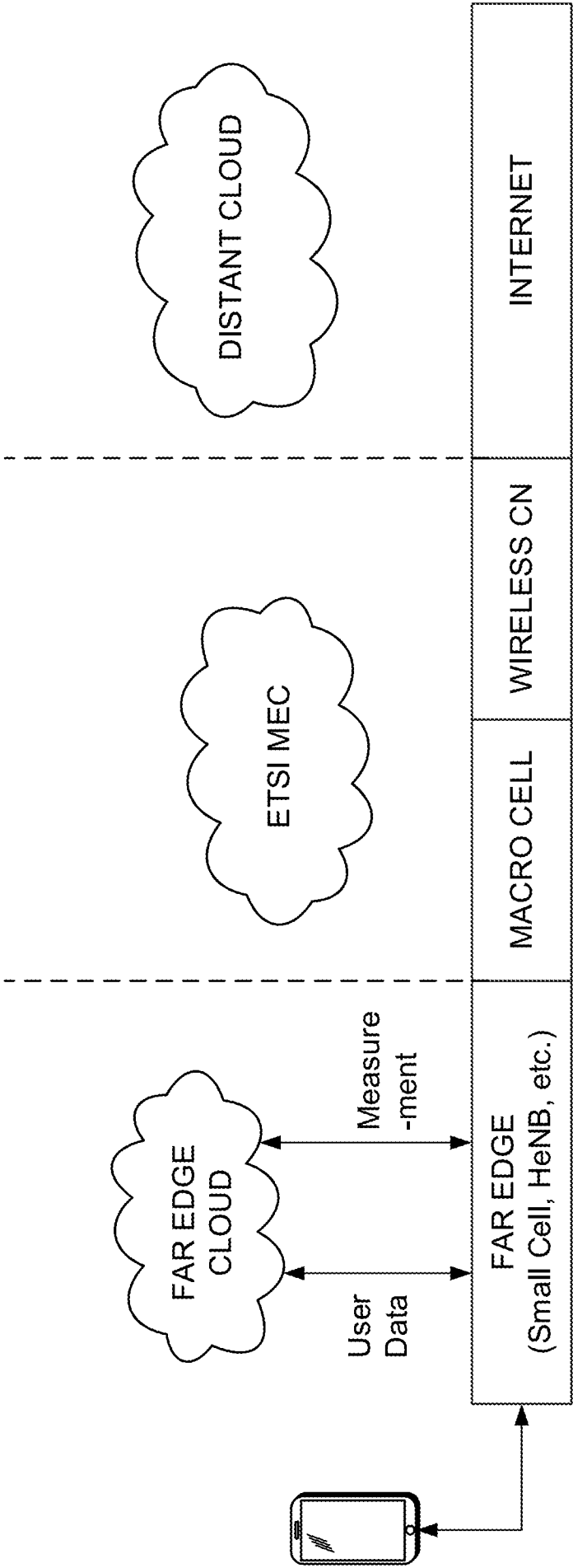


FIG. 2A

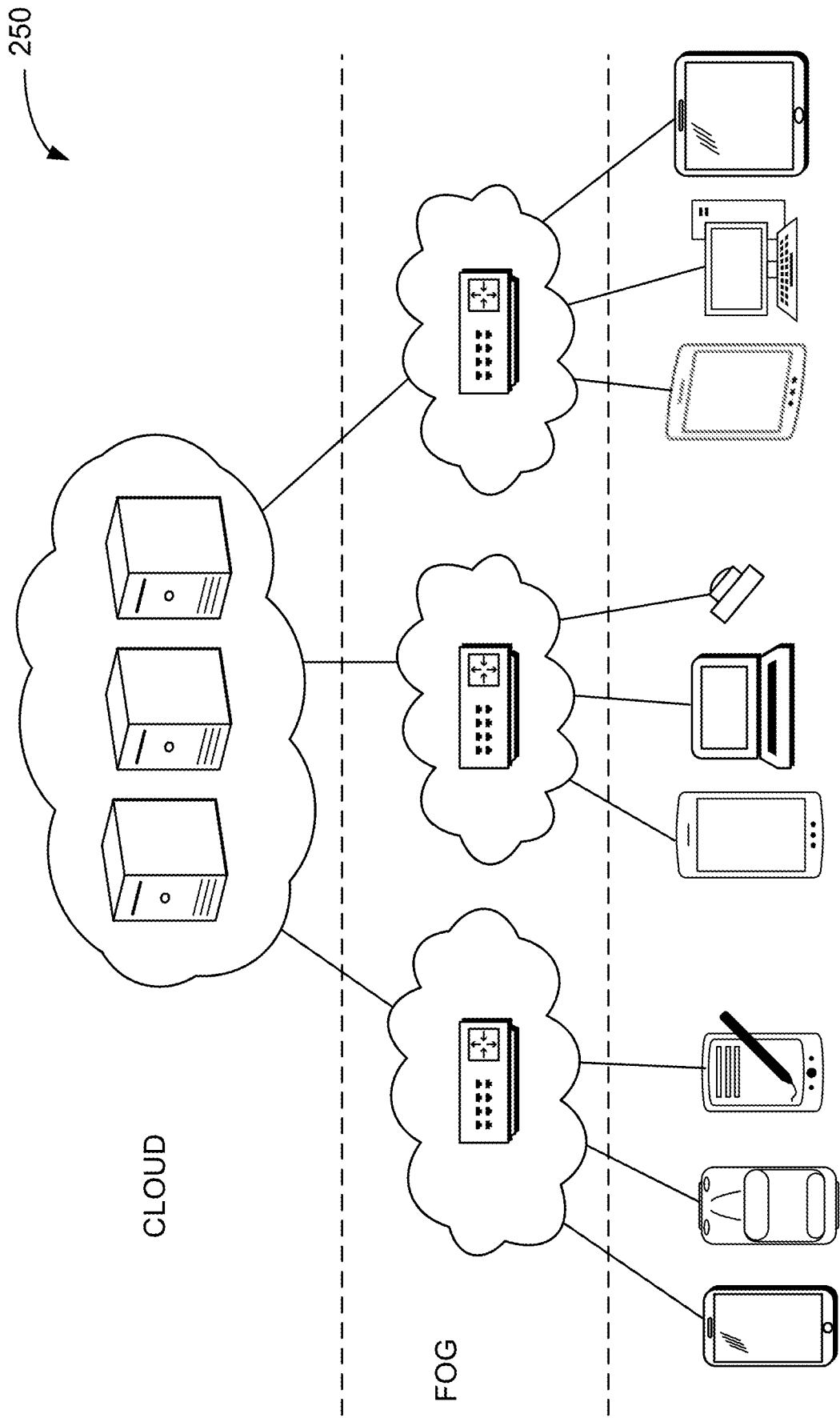


FIG. 2B

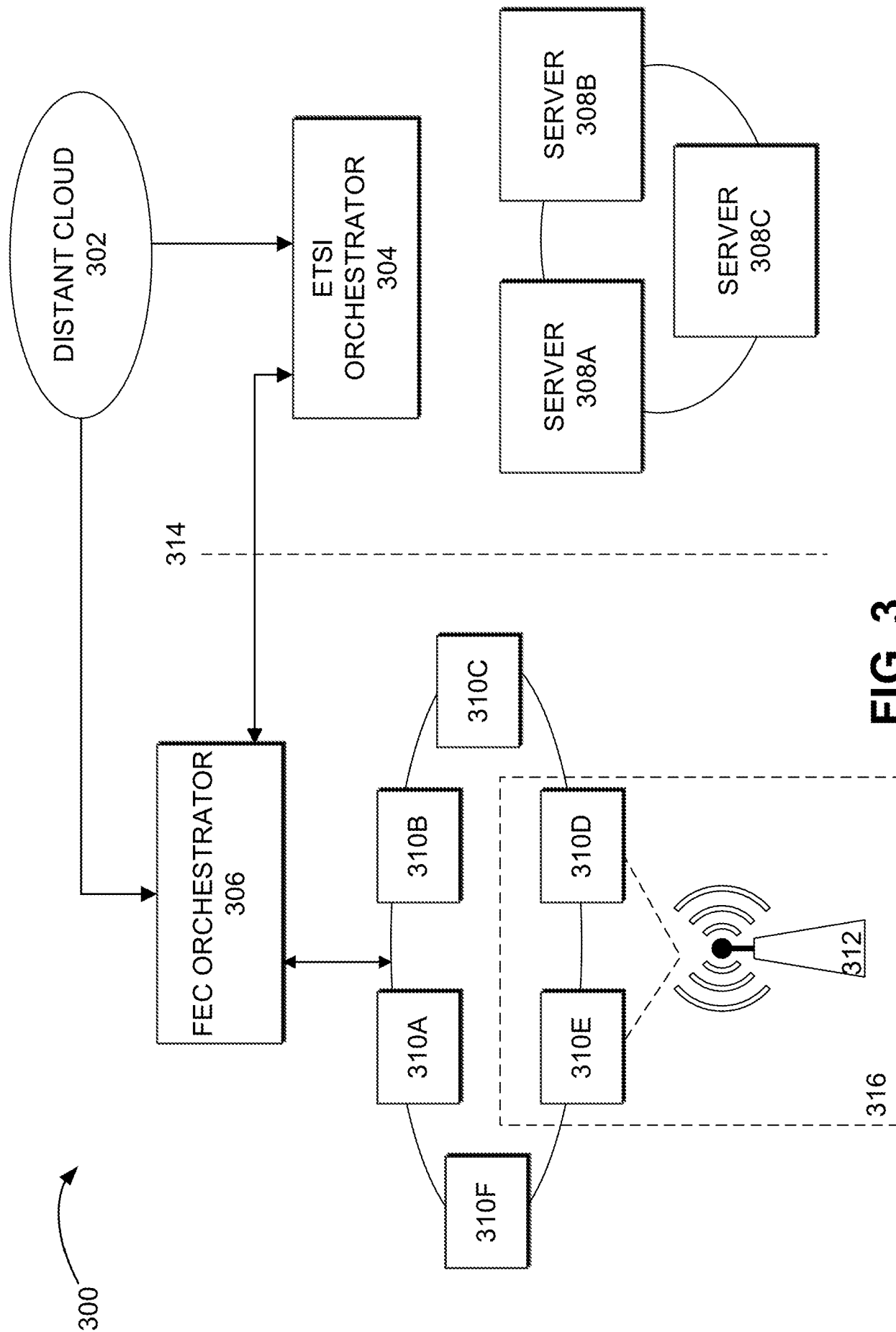
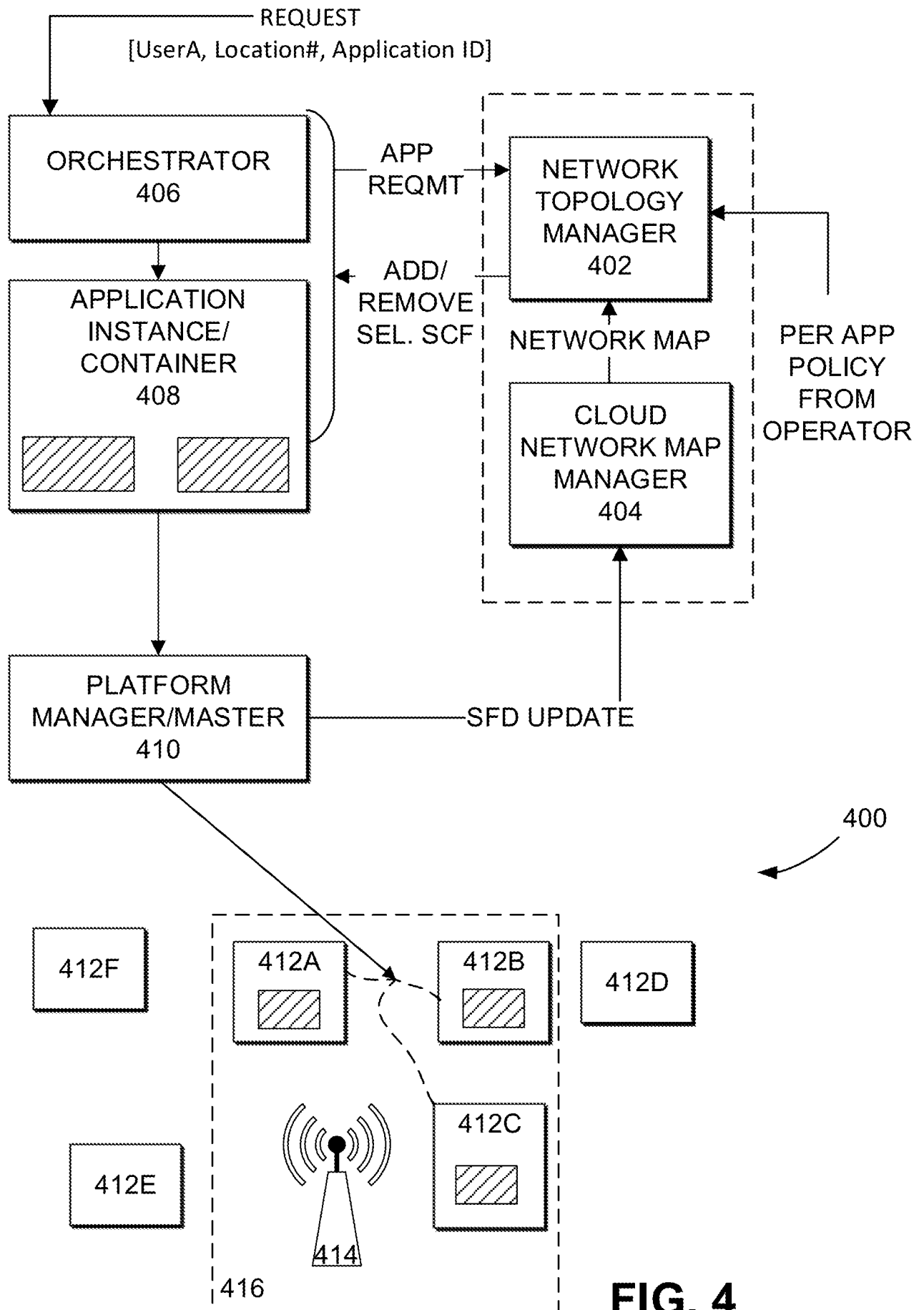


FIG. 3

**FIG. 4**

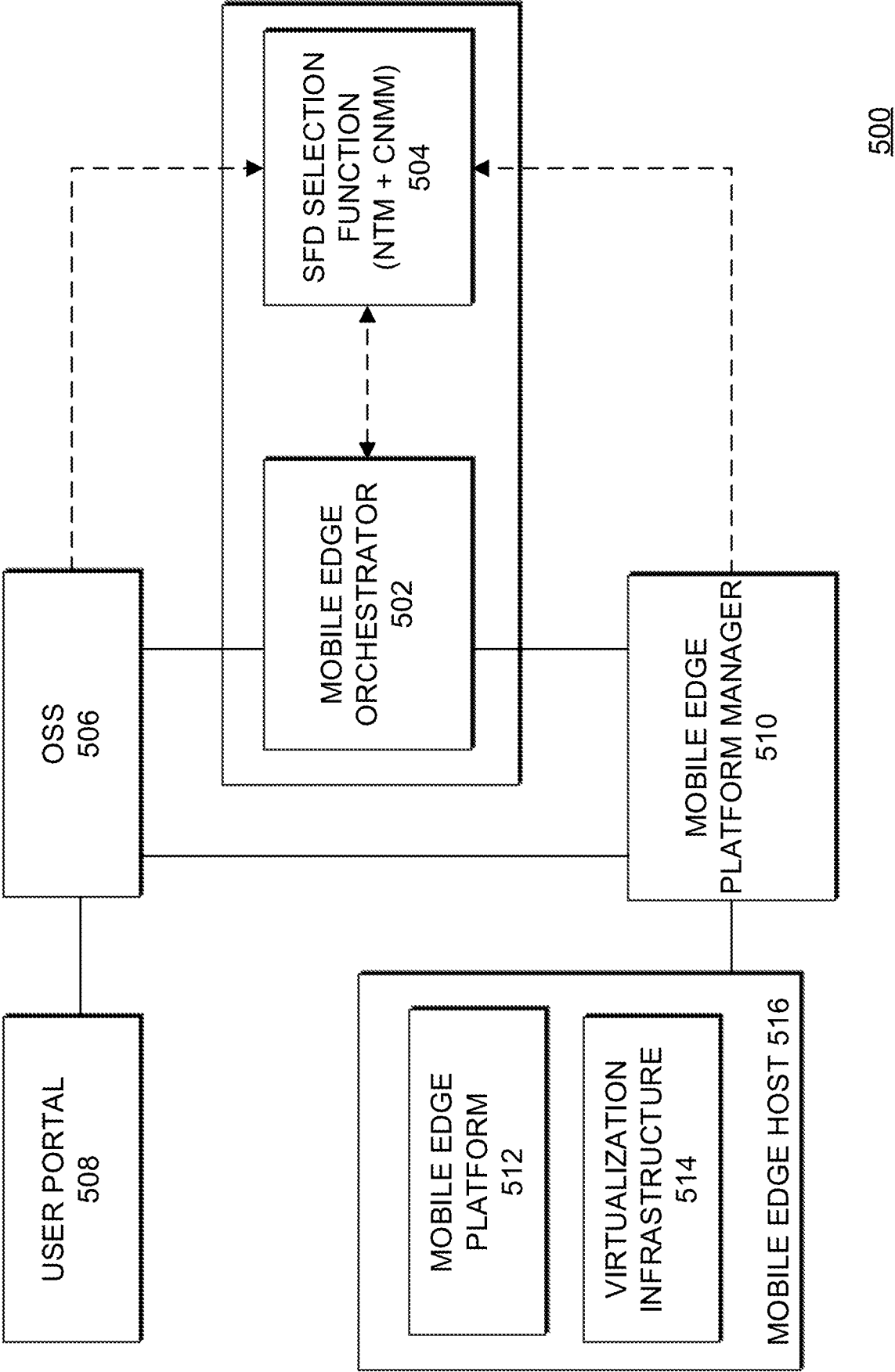


FIG. 5

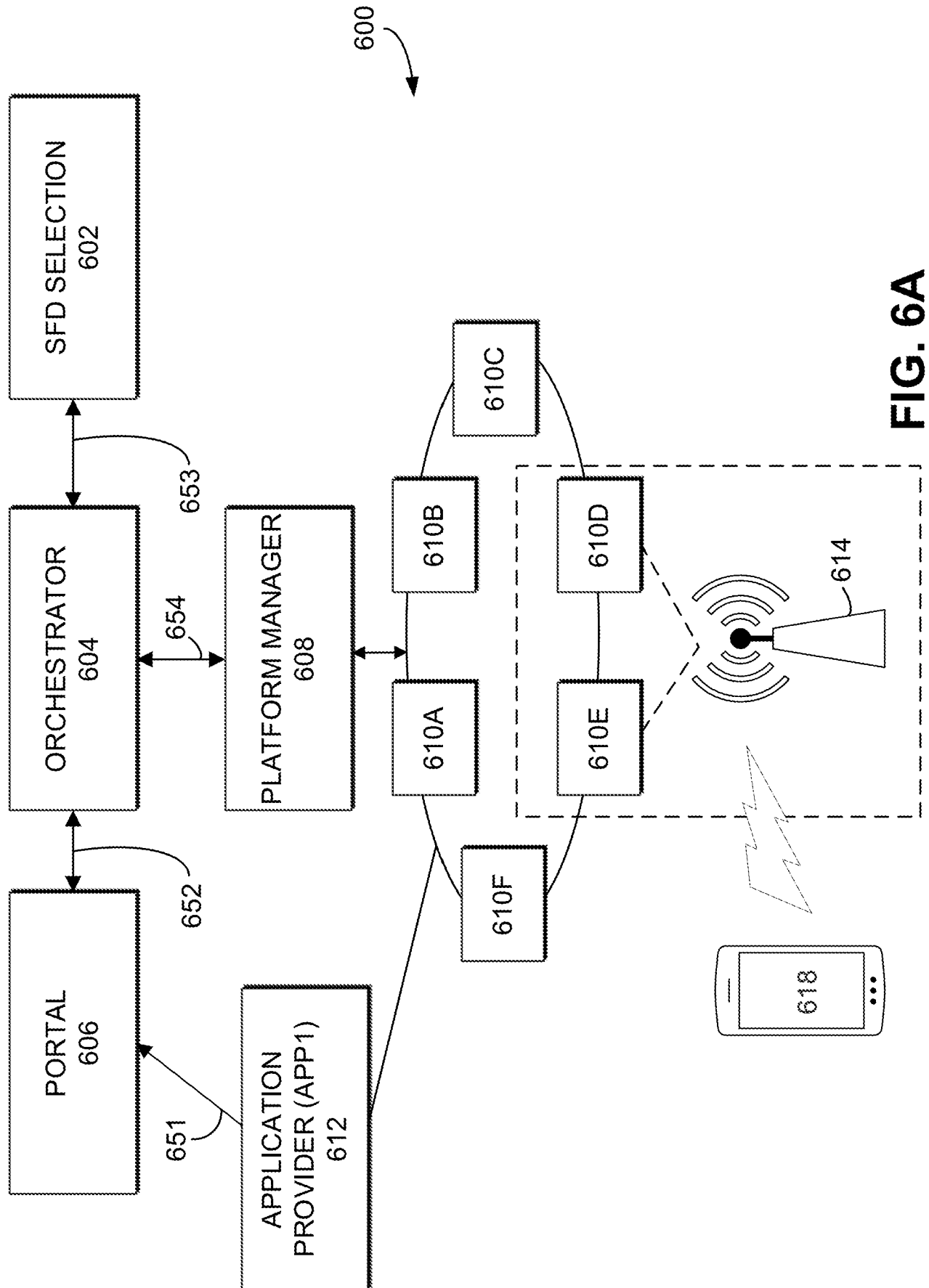


FIG. 6A

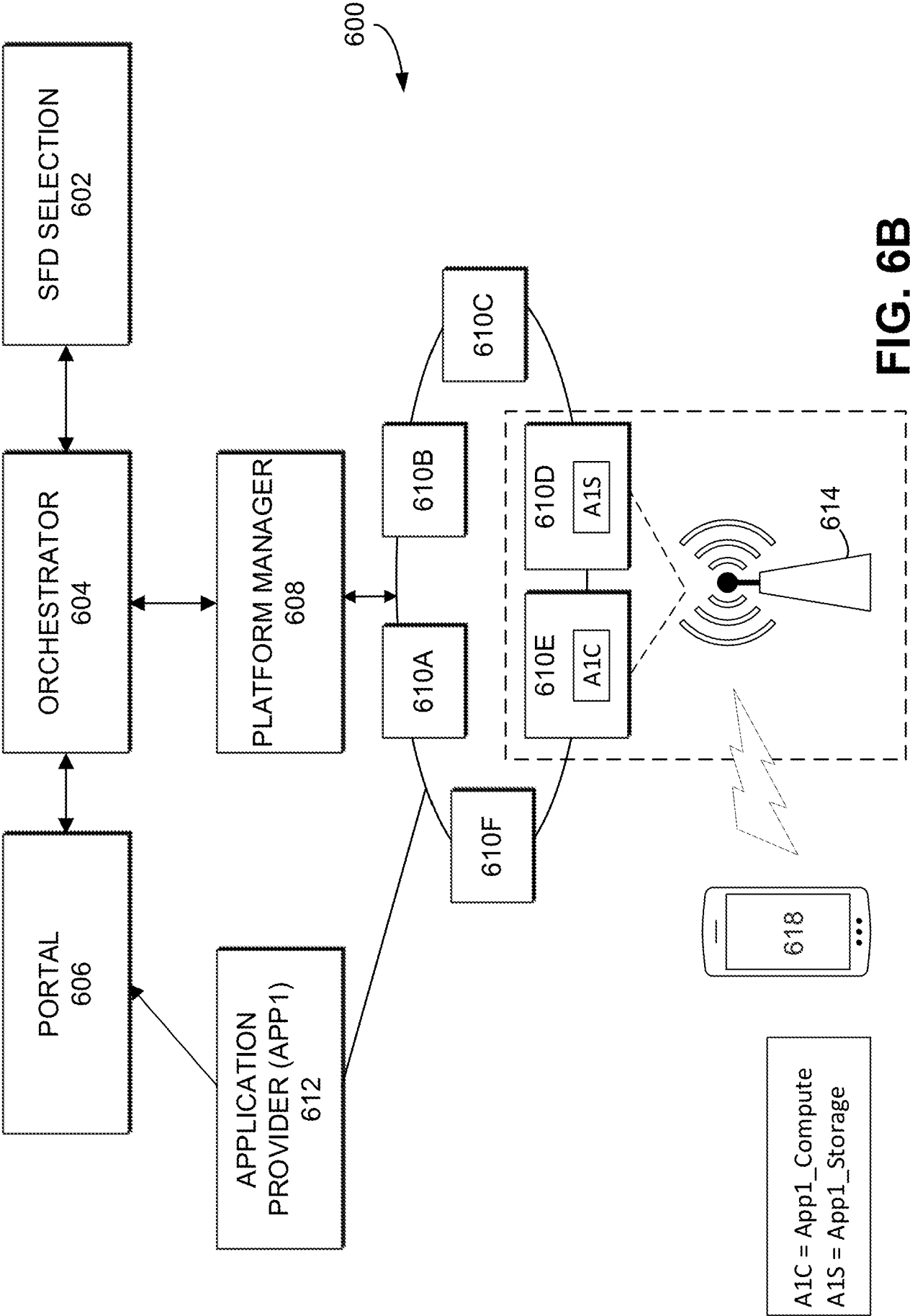


FIG. 6B

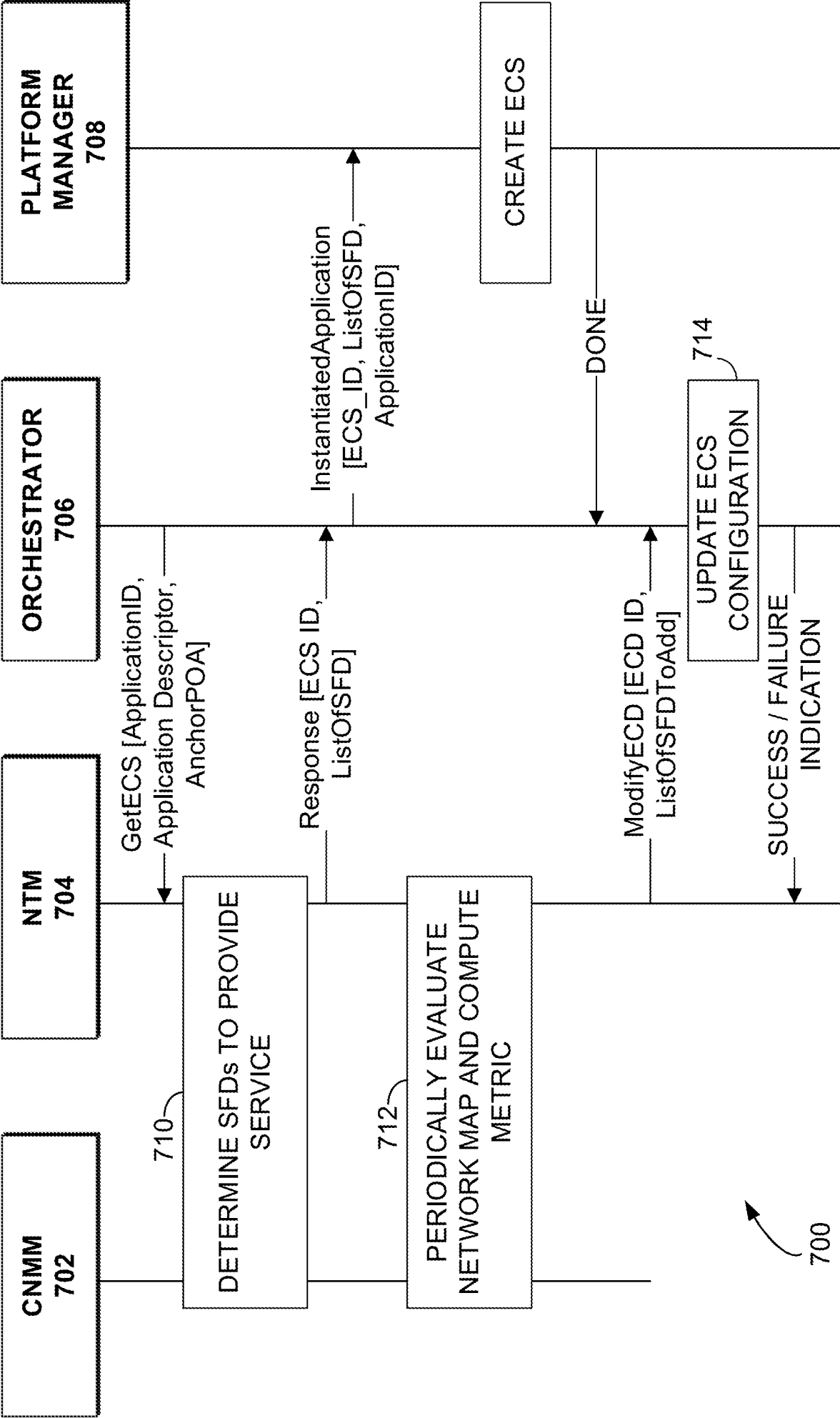


FIG. 7

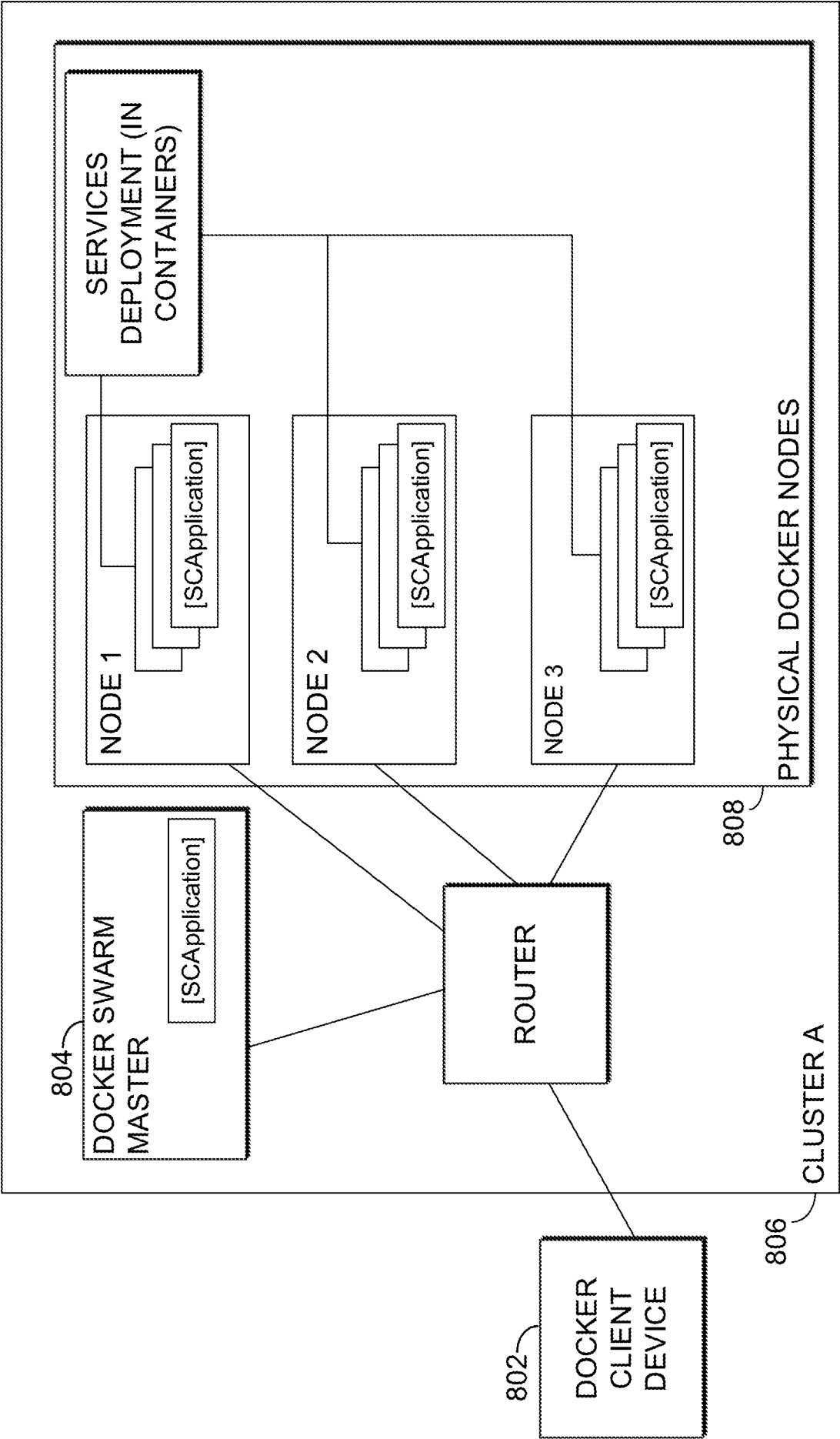


FIG. 8

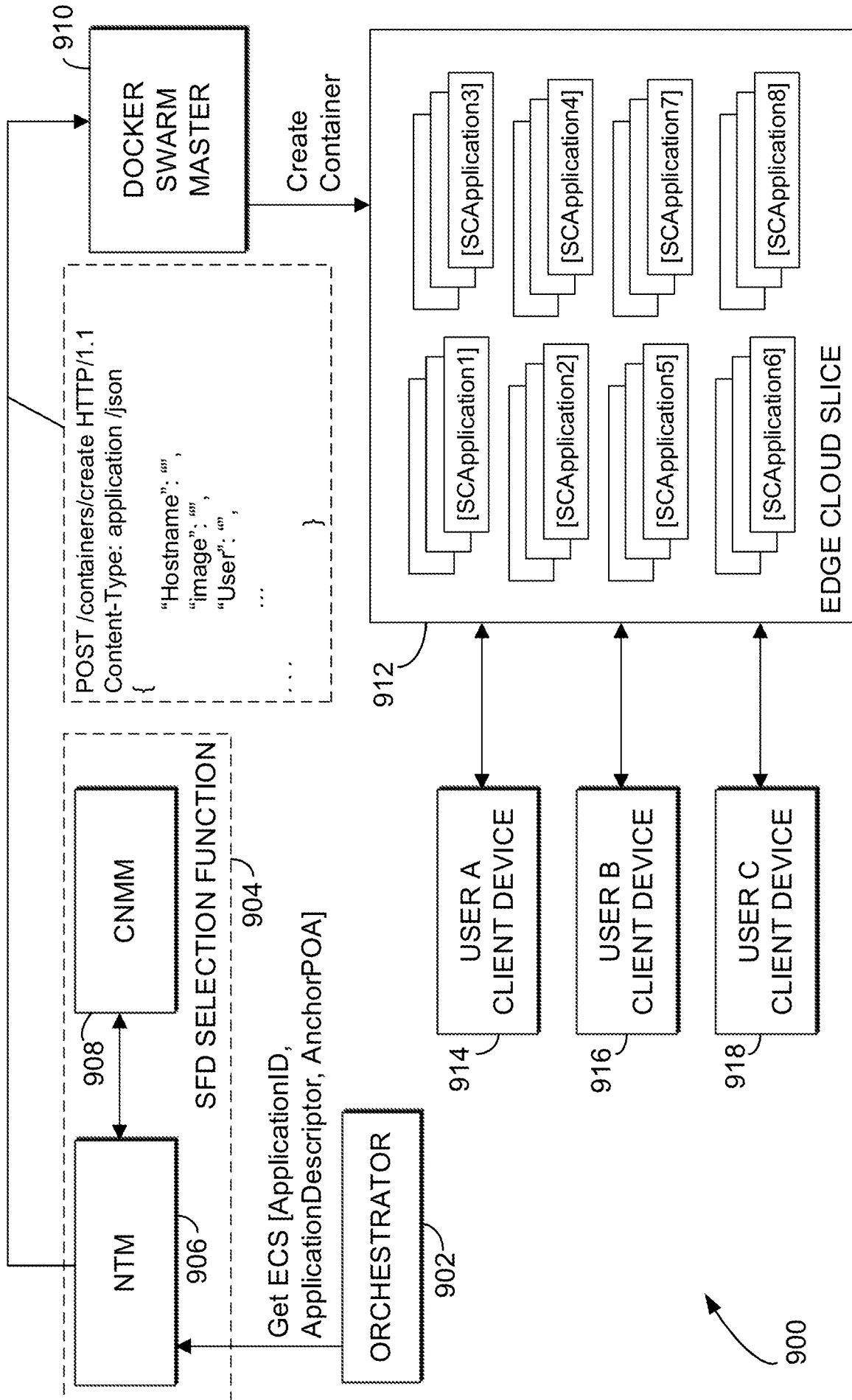
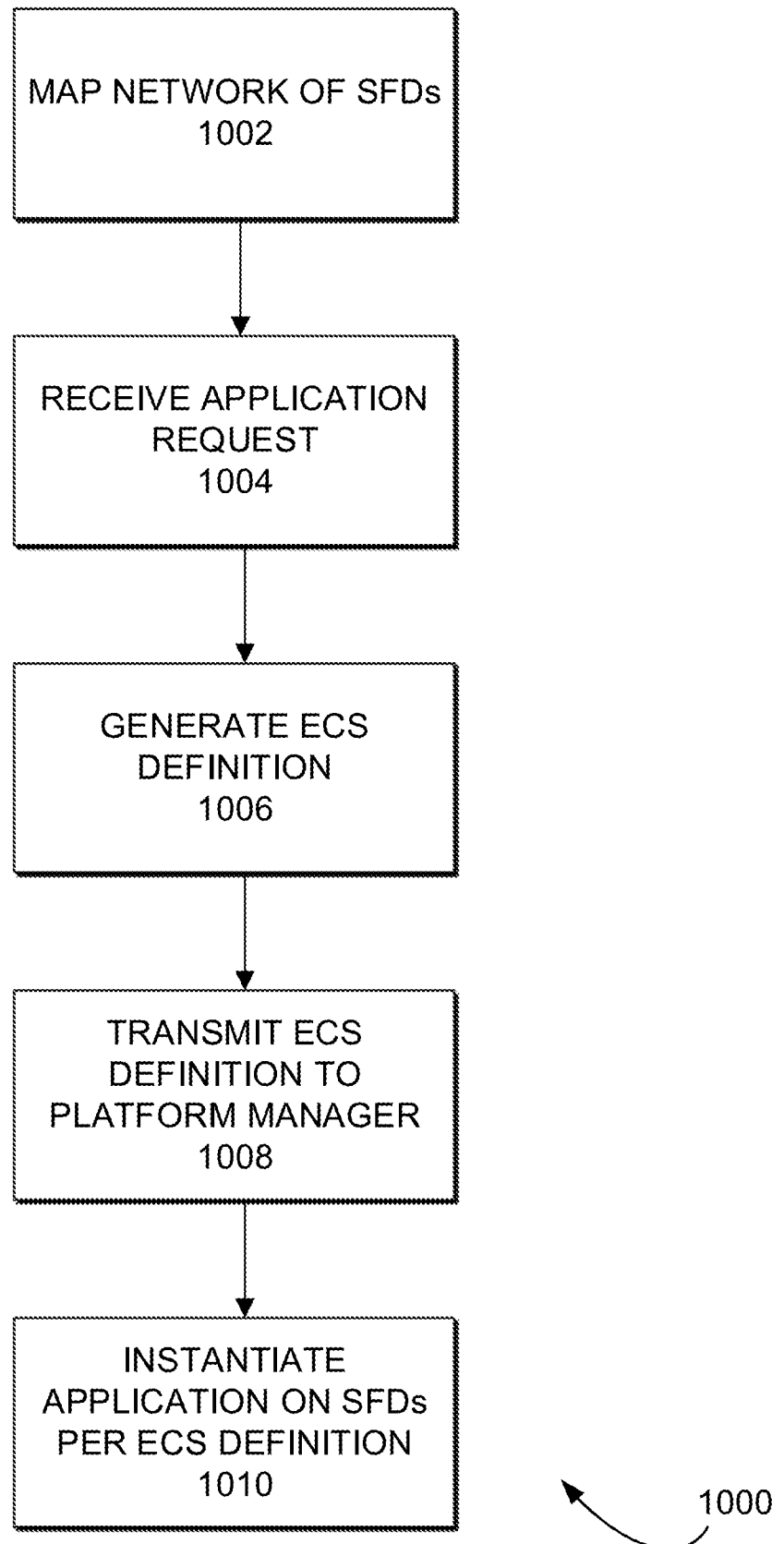


FIG. 9

15/15

**FIG. 10**

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2017/060528

A. CLASSIFICATION OF SUBJECT MATTER

INV. H04L29/08

ADD.

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)

H04L H04W

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)

EPO-Internal, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 2012/239792 A1 (BANERJEE SUBRATA [US] ET AL) 20 September 2012 (2012-09-20) abstract paragraphs [0044] - [0050]; figures 3,4,5 -----	1-15
A	US 2011/282975 A1 (CARTER STEPHEN R [US]) 17 November 2011 (2011-11-17) abstract paragraphs [0005] - [0072] -----	1-15



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

12 January 2018

Date of mailing of the international search report

30/01/2018

Name and mailing address of the ISA/

European Patent Office, P.B. 5818 Patentlaan 2
NL - 2280 HV Rijswijk
Tel. (+31-70) 340-2040,
Fax: (+31-70) 340-3016

Authorized officer

Hackl, Alexander

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2017/060528

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 2012239792	A1	20-09-2012	NONE

US 2011282975	A1	17-11-2011	EP 2410427 A1 25-01-2012
		US 2011282975 A1	17-11-2011
