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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
06 March 2025 (06.03.2025)

WIPO | PCT

(10) International Publication Number
WO 2025/045809 A1

(51) International Patent Classification:
G06N 3/098 (2023.01)

(21) International Application Number:
PCT/EP2024/073793

(22) International Filing Date:
26 August 2024 (26.08.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
23306451.8 31 August 2023 (31.08.2023) EP

(71) Applicant: **INTERDIGITAL CE PATENT HOLDINGS, SAS** [FR/FR]; 3 rue du Colonel Moll, 75017 Paris (FR).

(72) Inventors: **FILOCHE, Thierry**; c/o InterDigital, Immeuble ZEN 2, 845A Avenue des Champs Blancs, 35510 CESSON-SEVIGNE (FR). **QUINQUIS, Cyril**; c/o InterDigital, Immeuble ZEN 2, 845A Avenue des Champs Blancs, 35510 CESSON-SEVIGNE (FR). **ONNO, Stephane**; c/o InterDigital, Immeuble ZEN 2, 845A Avenue des Champs Blancs, 35510 CESSON-SEVIGNE (FR).

(74) Agent: **INTERDIGITAL**; Immeuble Zen 2, 845A Avenue des Champs Blancs, 35510 CESSON-SEVIGNE (FR).

(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, CV, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IQ, IR, IS, IT, JM, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, MG, MK, MN, MU, 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, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, CV, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SC, 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, ME, 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).

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

(54) Title: METHODS, ARCHITECTURES, APPARATUSES AND SYSTEMS FOR DISTRIBUTED ARTIFICIAL INTELLIGENCE

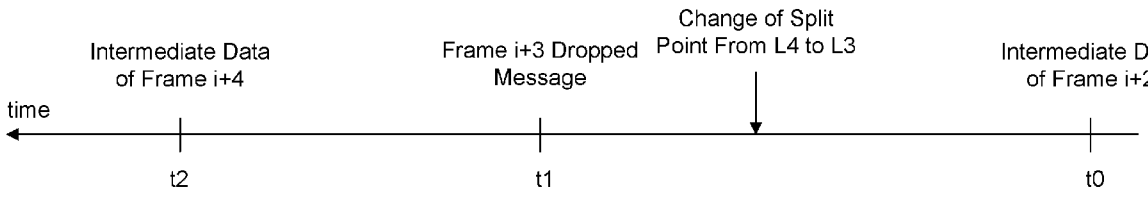

FIG. 3

(57) Abstract: Procedures, methods, architectures, apparatuses, systems, devices, and computer program products for distributed Artificial Intelligence, AI. A first device iteratively processes, using a machine learning model with a plurality of layers, a number of parts of sequentially organized data up to an intermediate layer of the machine learning model to obtain corresponding intermediate data, determines not to process of a subsequent part of the sequentially organized data, and transmits to a second device information indicating that the subsequent part has been dropped.

WO 2025/045809 A1

METHODS, ARCHITECTURES, APPARATUSES AND SYSTEMS FOR DISTRIBUTED ARTIFICIAL INTELLIGENCE

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of European Patent Application 23306451.8, filed 31 August 2023, which is incorporated herein by reference in its entirety.

BACKGROUND

[0002] The present disclosure is generally directed to the fields of communications, software and encoding, including, for example, to methods, architectures, apparatuses, systems directed to collaborative Artificial Intelligence (AI).

SUMMARY

[0003] In a first aspect, the present principles are directed to a first device including at least one hardware processor configured to, using a machine learning model with a plurality of layers, iteratively process a number of parts of sequentially organized data up to an intermediate layer of the machine learning model to obtain corresponding intermediate data, determine not to process of a subsequent part the sequentially organized data, and transmit to a second device information indicating that the subsequent part has been dropped.

[0004] In a second aspect, the present principles are directed to a method performed in a first device, the method including, iteratively processing, using a machine learning model with a plurality of layers, a number of parts of sequentially organized data up to an intermediate layer of the machine learning model to obtain corresponding intermediate data, determining not to process a subsequent part of the sequentially organized data, and transmitting to a second device information indicating that the subsequent part has been dropped.

[0005] In a third aspect, the present principles are directed to a non-transitory computer readable medium storing program code instructions that, when executed by a processor, implement a method including iteratively processing, using a machine learning model with a plurality of layers, a number of parts of sequentially organized data up to an intermediate layer of the machine learning model to obtain corresponding intermediate data, determining not to process a subsequent part of the sequentially organized data, and transmitting to a second device information indicating that the subsequent part has been dropped.

BRIEF DESCRIPTION OF THE DRAWINGS

[0006] A more detailed understanding may be had from the detailed description below, given by way of example in conjunction with drawings appended hereto. Figures in such drawings,

like the detailed description, are examples. As such, the Figures (FIGs.) and the detailed description are not to be considered limiting, and other equally effective examples are possible and likely. Furthermore, like reference numerals ("ref.") in the FIGs. indicate like elements, and wherein:

[0007] FIG. 1A is a system diagram illustrating an example communications system;

[0008] 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;

[0009] 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;

[0010] 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;

[0011] FIG. 2 illustrates different examples of splitting of an AI/ML model;

[0012] FIG. 3 illustrates an example of a scenario according to the first embodiment of the present principles; and

[0013] FIG. 4 illustrates an example of a scenario according to the second embodiment of the present principles.

DETAILED DESCRIPTION

[0014] In the following detailed description, numerous specific details are set forth to provide a thorough understanding of embodiments and/or examples disclosed herein. However, it will be understood that such embodiments and examples may be practiced without some or all of the specific details set forth herein. In other instances, well-known methods, procedures, components and circuits have not been described in detail, so as not to obscure the following description. Further, embodiments and examples not specifically described herein may be practiced in lieu of, or in combination with, the embodiments and other examples described, disclosed or otherwise provided explicitly, implicitly and/or inherently (collectively "provided") herein. Although various embodiments are described and/or claimed herein in which an apparatus, system, device, etc. and/or any element thereof carries out an operation, process, algorithm, function, etc. and/or any portion thereof, it is to be understood that any embodiments described and/or claimed herein assume that any apparatus, system, device, etc. and/or any element thereof is configured to carry out any operation, process, algorithm, function, etc. and/or any portion thereof.

[0015] Example Communications System

[0016] The methods, apparatuses and systems provided herein are well-suited for communications involving both wired and wireless networks. An overview of various types of wireless devices and infrastructure is provided with respect to FIGs. 1A-1D, where various elements of the network may utilize, perform, be arranged in accordance with and/or be adapted and/or configured for the methods, apparatuses and systems provided herein.

[0017] FIG. 1A is a system 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 (ZT) unique-word (UW) discreet Fourier transform (DFT) spread OFDM (ZT UW DTS-s OFDM), unique word OFDM (UW-OFDM), resource block-filtered OFDM, filter bank multicarrier (FBMC), and the like.

[0018] As shown in FIG. 1A, the communications system 100 may include wireless transmit/receive units (WTRUs) 102a, 102b, 102c, 102d, a radio access network (RAN) 104/113, a core network (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 (or be) 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.

[0019] 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, e.g., to facilitate access to one or more communication networks, such as the CN 106/115, the Internet 110, and/or the networks 112. By way of example, the base stations 114a, 114b may be any of a base transceiver station (BTS), a Node-B (NB), an eNode-B (eNB), a Home Node-B (HNB), a Home eNode-B (HeNB), a gNode-B (gNB), a NR Node-B (NR NB), 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.

[0020] 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 an 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 or any sector of the cell. For example, beamforming may be used to transmit and/or receive signals in desired spatial directions.

[0021] 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).

[0022] 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 116 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 Packet Access (HSDPA) and/or High-Speed Uplink Packet Access (HSUPA).

[0023] 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).

[0024] 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).

[0025] 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., an eNB and a gNB).

[0026] In an embodiment, the base station 114a and the WTRUs 102a, 102b, 102c may implement radio technologies such as IEEE 802.11 (i.e., Wireless Fidelity (Wi-Fi), 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.

[0027] 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 an 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 an 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 any of a small cell, 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.

[0028] 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 an NR radio technology, the CN 106/115 may also be in communication with another RAN (not shown) employing any of a GSM, UMTS, CDMA 2000, WiMAX, E-UTRA, or Wi-Fi radio technology.

[0029] 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 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 104/114 or a different RAT.

[0030] 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.

[0031] 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 elements/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.

[0032] 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, e.g., in an electronic package or chip.

[0033] 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 an 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 an 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.

[0034] 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. For example, the WTRU 102 may employ MIMO technology. Thus, in an 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.

[0035] 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.

[0036] 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).

[0037] 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.

[0038] 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.

[0039] The processor 118 may further be coupled to other elements/peripherals 138, which may include one or more software and/or hardware modules/units that provide additional features, functionality and/or wired or wireless connectivity. For example, the elements/peripherals 138 may include an accelerometer, an e-compass, a satellite transceiver, a digital camera (e.g., 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 elements/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.

[0040] 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 uplink (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 uplink (e.g., for transmission) or the downlink (e.g., for reception)).

[0041] 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, and 102c over the air interface 116. The RAN 104 may also be in communication with the CN 106.

[0042] 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 an 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 receive wireless signals from, the WTRU 102a.

[0043] Each of the eNode-Bs 160a, 160b, and 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 uplink (UL) and/or downlink (DL), and the like. As shown in FIG. 1C, the eNode-Bs 160a, 160b, 160c may communicate with one another over an X2 interface.

[0044] 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 (PGW) 166. While each of the foregoing elements are depicted as part of the CN 106, it will be appreciated that any one of these elements may be owned and/or operated by an entity other than the CN operator.

[0045] The MME 162 may be connected to each of the eNode-Bs 160a, 160b, and 160c 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.

[0046] 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.

[0047] 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.

[0048] 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.

[0049] 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.

[0050] In representative embodiments, the other network 112 may be a WLAN.

[0051] 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 into 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.

[0052] 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 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.

[0053] 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.

[0054] Very high throughput (VHT) STAs may support 20 MHz, 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 a medium access control (MAC) layer, entity, etc.

[0055] 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 (MTC), 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).

[0056] 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.

[0057] 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.

[0058] 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.

[0059] 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 an 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 WTRUs 102a, 102b, 102c. 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).

[0060] The WTRUs 102a, 102b, 102c may communicate with gNBs 180a, 180b, 180c using transmissions associated with a scalable numerology. For example, 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., including a varying number of OFDM symbols and/or lasting varying lengths of absolute time).

[0061] 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.

[0062] 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 functions (UPFs) 184a, 184b, routing of control plane information towards access and mobility management functions (AMFs) 182a, 182b, and the like. As shown in FIG. 1D, the gNBs 180a, 180b, 180c may communicate with one another over an Xn interface.

[0063] 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 at least one 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.

[0064] 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 protocol data unit (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, e.g., 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 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 Wi-Fi.

[0065] 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.

[0066] 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, e.g., to facilitate communications between the WTRUs 102a, 102b, 102c and IP-enabled devices. The UPF 184, 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.

[0067] 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 an 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.

[0068] In view of FIGs. 1A-1D, and the corresponding description of FIGs. 1A-1D, one or more, or all, of the functions described herein with regard to any of: WTRUs 102a-d, base stations 114a-b, eNode-Bs 160a-c, MME 162, SGW 164, PGW 166, gNBs 180a-c, AMFs 182a-b, UPFs 184a-b, SMFs 183a-b, DNs 185a-b, and/or any other element(s)/device(s) described herein, may be performed by one or more emulation elements/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.

[0069] 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.

[0070] 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.

[0071] **Introduction**

[0072] While Artificial Intelligence (AI), in particular Machine Learning (ML) can be a very powerful tool in various devices, it will be understood that the resource (e.g. processing) requirements can be too big for devices with relatively limited resources. Such devices, in this description denoted UEs, can be end user devices, in particular mobile devices such as smartphones and tablets. For this reason, it is a common solution to let one or more other

devices perform at least part of the calculations. It is noted that 'device' (also called 'node') in this context may mean a plurality of devices acting as one, for example in the case of server banks and cloud computing.

[0073] Apart from being done on a single device, AI/ML inference can thus be divided (i.e. split) over different points (e.g. from UEs to Edge or Cloud devices) for a certain AI/ML application that for example may be based on a Deep Neural Network (DNN). The AI/ML application may for example be split along the interface between two layers in a DNN, or between different parts where one (e.g. a part detecting facial features) can provide a result as input to a following part (e.g. using detected facial features to a person's mood). The corresponding AI/ML model, i.e. the computer program (e.g. trained, i.e. with proper parameters), can thus be split into several AI/ML model subsets to be run on different devices/servers. Each AI/ML model subset is a piece of independent software to run on a split node such as for example a UE, an Edge device, in the Cloud or in an Access Network.

[0074] FIG. 2 illustrates different examples of splitting of an AI/ML model. An example AI/ML model can be split in different ways. For example, split model M can include AI/ML model subsets {M0, M1}. Similarly for split model M' can include {M'0, M'1} and split model M" {M0", M1", M2"}. These three examples of split models M, M', M" provide the same service and, as can be seen, can be split into a different number of subsets. The border between different subsets will be referred to as a splitting point (split point, partition point), indicated by an illustrative arrow in one instance. Similarly, another model M can include AI/ML model subsets {N0, N1} while a model N' can include {N'0, N'1,N'2}. The Model M and N can have the same DNN layers composition with same split points, but they may differ for example, in the weight values of the neurons, the bias values, or the quantization level of any input value.

[0075] In addition, different subsets can run or be intended to be run on different devices. Which device a subset runs on can depend on different conditions. Example conditions include device capabilities, device load, and network load. A specific subset, say M0, can thus for example run either on an edge device or on the UE.

[0076] When processing a video sequence using a split AI/ML model (i.e., a partitioned AI/ML pipeline), a first device (e.g., the UE) can perform inference, for example individually, on a series of frames and transmit to a second device (e.g., an edge device or an Internet server) resulting intermediate data with additional information (e.g., split point ID, sequence number, timestamp).

[0077] An example message structure to wrap data for exchange within the AI/ML partitioned pipeline is:

[0078] **IntermediateDataWrapper:**

```
Struct IntermediateDataWrapper {
    Int    Model_RefID;              # Model Identifier
    bit    ChangeOfSplitPoint;       # 1 if the splitpoint is changed from the previous message
    Int SplitPointID;                # Split point used by the UE
    Int    IntermediateDataLength;   # Length of Intermediate Data
    Dim    IntermediateDataDim;      # Array dimension of intermediate data
    byte   EncodingMethod;           # Indicate if data are compressed and by which algorithm
    int    SequenceNumber;           # Sequential identifier of input data
    double  TimeStamp;               # Timestamp of the intermediate data
    Byte  IntermediateData[]         # Intermediate Data
    };
```

[0079] The UE can thus perform inference up to the split point (i.e., up to a given AI/ML layer) on a given frame, e.g., frame i, and transmit the resulting intermediate data to the network side device together with additional intermediate data information for example in a message in the intermediate data wrapper format. The network side device receiving the intermediate data corresponding to frame i and processes it using the following layer(s) of the model (according to the split point identifier).

[0080] It is noted that the split point may be different between frames and that the change of split point may occur while a frame is being processed. It is also noted that the UE can process a plurality of frames [i, i+1, …, i+k] simultaneously, as can the network side (although typically not the same frames as the network side needs the intermediate data from the UE).

[0081] As it can be possible to have different split points for different frames, a first frame can be processed using a model with a first split point (e.g., split==4) and the subsequent frame can be processed using a model with a first split point (e.g., split==3). This change of split point can be handled in different ways depending on settings determining priorities.

[0082] If the main priority is not to lose a frame, the UE may wait for the end of processing of the frame in the current layer. As an example, the UE has output intermediate data

corresponding to frame i+2 to the network device. A change of split point (e.g., from 4 to 3) is determined while the UE is processing frames i+3 and i+4. The UE processes (i.e., performs inference on) the currently processed frames to the end and switches to the new split point for subsequent frames.

[0083] If main priority is low latency, the UE may skip one or more frames under processing and start the processing, using the new split point, of the subsequent frames. As an example, the UE sends intermediate data for frame i+2 to the network device, the split point is changed (e.g., from 4 to 3), which causes the UE to drop the currently processed frame(s) (e.g., frame i+3) and start processing frame i+4 using split point 3.

[0084] Using timestamps, it is possible to combine both ways of handling a change of split point. The UE sends the intermediate data with at least one of a timestamp and a sequence number and it is up to the device on the network side to skip or not the processing of delayed frame(s), for example based on the timestamp and/or the sequence number. As an example, the network device receives intermediate data for frame i+2 from the UE, a change of split point occurs on the first device (e.g., from 4 to 3), the network device receives intermediate data first for frame i+4, then for frame i+3. The network device can then, for example depending on its policy, determine to skip frame i+3 or to wait for its data and put the intermediate data in the correct order.

[0085] As can be seen, it is possible to transmit metadata associated with intermediate data for enabling dynamic split processing where the split point may change overtime. The processing of input data (e.g., a video sequence, an audio sequence or haptic data) at the first device may change underway depending on the split point and resulting intermediate data for a given part of the input data (e.g., a video frame) may be transmitted to the second device before the resulting intermediate data for a previous (e.g., chronologically anterior) part. Further, the use of sequence numbers in the metadata associated with intermediate data can enable the second device to reorder the received sequence of intermediate data (i.e., put it back in order).

[0086] However, the first device may decide to drop some previous frames to meet latency constraints, which, in case the second device is not informed of the dropped frames, may cause the second device wait for a frame (for reordering) that will never be received. This may cause inconsistent. As an example, the second device first receives the intermediate data for frame i+2 and then for frame i+4 (instead of the expected frame i+3). There are two possibilities: the intermediate data for frame i+3 will be received later, or it will never be received. Given the

two mutually exclusive possibilities, the second device may not handle the situation in an appropriate manner.

[0087] At the very least, if the second device has a FIFO buffer with a capacity to receive intermediate data from a plurality n (e.g., three) frames, it may need to wait until the buffer is full to determine that a given frame probably has been dropped, which can lead to delays in processing.

[0088] **Overview**

[0089] According to embodiments of the present principles, the first device (sender) transmits information about dropped frames (for example upon a change of split point) to the second device (receiver). When a frame is dropped, the first device can transmit to the second device information indicative of a dropped frame instead of intermediate data. The first device can also transmit to the second device, together with the intermediate data for a subsequent frame, information indicating that a previous frame (or frames) has been dropped.

[0090] In a first embodiment, information about dropped frame(s) is transmitted instead of (expected) intermediate data for the frame(s) by the first device to the second device that then can take appropriate action, for example reconsidering the order of the received metadata to take into account dropped frames. If, for instance, the second device has received frames 14 and 16 as well as an indication that frame 15 has been dropped, the second device can then determine that frame 16 is the frame following frame 14 (instead of 15).

[0091] The information about one or more dropped frame may be sent in a message that can include:

Model Identifier: unique identifier of the AI/ML model;

Split point Identifier: unique identifier of the model split point;

Sequence Number: sequential identifier of input data (e.g., frame number); and

Dropped indication: indicates whether the identified sequence of input data (e.g., frame) has been dropped.

[0092] In a first variant, the metadata message can further include:

Dropped Sequence Number: the sequence identifier of the dropped frame

[0093] The first variant can provide information, for example in case the 'dropped' message for a previous frame has not been sent or is lost in transmission.

[0094] In a second variant, the metadata message can include:

Dropped Sequence list: the list of the sequence number of all the dropped frames.

[0095] Examples of adapted frame wrappers are proposed in the annex hereto.

[0096] FIG. 3 illustrates an example of a scenario according to the first embodiment. In the scenario, due to the change of a split point from layer L4 to layer L3, the intermediate data associated to the frame i+3 is not sent.

[0097] At time t_0, the second device receives intermediate data and metadata corresponding to frame i+2.

[0098] During processing of frame i+3, a change of split point from Layer4 to Layer3 occurs on the first device. As a result, the frame i+3 will be dropped by the first device and the intermediate data will not be sent to the second device. Instead, at time t_1, the first device sends a message to the second device to inform that frame i+3 has been dropped.

[0099] The message can include:

Sequence number of the dropped frame;

Dropped indication set to "sequence dropped";

IntermediateDataLength set to 0;

SplitPoint set to the current SplitPoint (e.g., L4); and

NextSplitPoint set to the next SplitPoint that will be used for the next frame (e.g., L3).

[0100] Then the first device processes frame i+4 and sends the resulting intermediate data to the second device at time t_2.

[0101] Upon reception of this message, the second device knows that the sequence "sequence number" (e.g., frame i+3) has been dropped and that the next intermediate data will concern another layer (e.g., Layer L3). As a consequence, the second device may anticipate memory management by either unloading from memory a layer (e.g., unload L4, if the change at sender side is from L3 to L4) or load a new layer in memory (e.g., load L4, if the change at sender side is from L4 to L3). This anticipation can help decrease the latency at the second device.

[0102] This scenario may also be applicable when the sender for any reason determines to not process a specific frame (e.g., when it is not able to process all the frames at a certain frame rate) even if there is no change of split point.

[0103] In a second embodiment, information about dropped frame(s) is transmitted with intermediate data of at least one subsequent frame by the first device to the second device that then can take appropriate action.

[0104] The information about dropped frames can include:

An identifier of the AI/ML model;

An identifier of a previous split point;

An identifier of a present split point;

The sequence number of the intermediate data;

The number of sequences dropped;

The sequence number(s) of the frame(s) dropped;

The intermediate data of the subsequent frame(s); and

The encoding method used to encode the intermediate data of undropped frames.

[0105] FIG. 4 illustrates an example of a scenario according to the second embodiment of the present principles.

[0106] At time t_0, the second device receives intermediate data and metadata related to frame i+2.

[0107] During processing of frame i+3, a change of split point from Layer4 to Layer3 occurs on the first device. As a result, the first device drops frame i+3 that hence will not be sent to the second device.

[0108] The first device processes frame i+4 and sends to the second device information indicative of the intermediate data and metadata for frame i+4, as well as that frame i+3 has been dropped. The message can be sent with NbSequenceDropped set to 1 indicating that a previous frame has been dropped and Dropped Sequence number list set to [frame i+3].

[0109] Upon reception of this information, the second device is informed that frame i+3 has been dropped and can then take appropriate action.

[0110] **Conclusion**

[0111] It will be appreciated that the present principles provide the receiver in a distributed AI/ML processing system with information about sequences (e.g., frames) dropped by the sender. There is thus no need to wait for intermediate data of expected but dropped frames, which can avoid increasing the latency in the system.

[0112] Although features and elements are provided 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. The present disclosure is not to be limited in terms of the particular embodiments described in this application, which are intended as illustrations of various aspects. Many modifications and variations may be made without departing from its spirit and scope, as will be apparent to those skilled in the art. No element, act, or instruction used in the description of the present application should be construed as critical or essential to the invention unless explicitly provided as such.

Functionally equivalent methods and apparatuses within the scope of the disclosure, in addition to those enumerated herein, will be apparent to those skilled in the art from the foregoing descriptions. Such modifications and variations are intended to fall within the scope of the appended claims. The present disclosure is to be limited only by the terms of the appended claims, along with the full scope of equivalents to which such claims are entitled. It is to be understood that this disclosure is not limited to particular methods or systems.

[0113] The foregoing embodiments are discussed, for simplicity, with regard to the terminology and structure of infrared capable devices, i.e., infrared emitters and receivers. However, the embodiments discussed are not limited to these systems but may be applied to other systems that use other forms of electromagnetic waves or non-electromagnetic waves such as acoustic waves.

[0114] It is also to be understood that the terminology used herein is for the purpose of describing particular embodiments only, and is not intended to be limiting. As used herein, the term "video" or the term "imagery" may mean any of a snapshot, single image and/or multiple images displayed over a time basis. As another example, when referred to herein, the terms "user equipment" and its abbreviation "UE", the term "remote" and/or the terms "head mounted display" or its abbreviation "HMD" may mean or include (i) a wireless transmit and/or receive unit (WTRU); (ii) any of a number of embodiments of a WTRU; (iii) a wireless-capable and/or wired-capable (e.g., tetherable) device configured with, inter alia, some or all structures and functionality of a WTRU; (iii) a wireless-capable and/or wired-capable device configured with less than all structures and functionality of a WTRU; or (iv) the like. Details of an example WTRU, which may be representative of any WTRU recited herein, are provided herein with respect to FIGs. 1A-1D. As another example, various disclosed embodiments herein *supra* and *infra* are described as utilizing a head mounted display. Those skilled in the art will recognize that a device other than the head mounted display may be utilized and some or all of the disclosure and various disclosed embodiments can be modified accordingly without undue experimentation. Examples of such other device may include a drone or other device configured to stream information for providing the adapted reality experience.

[0115] In addition, the methods provided 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 media include electronic signals (transmitted over wired or wireless connections) and computer-readable storage media. 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.

[0116] Variations of the method, apparatus and system provided above are possible without departing from the scope of the invention. In view of the wide variety of embodiments that can be applied, it should be understood that the illustrated embodiments are examples only, and should not be taken as limiting the scope of the following claims. For instance, the embodiments provided herein include handheld devices, which may include or be utilized with any appropriate voltage source, such as a battery and the like, providing any appropriate voltage.

[0117] Moreover, in the embodiments provided above, processing platforms, computing systems, controllers, and other devices that include processors are noted. These devices may include at least one Central Processing Unit ("CPU") and memory. In accordance with the practices of persons skilled in the art of computer programming, reference to acts and symbolic representations of operations or instructions may be performed by the various CPUs and memories. Such acts and operations or instructions may be referred to as being "executed," "computer executed" or "CPU executed."

[0118] One of ordinary skill in the art will appreciate that the acts and symbolically represented operations or instructions include the manipulation of electrical signals by the CPU. An electrical system represents data bits that can cause a resulting transformation or reduction of the electrical signals and the maintenance of data bits at memory locations in a memory system to thereby reconfigure or otherwise alter the CPU's operation, as well as other processing of signals. The memory locations where data bits are maintained are physical locations that have particular electrical, magnetic, optical, or organic properties corresponding to or representative of the data bits. It should be understood that the embodiments are not limited to the above-mentioned platforms or CPUs and that other platforms and CPUs may support the provided methods.

[0119] The data bits may also be maintained on a computer readable medium including magnetic disks, optical disks, and any other volatile (e.g., Random Access Memory (RAM)) or non-volatile (e.g., Read-Only Memory (ROM)) mass storage system readable by the CPU. The computer readable medium may include cooperating or interconnected computer readable

medium, which exist exclusively on the processing system or are distributed among multiple interconnected processing systems that may be local or remote to the processing system. It should be understood that the embodiments are not limited to the above-mentioned memories and that other platforms and memories may support the provided methods.

[0120] In an illustrative embodiment, any of the operations, processes, etc. described herein may be implemented as computer-readable instructions stored on a computer-readable medium. The computer-readable instructions may be executed by a processor of a mobile unit, a network element, and/or any other computing device.

[0121] There is little distinction left between hardware and software implementations of aspects of systems. The use of hardware or software is generally (but not always, in that in certain contexts the choice between hardware and software may become significant) a design choice representing cost versus efficiency tradeoffs. There may be various vehicles by which processes and/or systems and/or other technologies described herein may be effected (e.g., hardware, software, and/or firmware), and the preferred vehicle may vary with the context in which the processes and/or systems and/or other technologies are deployed. For example, if an implementer determines that speed and accuracy are paramount, the implementer may opt for a mainly hardware and/or firmware vehicle. If flexibility is paramount, the implementer may opt for a mainly software implementation. Alternatively, the implementer may opt for some combination of hardware, software, and/or firmware.

[0122] The foregoing detailed description has set forth various embodiments of the devices and/or processes via the use of block diagrams, flowcharts, and/or examples. Insofar as such block diagrams, flowcharts, and/or examples include one or more functions and/or operations, it will be understood by those within the art that each function and/or operation within such block diagrams, flowcharts, or examples may be implemented, individually and/or collectively, by a wide range of hardware, software, firmware, or virtually any combination thereof. In an embodiment, several portions of the subject matter described herein may be implemented via Application Specific Integrated Circuits (ASICs), Field Programmable Gate Arrays (FPGAs), digital signal processors (DSPs), and/or other integrated formats. However, those skilled in the art will recognize that some aspects of the embodiments disclosed herein, in whole or in part, may be equivalently implemented in integrated circuits, as one or more computer programs running on one or more computers (e.g., as one or more programs running on one or more computer systems), as one or more programs running on one or more processors (e.g., as one or more programs running on one or more microprocessors), as

firmware, or as virtually any combination thereof, and that designing the circuitry and/or writing the code for the software and or firmware would be well within the skill of one of skill in the art in light of this disclosure. In addition, those skilled in the art will appreciate that the mechanisms of the subject matter described herein may be distributed as a program product in a variety of forms, and that an illustrative embodiment of the subject matter described herein applies regardless of the particular type of signal bearing medium used to actually carry out the distribution. Examples of a signal bearing medium include, but are not limited to, the following: a recordable type medium such as a floppy disk, a hard disk drive, a CD, a DVD, a digital tape, a computer memory, etc., and a transmission type medium such as a digital and/or an analog communication medium (e.g., a fiber optic cable, a waveguide, a wired communications link, a wireless communication link, etc.).

[0123] Those skilled in the art will recognize that it is common within the art to describe devices and/or processes in the fashion set forth herein, and thereafter use engineering practices to integrate such described devices and/or processes into data processing systems. That is, at least a portion of the devices and/or processes described herein may be integrated into a data processing system via a reasonable amount of experimentation. Those having skill in the art will recognize that a typical data processing system may generally include one or more of a system unit housing, a video display device, a memory such as volatile and non-volatile memory, processors such as microprocessors and digital signal processors, computational entities such as operating systems, drivers, graphical user interfaces, and applications programs, one or more interaction devices, such as a touch pad or screen, and/or control systems including feedback loops and control motors (e.g., feedback for sensing position and/or velocity, control motors for moving and/or adjusting components and/or quantities). A typical data processing system may be implemented utilizing any suitable commercially available components, such as those typically found in data computing/communication and/or network computing/communication systems.

[0124] The herein described subject matter sometimes illustrates different components included within, or connected with, different other components. It is to be understood that such depicted architectures are merely examples, and that in fact many other architectures may be implemented which achieve the same functionality. In a conceptual sense, any arrangement of components to achieve the same functionality is effectively "associated" such that the desired functionality may be achieved. Hence, any two components herein combined to achieve a particular functionality may be seen as "associated with" each other such that the desired

functionality is achieved, irrespective of architectures or intermedial components. Likewise, any two components so associated may also be viewed as being "operably connected", or "operably coupled", to each other to achieve the desired functionality, and any two components capable of being so associated may also be viewed as being "operably couplable" to each other to achieve the desired functionality. Specific examples of operably couplable include but are not limited to physically mateable and/or physically interacting components and/or wirelessly interactable and/or wirelessly interacting components and/or logically interacting and/or logically interactable components.

[0125] With respect to the use of substantially any plural and/or singular terms herein, those having skill in the art can translate from the plural to the singular and/or from the singular to the plural as is appropriate to the context and/or application. The various singular/plural permutations may be expressly set forth herein for sake of clarity.

[0126] It will be understood by those within the art that, in general, terms used herein, and especially in the appended claims (e.g., bodies of the appended claims) are generally intended as "open" terms (e.g., the term "including" should be interpreted as "including but not limited to," the term "having" should be interpreted as "having at least," the term "includes" should be interpreted as "includes but is not limited to," etc.). It will be further understood by those within the art that if a specific number of an introduced claim recitation is intended, such an intent will be explicitly recited in the claim, and in the absence of such recitation no such intent is present. For example, where only one item is intended, the term "single" or similar language may be used. As an aid to understanding, the following appended claims and/or the descriptions herein may include usage of the introductory phrases "at least one" and "one or more" to introduce claim recitations. However, the use of such phrases should not be construed to imply that the introduction of a claim recitation by the indefinite articles "a" or "an" limits any particular claim including such introduced claim recitation to embodiments including only one such recitation, even when the same claim includes the introductory phrases "one or more" or "at least one" and indefinite articles such as "a" or "an" (e.g., "a" and/or "an" should be interpreted to mean "at least one" or "one or more"). The same holds true for the use of definite articles used to introduce claim recitations. In addition, even if a specific number of an introduced claim recitation is explicitly recited, those skilled in the art will recognize that such recitation should be interpreted to mean at least the recited number (e.g., the bare recitation of "two recitations," without other modifiers, means at least two recitations, or two or more recitations). Furthermore, in those instances where a convention analogous to "at least one of

A, B, and C, etc." is used, in general such a construction is intended in the sense one having skill in the art would understand the convention (e.g., "a system having at least one of A, B, and C" would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). In those instances where a convention analogous to "at least one of A, B, or C, etc." is used, in general such a construction is intended in the sense one having skill in the art would understand the convention (e.g., "a system having at least one of A, B, or C" would include but not be limited to systems that have A alone, B alone, C alone, A and B together, A and C together, B and C together, and/or A, B, and C together, etc.). It will be further understood by those within the art that virtually any disjunctive word and/or phrase presenting two or more alternative terms, whether in the description, claims, or drawings, should be understood to contemplate the possibilities of including one of the terms, either of the terms, or both terms. For example, the phrase "A or B" will be understood to include the possibilities of "A" or "B" or "A and B." Further, the terms "any of" followed by a listing of a plurality of items and/or a plurality of categories of items, as used herein, are intended to include "any of," "any combination of," "any multiple of," and/or "any combination of multiples of" the items and/or the categories of items, individually or in conjunction with other items and/or other categories of items. Moreover, as used herein, the term "set" is intended to include any number of items, including zero. Additionally, as used herein, the term "number" is intended to include any number, including zero. And the term "multiple", as used herein, is intended to be synonymous with "a plurality".

[0127] In addition, where features or aspects of the disclosure are described in terms of Markush groups, those skilled in the art will recognize that the disclosure is also thereby described in terms of any individual member or subgroup of members of the Markush group.

[0128] As will be understood by one skilled in the art, for any and all purposes, such as in terms of providing a written description, all ranges disclosed herein also encompass any and all possible subranges and combinations of subranges thereof. Any listed range can be easily recognized as sufficiently describing and enabling the same range being broken down into at least equal halves, thirds, quarters, fifths, tenths, etc. As a non-limiting example, each range discussed herein may be readily broken down into a lower third, middle third and upper third, etc. As will also be understood by one skilled in the art all language such as "up to," "at least," "greater than," "less than," and the like includes the number recited and refers to ranges which can be subsequently broken down into subranges as discussed above. Finally, as will be

understood by one skilled in the art, a range includes each individual member. Thus, for example, a group having 1-3 cells refers to groups having 1, 2, or 3 cells. Similarly, a group having 1-5 cells refers to groups having 1, 2, 3, 4, or 5 cells, and so forth.

[0129] Moreover, the claims should not be read as limited to the provided order or elements unless stated to that effect. In addition, use of the terms "means for" in any claim is intended to invoke 35 U.S.C. §112, ¶ 6 or means-plus-function claim format, and any claim without the terms "means for" is not so intended.

ANNEXES

First example of a data wrapper of the first embodiment:

IntermediateDataWrapper:

```
Struct IntermediateDataWrapper {
Int      Model_RefID;              # Model Identifier
bit      ChangeOfSplitPoint;       # 1 if the splitpoint is changed from the previous
         message
Int      SplitPointID;             # Split point used by the UE
Int      NextSplitpointID;         # Next split point used by the UE (optional)
Int      IntermediateDataLength;   # Length of Intermediate Data
Dim      IntermediateDataDim;      # Array dimension of intermediate data
byte     EncodingMethod;           # Indicate if data are compressed and by which
algorithm
int      SequenceNumber;           # Sequential identifier of input data
Int      Dropped indication ;      # 0: sequence SequenceNumber  is processed ;
                                   # 1: sequence SequenceNumber  is dropped
double   TimeStamp;                # Timestamp of the message
Byte     IntermediateData[]        # Empty if Dropped indication is set to 1
};
```

Second example of a data wrapper of the first embodiment, based on JSON structured message:

{

"Model_RefID": < Model Identifier>,

"ChangeOfSplitPoint": < 1 if the splitpoint is changed from the previous message, else 0>,

"SplitPointID":" < Split point used by the UE>,

"NextSplitpointID": < Next split point used by the UE (optional)>,

"IntermediateDataLength": <Length of Intermediate Data>,

"IntermediateDataDim": < Array dimension of intermediate data>,

"EncodingMethod": <Indicate if data are compressed and by which algorithm>,

"SequenceNumber": < Sequential identifier of input data >,

"Dropped indication": < 0: sequence SequenceNumber is processed, 1: sequence SequenceNumber is dropped >,

"TimeStamp": < Timestamp of the message>,

"IntermediateData": < Intermediate Data, Empty if Dropped indication is set to 1>

}

The JSON format can offer a better resilience to the message content evolution compared to a C like structure as it is possible to add new fields without consequences on the receiver. If the receiver is compliant with an older JSON message structure, it will simply ignore the new field; if the receiver is aware of the new message structure it will correctly manage the new field.

First example of a data wrapper of the second embodiment:

IntermediateDataWrapper:

Struct IntermediateDataWrapper {

```
Int      Model_RefID;              # Model Identifier
bit      ChangeOfSplitPoint;       # 1 if the split point is changed from the previous
         message
Int      SplitPointID;             # Split point used by the UE
Int      NextSplitpointID;         # Next split point used by the UE (optional)
Int      IntermediateDataLength;   # Length of Intermediate Data
Dim      IntermediateDataDim;      # Array dimension of intermediate data
byte     EncodingMethod;           # Indicate if data are compressed and by which algorithm
int      SequenceNumber;           # Sequential identifier of input data
Int      NbSequenceDropped;        # Number of sequence dropped
List     Dropped Sequence list;    # List of sequences dropped
double   TimeStamp;                # Timestamp of the intermediate data
Byte     IntermediateData[]        # Intermediate Data
   };
```

In this example, the NbSequenceDropped field indicates the number of sequences dropped since the previous message.

Second example of a data wrapper of the second embodiment, based on JSON structured message:

```
{
      "Model_RefID":              < Model Identifier>,
      "ChangeOfSplitPoint":       < 1 if the splitpoint is changed from the previous
message, else 0>,
      "SplitPointID":"            < Split point used by the UE>,
      "NextSplitpointID":         < Next split point used by the UE (optional)>,
      "IntermediateDataLength":   <Length of Intermediate Data>,
```

"IntermediateDataDim": " < Array dimension of intermediate data>,

"EncodingMethod": <Indicate if data are compressed and by which algorithm>,

"SequenceNumber": < Sequential identifier of input data >,

"NbSequenceDropped": <Number of sequence dropped>,

"Dropped Sequence list": <List of sequences dropped>,

"TimeStamp": < Timestamp of the intermediate data>,

"IntermediateData": < Intermediate Data>

}

CLAIMS

What is claimed is:

1. A first device comprising at least one hardware processor configured to:

using a machine learning model with a plurality of layers, iteratively process a number of parts of sequentially organized data up to an intermediate layer of the machine learning model to obtain corresponding intermediate data;

determine not to process a subsequent part of the sequentially organized data; and

transmit to a second device information indicating that the subsequent part has been dropped.

2. The first device of claim 1, wherein:

the information indicating that the subsequent part has been dropped is transmitted instead of corresponding intermediate data for the subsequent part.

3. The first device of claim 1, wherein:

the information indicating that the subsequent part has been dropped is transmitted with corresponding intermediate data for at least one part following the subsequent part.

4. The first device of claim 1, wherein the sequentially organized data is a series of video frames.

5. The first device of claim 1, wherein the sequentially organized data is representative of audio or haptic data.

6. The first device of claim 1, wherein the first device determines not to process the subsequent part upon a change of processing to a different intermediate layer.

7. The first device of claim 1, wherein the intermediate data is output for further processing, by the second device using following layers of the machine learning model, to obtain at least one of further intermediate data and a result.

8. A method performed in a first device and comprising:

iteratively processing, using a machine learning model with a plurality of layers, a number of parts of sequentially organized data up to an intermediate layer of the machine learning model to obtain corresponding intermediate data;

determining not to process a subsequent part of the sequentially organized data; and

transmitting to a second device information indicating that the subsequent part has been dropped.

9. The method of claim 8, wherein:

the information indicating that the subsequent part has been dropped is transmitted instead of corresponding intermediate data for the subsequent part.

10. The method of claim 8, wherein:

the information indicating that the subsequent part has been dropped is transmitted with corresponding intermediate data for at least one part following the subsequent part.

11. The method of claim 8, wherein the sequentially organized data is a series of video frames.

12. The method of claim 8, wherein the sequentially organized data is representative of audio or haptic data.

13. The method of claim 8, wherein the first device determines not to process the subsequent part upon a change of processing to a different intermediate layer.

14. The method of claim 8, further comprising outputting the intermediate data for further processing, by the second device using following layers of the machine learning model, to obtain at least one of further intermediate data and a result.

15. A non-transitory computer readable medium storing program code instructions that, when executed by a processor, implement a method comprising:

iteratively processing, using a machine learning model with a plurality of layers, a number of parts of sequentially organized data up to an intermediate layer of the machine learning model to obtain corresponding intermediate data;

determining not to process a subsequent part of the sequentially organized data; and

transmitting to a second device information indicating that the subsequent part has been dropped.

5

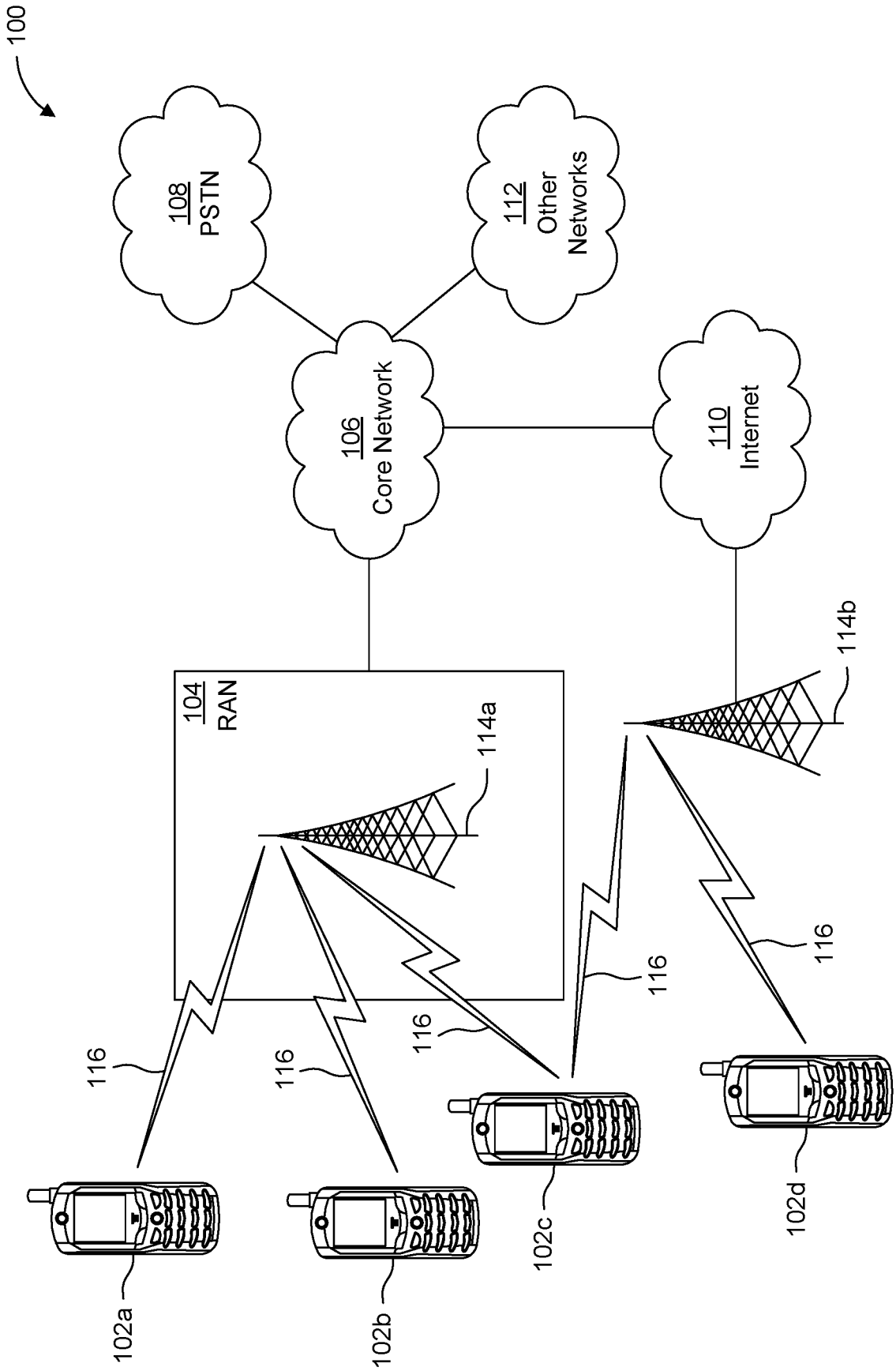
100
108 PSTN
112 Other Networks
106 Core Network
110 Internet
104 RAN
114a
114b
116
116
116
116
102a
102b
102c
102d

FIG. 1A

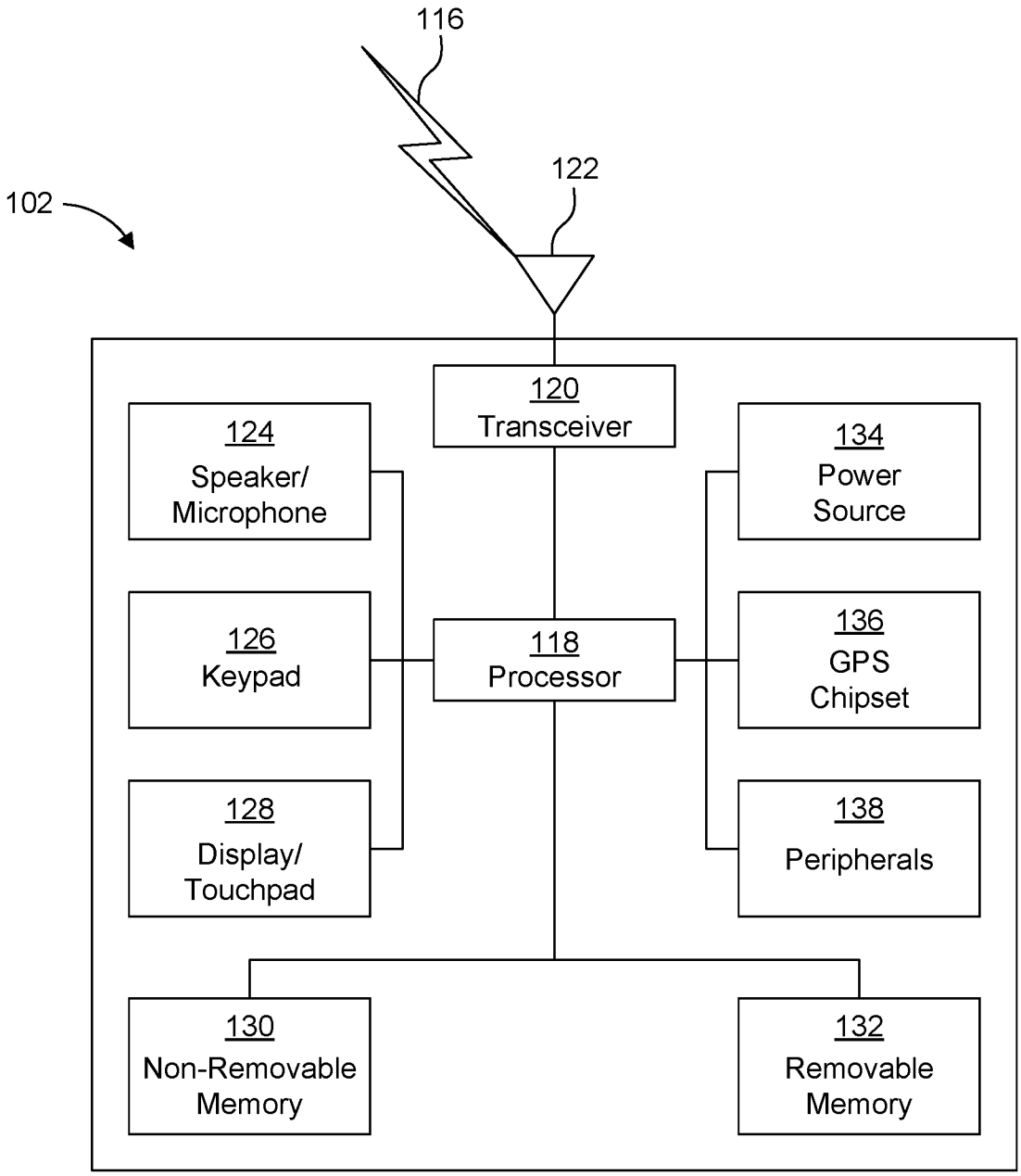
116
122
102
120
Transceiver
124
Speaker/
Microphone
134
Power
Source
126
Keypad
118
Processor
136
GPS
Chipset
128
Display/
Touchpad
138
Peripherals
130
Non-Removable
Memory
132
Removable
Memory

FIG. 1B

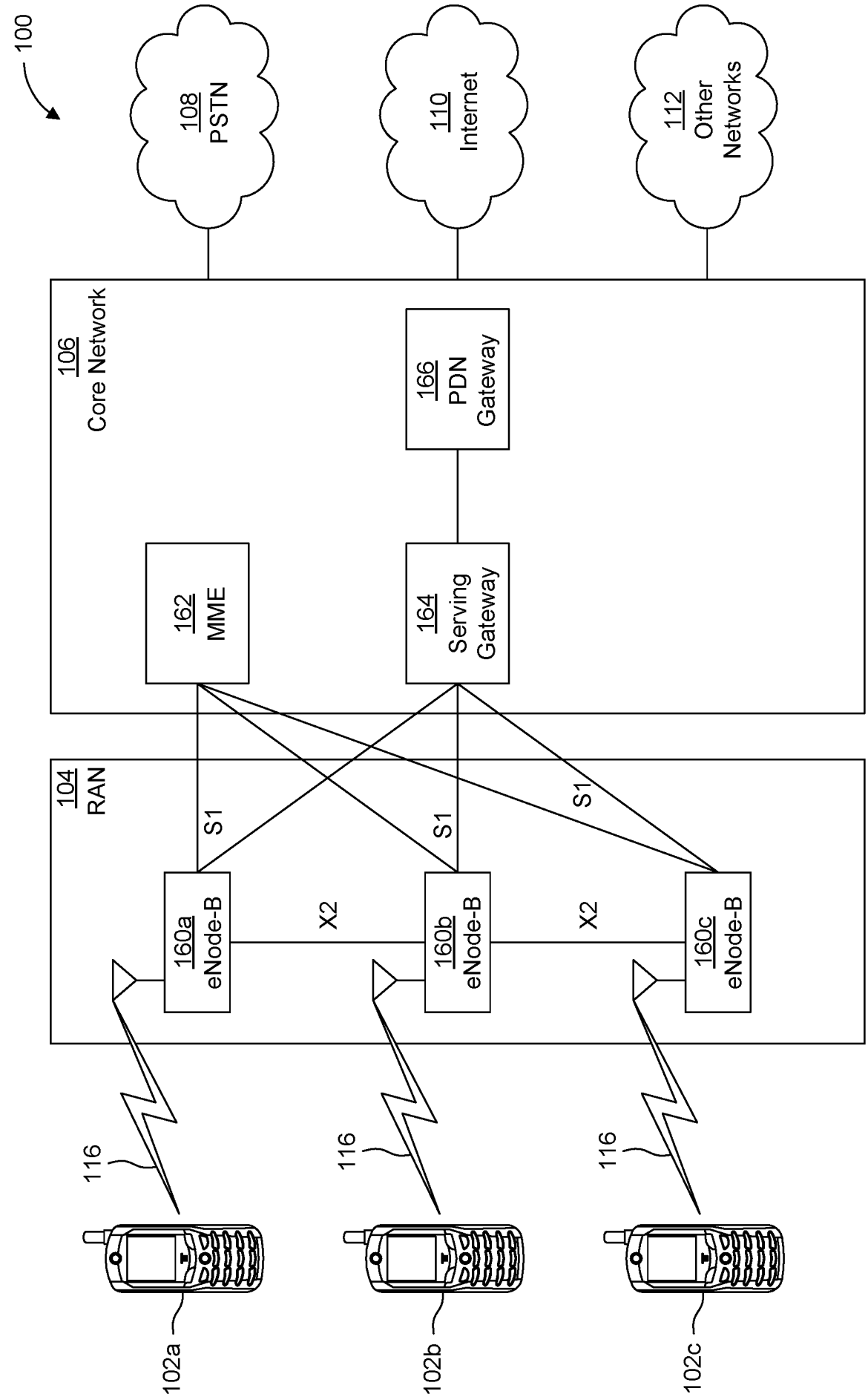
100
108 PSTN
110 Internet
112 Other Networks
106 Core Network
166 PDN Gateway
162 MME
164 Serving Gateway
104 RAN
S1
S1
S1
160a eNode-B
160b eNode-B
160c eNode-B
X2
X2
116
116
116
102a
102b
102c

FIG. 1C

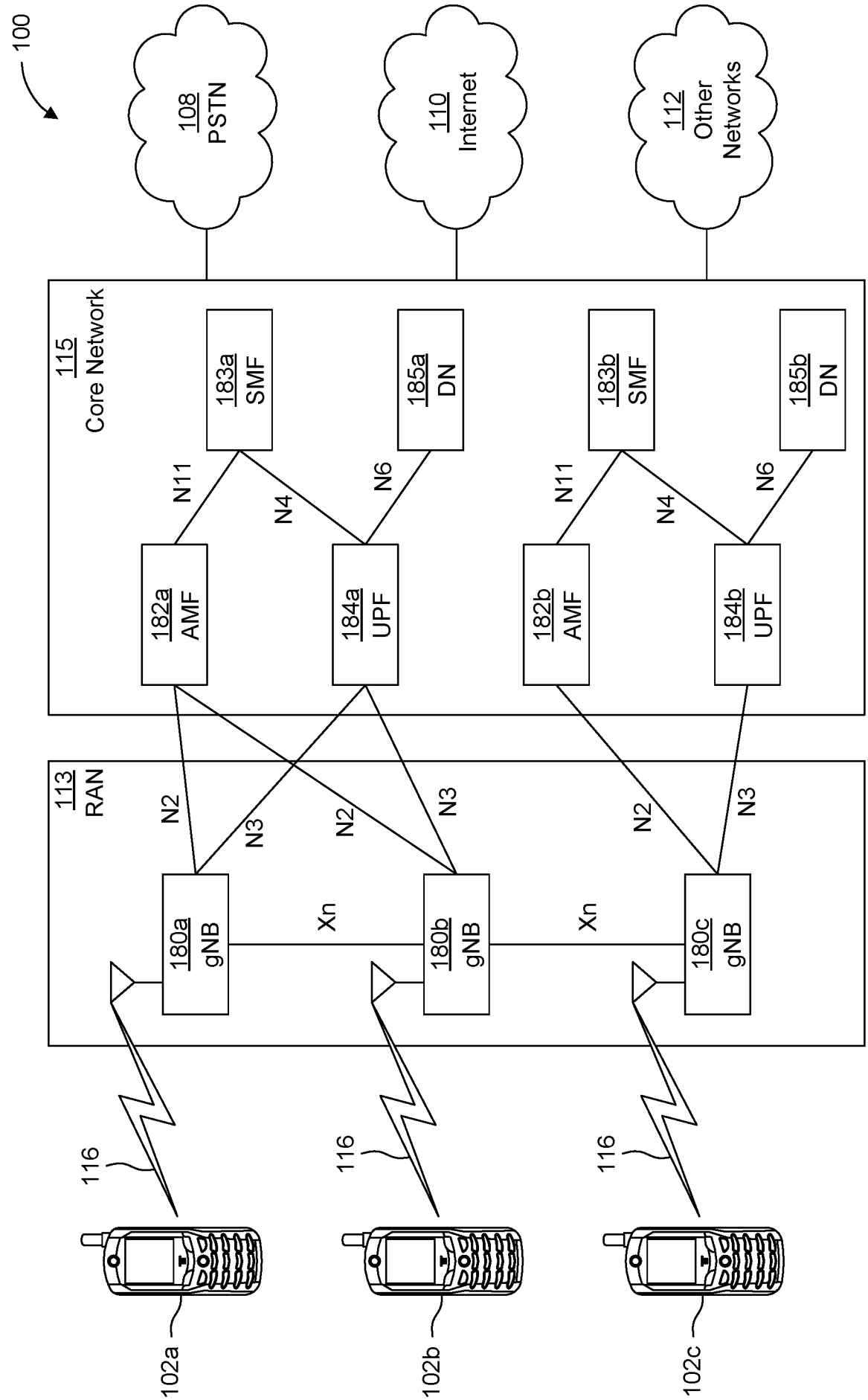
100
108 PSTN
110 Internet
112 Other Networks
115 Core Network
183a SMF
185a DN
183b SMF
185b DN
N11
N4
N6
N11
N4
N6
182a AMF
184a UPF
182b AMF
184b UPF
113 RAN
N2
N3
N2
N3
N2
N3
180a gNB
Xn
180b gNB
Xn
180c gNB
116
116
116
102a
102b
102c

FIG. 1D

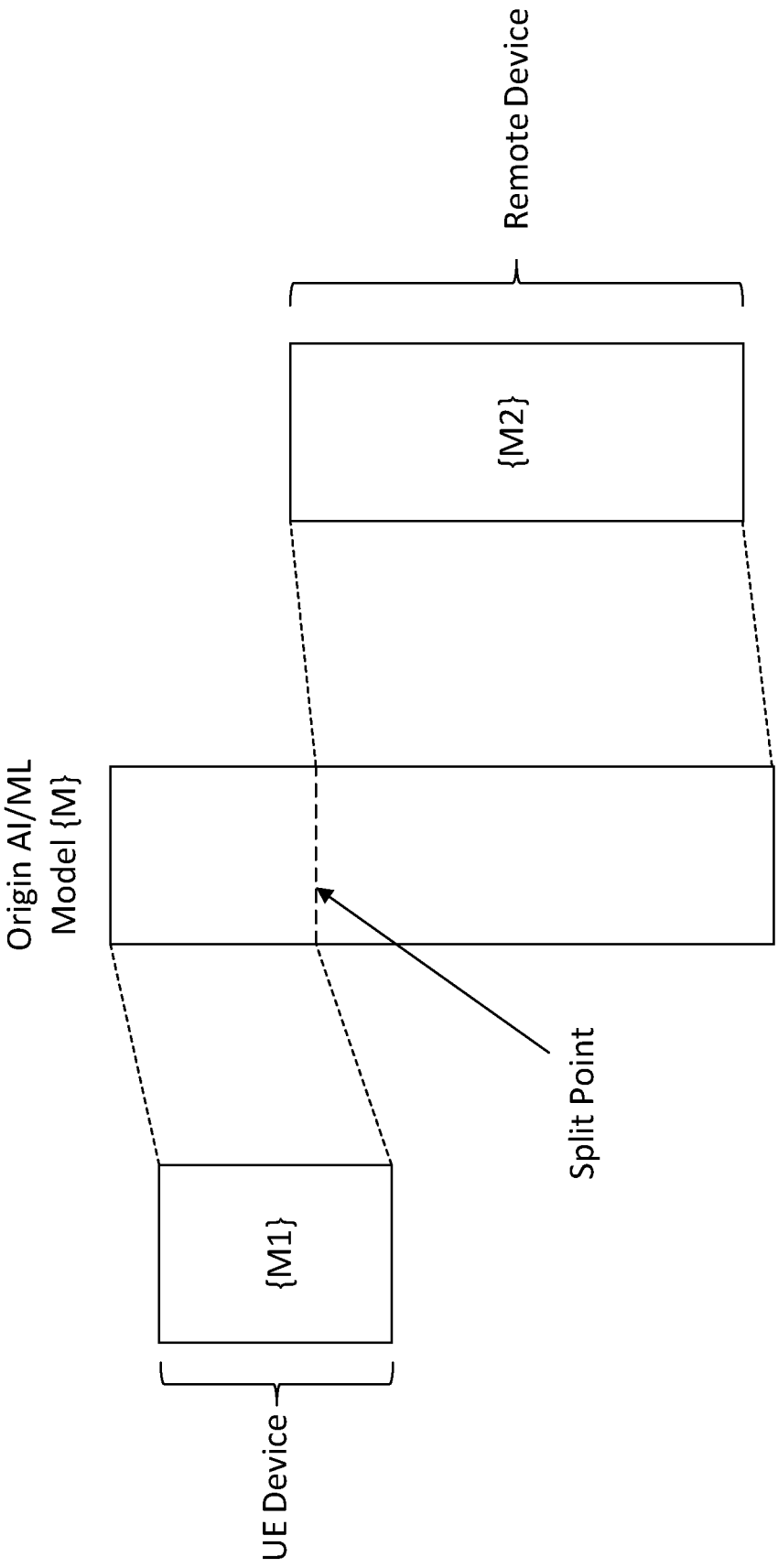
Remote Device
{M2}
Origin AI/ML Model {M}
Split Point
{M1}
UE Device

FIG. 2

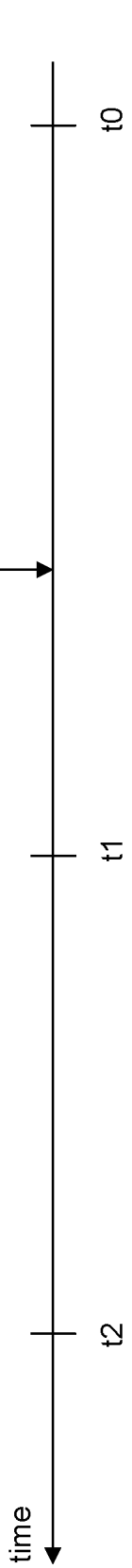
Intermediate Data of Frame i+2
t0
Change of Split Point From L4 to L3
Frame i+3 Dropped Message
t1
Intermediate Data of Frame i+4
t2
time

FIG. 3

Intermediate Data of Frame i+2
t0
Change of Split Point from L4 to L3
Intermediate Data of Frame i+4; Frame i+3 Dropped Information
t1
time

FIG. 4

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2024/073793

A. CLASSIFICATION OF SUBJECT MATTER
INV. G06N3/098
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)
G06N

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

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	STEPHANE ONNO ET AL: "[FS_AI4Media] Metadata for Split inferencing", 3GPP DRAFT; S4-231567; TYPE DISCUSSION, 3RD GENERATION PARTNERSHIP PROJECT (3GPP), MOBILE COMPETENCE CENTRE ; 650, ROUTE DES LUCIOLES ; F-06921 SOPHIA-ANTIPOLIS CEDEX ; FRANCE , vol. SA WG4, no. Goteborg, SE; 20230821 - 20230825 25 August 2023 (2023-08-25), XP052451591, Retrieved from the Internet: URL:https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_125_Gotheneburg/Docs/S4-231567.zip S4-231567 FS_AI4Media Metadata for Split inferencing.docx [retrieved on 2023-08-25] Table on Section 6.4.2; page first - page third	1-15

-/--

[x] Further documents are listed in the continuation of Box C.

[ ] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier application or patent but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance;; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance;; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
14 November 2024	22/11/2024

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Fonseca dos Santos

2

Form PCT/ISA/210 (second sheet) (April 2005)

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
	Section 1; page first -----	
A	ERIC YIP ET AL: "[FS_AI4Media] Permanent Document v0.8", 3GPP DRAFT; S4-231011; TYPE DISCUSSION, 3RD GENERATION PARTNERSHIP PROJECT (3GPP), MOBILE COMPETENCE CENTRE ; 650, ROUTE DES LUCIOLES ; F-06921 SOPHIA-ANTIPOLIS CEDEX ; FRANCE , vol. SA WG4, no. Berlin, DE; 20230522 - 20230526 26 May 2023 (2023-05-26), XP052378545, Retrieved from the Internet: URL:https://www.3gpp.org/ftp/TSG_SA/WG4_CODEC/TSGS4_124_Berlin/Docs/S4-231011.zip S4-231011 [FS_AI4Media] PD v0.8.docx [retrieved on 2023-05-26] "Intermediate Data" on the Table of the 55th page; page 55th page 13th; figures 5.1-1 -----	1-15