



- (51) International Patent Classification:
H04N 19/70 (2014.01) *G06N 3/045* (2023.01)

(21) International Application Number:
PCT/US2024/037509

(22) International Filing Date:
11 July 2024 (11.07.2024)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
63/527,567 18 July 2023 (18.07.2023) US

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

(72) Inventors: **ONNO, Stephane**; 845A, avenue des Champs
Blancs, 35510 Cesson-Sevigne (FR). **FILOCHE, Thier-**
ry; 845A, avenue des Champs Blancs, 35510 Cesson-Sevi-
- gne (FR). **QUINQUIS, Cyril**; 845A, avenue des Champs
Blancs, 35510 Cesson-Sevigne (FR). **RACAPE, Fabien**;
4410 El Camino Real, Los Altos, CA 94022 (US).

(74) Agent: **ROCCIA, Vincent, J.** et al.; Condo Roccia Kop-
tiw LLP, 1800 JFK Boulevard, Suite 1700, Philadelphia, PA
19103 (US).

(81) Designated States (*unless otherwise indicated, for every
kind of national protection available*): AE, AG, AL, AM,
AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ,
CA, CH, CL, CN, CO, CR, CU, 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.

(54) Title: TENSOR INFORMATION FOR INTERMEDIATE DATA

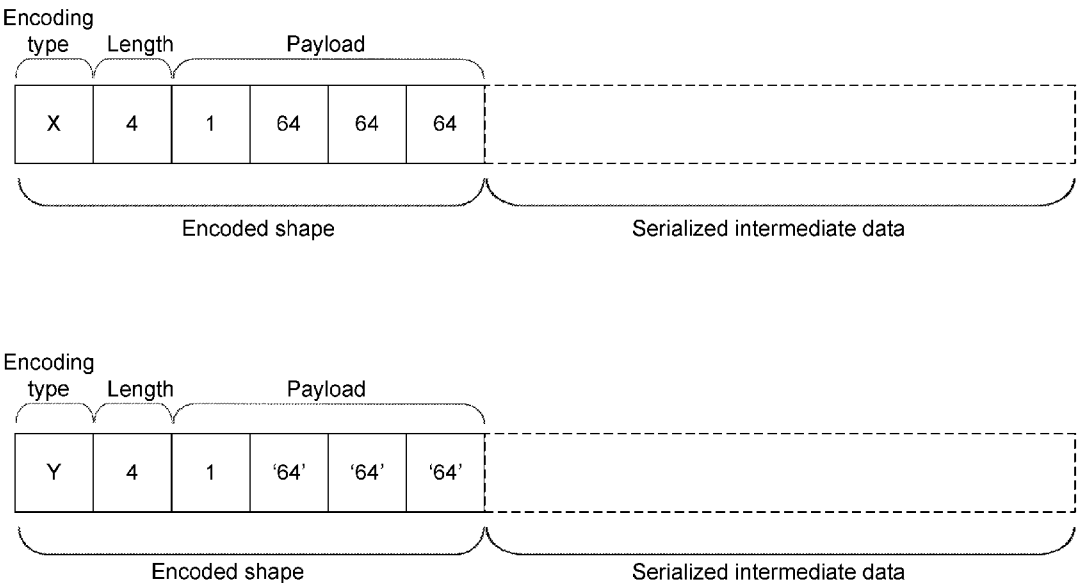


FIG. 7

(57) Abstract: System, methods, and instrumentalities are disclosed for encoding and decoding tensor information for intermediate data. An example device for video encoding may determine tensor metadata associated with intermediate data; encode the intermediate data and the tensor metadata associated with the intermediate data; and send the encoded intermediate data and the encoded tensor metadata to a second device. An example device for video decoding may receive encoded intermediate data and encoded tensor metadata from a second device; decode the encoded intermediate data and the encoded tensor metadata; and reconstruct a tensor based on the intermediate data and the tensor metadata.

(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))*

TENSOR INFORMATION FOR INTERMEDIATE DATA

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Patent Application No. 63/527,567, filed July 18, 2023, the contents of which are hereby incorporated by reference herein.

BACKGROUND

[0002] Video coding systems may be used to compress digital video signals, e.g., to reduce the storage and/or transmission bandwidth needed for such signals. Video coding systems may include, for example, block-based, wavelet-based, and/or object-based systems.

SUMMARY

[0003] System, methods, and instrumentalities are disclosed for encoding and decoding tensor information for intermediate data.

[0004] An example device (e.g., for video encoding) may be a first device. The first device may determine tensor metadata associated with intermediate data. The first device may encode the intermediate data and the tensor metadata associated with the intermediate data. The first device may send the encoded intermediate data and the encoded tensor metadata to a second device.

[0005] The first device may be associated with a first portion of an artificial intelligence (AI) model. The first portion may be an initial portion of the AI model. The second device may be associated with a second portion of the AI model. The second portion may be a latter portion of the AI model. The first device may receive the intermediate data as an output of the first portion of the AI model.

[0006] The intermediate data may be first intermediate data. The tensor metadata may be first tensor metadata indicative of a first tensor shape. The first device may determine a second tensor shape associated with second intermediate data. The second tensor shape may be different from the first tensor shape. The first device may encode the second intermediate data and second tensor metadata. The second tensor metadata

may be indicative of the second tensor shape. The first device may send the encoded second intermediate data and the encoded second tensor metadata to the second device.

[0007] The intermediate data may be first intermediate data. The tensor metadata may be indicative of a tensor shape. The first device may determine that second intermediate data is associated with the tensor shape. The first device may encode the second intermediate data and an indication that the tensor metadata has not been included because the tensor shape has not changed. The first device may send the encoded second intermediate data and the encoded indication to the second device.

[0008] The tensor metadata may be indicative of at least one of: a tensor shape, tensor structure information, a data type, or an encoding algorithm.

[0009] The first device may encode the intermediate data by transforming a multi-dimensional tensor into a one-dimensional tensor, and encoding the one-dimensional tensor in accordance with a binary structure.

[0010] The first device may send the encoded intermediate data and the encoded tensor metadata to the second device by sending the encoded intermediate data in a first band associated with a data channel; and sending the encoded tensor metadata in a second band associated with a control channel.

[0011] The first device may encode the intermediate data and the tensor metadata associated with the intermediate data by generating an encoded bitstream. The encoded bitstream may include the encoded intermediate data and the encoded tensor metadata. The first device may send the encoded intermediate data and the encoded tensor metadata to the second device by sending the encoded bitstream to the second device.

[0012] The first device may be a wireless transmit/receive unit (WTRU) or a network entity.

[0013] An example device (e.g., for video decoding) may be a first device. The first device may receive encoded intermediate data and encoded tensor metadata from a second device. The first device may decode the encoded intermediate data and the encoded tensor metadata. The first device may reconstruct a tensor based on the intermediate data and the tensor metadata.

[0014] The second device may be associated with a first portion of an artificial intelligence (AI) model. The first portion may be an initial portion of the AI model. The first device may be associated with a second portion of the AI model. The second portion may be a latter portion of the AI model. The first device may input the reconstructed tensor into the second portion of the AI model.

[0015] The intermediate data may be first intermediate data. The tensor metadata may be first tensor metadata indicative of a first tensor shape, the tensor is a first tensor. The first device may receive encoded

second intermediate data and encoded second tensor metadata from the second device. The first device may decode the encoded second intermediate data and the encoded second tensor metadata. The second tensor metadata may be indicative of a second tensor shape. The first device may reconstruct a second tensor based on the second intermediate data and the second tensor metadata.

[0016] The intermediate data may be first intermediate data. The tensor metadata may be indicative of a tensor shape. The first device may receive encoded second intermediate data and an encoded indication that the tensor metadata has not been included because the tensor shape has not changed. The first device may decode the encoded second intermediate data and the indication. The first device may determine that the second intermediate data is associated with the tensor shape based on the indication.

[0017] The tensor metadata may be indicative of at least one of: a tensor shape, tensor structure information, a data type, or an encoding algorithm.

[0018] The first device may decode the encoded intermediate data by transforming a one-dimensional tensor into a multi-dimensional tensor, and decoding the multi-dimensional tensor in accordance with a binary structure.

[0019] The first device may receive the encoded intermediate data and the encoded tensor metadata from the second device by receiving the encoded intermediate data in a first band associated with a data channel; and receiving the encoded tensor metadata in a second band associated with a control channel.

[0020] The first device may receive the encoded intermediate data and the encoded tensor metadata from the second device by receiving an encoded bitstream from the second device. The encoded bitstream may include the encoded intermediate data and the encoded tensor metadata. The first device may decode the encoded intermediate data and the encoded tensor metadata by decoding the encoded bitstream.

[0021] The first device may decode the encoded intermediate data and the encoded tensor metadata by decoding one or more of: the encoded tensor metadata according to a tensor shape structure or algorithm to obtain the tensor shape; or the encoded intermediate data according to a bitstream structure to obtain a series of intermediate data bytes.

[0022] The first device may reconstruct the tensor based on the intermediate data and the tensor metadata by converting binary data, in the intermediate data, to tensor data, and reshaping the tensor to an original tensor shape based on the tensor data and the tensor metadata.

[0023] The first device may be a wireless transmit/receive unit (WTRU) or a network entity.

[0024] An example device for encoding intermediate data (e.g., comprising tensors) may receive intermediate data generated by an initial part of an artificial intelligence (AI) model. The device may encode a tensor shape associated with the intermediate data as a tensor shape metadata in accordance with a tensor shape structure or an encoding algorithm. The device may encode the intermediate data as an encoded intermediate data bitstream in accordance with a bitstream structure. The device may send the tensor shape metadata and the encoded intermediate data bitstream to a second device.

[0025] Encoding the intermediate data may involve transforming a multi-dimensional tensor into a one-dimensional tensor, and encoding the one-dimensional tensor in accordance with a binary structure.

[0026] An example device for decoding intermediate data may receive, from a second device, a data bitstream comprising encoded intermediate data and encoded tensor shape metadata associated with the encoded intermediate data. The device may decode the encoded tensor shape metadata based on a tensor shape structure or an encoding algorithm to determine a tensor shape. The device may decode the encoded intermediate data based on a bitstream structure to determine a series of intermediate data bytes. The device may reconstruct a tensor from the series of intermediate data bytes based on the tensor shape. The device may input the reconstructed tensor to a last part of an artificial intelligence (AI) model.

[0027] The tensor shape structure and the bitstream structure may be preconfigured. The encoding algorithm may be a binary structure algorithm, a binary serialization algorithm, or a textual serialization algorithm.

BRIEF DESCRIPTION OF THE DRAWINGS

[0028] FIG. 1A is a system diagram illustrating an example communications system in which one or more disclosed embodiments may be implemented.

[0029] FIG. 1B is a system diagram illustrating an example wireless transmit/receive unit (WTRU) that may be used within the communications system illustrated in FIG. 1A according to an embodiment.

[0030] FIG. 1C is a system diagram illustrating an example radio access network (RAN) and an example core network (CN) that may be used within the communications system illustrated in FIG. 1A according to an embodiment.

[0031] FIG. 1D is a system diagram illustrating a further example RAN and a further example CN that may be used within the communications system illustrated in FIG. 1A according to an embodiment.

[0032] FIG. 2 illustrates an example video encoder.

[0033] FIG. 3 illustrates an example video decoder.

[0034] FIG. 4 illustrates an example of a system in which various aspects and examples may be implemented.

[0035] FIG. 5 illustrates a split inference between a WTRU and a network, with the media data source originating from the WTRU.

[0036] FIG. 6 illustrates a split inference between the WTRU and the network, with the media data source originating from the network.

[0037] FIG. 7 illustrates an example encoded tensor shape.

[0038] FIG. 8 illustrates an example change mode for encoding a tensor shape.

[0039] FIG. 9 illustrates an example split inference between a WTRU and a network, in which the media data source and tensor shape metadata come from the WTRU.

[0040] FIG. 10 illustrates an example split inference between a WTRU and a network, in which the media data source and tensor shape metadata come from the network.

[0041] FIG. 11 illustrates examples of the intermediate data access and delivery functions.

DETAILED DESCRIPTION

[0042] A more detailed understanding may be had from the following description, given by way of example in conjunction with the accompanying drawings.

[0043] FIG. 1A is a diagram illustrating an example communications system 100 in which one or more disclosed embodiments may be implemented. The communications system 100 may be a multiple access system that provides content, such as voice, data, video, messaging, broadcast, etc., to multiple wireless users. The communications system 100 may enable multiple wireless users to access such content through the sharing of system resources, including wireless bandwidth. For example, the communications systems 100 may employ one or more channel access methods, such as code division multiple access (CDMA), time division multiple access (TDMA), frequency division multiple access (FDMA), orthogonal FDMA (OFDMA), single-carrier FDMA (SC-FDMA), zero-tail unique-word DFT-Spread OFDM (ZT UW DTS-s OFDM), unique word OFDM (UW-OFDM), resource block-filtered OFDM, filter bank multicarrier (FBMC), and the like.

[0044] As shown in FIG. 1A, the communications system 100 may include wireless transmit/receive units (WTRUs) 102a, 102b, 102c, 102d, a RAN 104/113, a CN 106/115, a public switched telephone network (PSTN) 108, the Internet 110, and other networks 112, though it will be appreciated that the disclosed embodiments contemplate any number of WTRUs, base stations, networks, and/or network elements. Each of

the WTRUs 102a, 102b, 102c, 102d may be any type of device configured to operate and/or communicate in a wireless environment. By way of example, the WTRUs 102a, 102b, 102c, 102d, any of which may be referred to as a “station” and/or a “STA,” may be configured to transmit and/or receive wireless signals and may include a user equipment (UE), a mobile station, a fixed or mobile subscriber unit, a subscription-based unit, a pager, a cellular telephone, a personal digital assistant (PDA), a smartphone, a laptop, a netbook, a personal computer, a wireless sensor, a hotspot or Mi-Fi device, an Internet of Things (IoT) device, a watch or other wearable, a head-mounted display (HMD), a vehicle, a drone, a medical device and applications (e.g., remote surgery), an industrial device and applications (e.g., a robot and/or other wireless devices operating in an industrial and/or an automated processing chain contexts), a consumer electronics device, a device operating on commercial and/or industrial wireless networks, and the like. Any of the WTRUs 102a, 102b, 102c and 102d may be interchangeably referred to as a UE.

[0045] The communications systems 100 may also include a base station 114a and/or a base station 114b. Each of the base stations 114a, 114b may be any type of device configured to wirelessly interface with at least one of the WTRUs 102a, 102b, 102c, 102d to facilitate access to one or more communication networks, such as the CN 106/115, the Internet 110, and/or the other networks 112. By way of example, the base stations 114a, 114b may be a base transceiver station (BTS), a Node-B, an eNode B, a Home Node B, a Home eNode B, a gNB, a NR NodeB, a site controller, an access point (AP), a wireless router, and the like. While the base stations 114a, 114b are each depicted as a single element, it will be appreciated that the base stations 114a, 114b may include any number of interconnected base stations and/or network elements.

[0046] The base station 114a may be part of the RAN 104/113, which may also include other base stations and/or network elements (not shown), such as a base station controller (BSC), a radio network controller (RNC), relay nodes, etc. The base station 114a and/or the base station 114b may be configured to transmit and/or receive wireless signals on one or more carrier frequencies, which may be referred to as a cell (not shown). These frequencies may be in licensed spectrum, unlicensed spectrum, or a combination of licensed and unlicensed spectrum. A cell may provide coverage for a wireless service to a specific geographical area that may be relatively fixed or that may change over time. The cell may further be divided into cell sectors. For example, the cell associated with the base station 114a may be divided into three sectors. Thus, in one embodiment, the base station 114a may include three transceivers, i.e., one for each sector of the cell. In an embodiment, the base station 114a may employ multiple-input multiple output (MIMO) technology and may utilize multiple transceivers for each sector of the cell. For example, beamforming may be used to transmit and/or receive signals in desired spatial directions.

[0047] 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).

[0048] More specifically, as noted above, the communications system 100 may be a multiple access system and may employ one or more channel access schemes, such as CDMA, TDMA, FDMA, OFDMA, SC-FDMA, and the like. For example, the base station 114a in the RAN 104/113 and the WTRUs 102a, 102b, 102c may implement a radio technology such as Universal Mobile Telecommunications System (UMTS) Terrestrial Radio Access (UTRA), which may establish the air interface 115/116/117 using wideband CDMA (WCDMA). WCDMA may include communication protocols such as High-Speed Packet Access (HSPA) and/or Evolved HSPA (HSPA+). HSPA may include High-Speed Downlink (DL) Packet Access (HSDPA) and/or High-Speed UL Packet Access (HSUPA).

[0049] 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).

[0050] 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).

[0051] 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).

[0052] In other embodiments, the base station 114a and the WTRUs 102a, 102b, 102c may implement radio technologies such as IEEE 802.11 (i.e., Wireless Fidelity (WiFi)), IEEE 802.16 (i.e., Worldwide Interoperability for Microwave Access (WiMAX)), CDMA2000, CDMA2000 1X, CDMA2000 EV-DO, Interim Standard 2000 (IS-2000), Interim Standard 95 (IS-95), Interim Standard 856 (IS-856), Global System for Mobile communications (GSM), Enhanced Data rates for GSM Evolution (EDGE), GSM EDGE (GERAN), and the like.

[0053] The base station 114b in FIG. 1A may be a wireless router, Home Node B, Home eNode B, or access point, for example, and may utilize any suitable RAT for facilitating wireless connectivity in a localized area, such as a place of business, a home, a vehicle, a campus, an industrial facility, an air corridor (e.g., for

use by drones), a roadway, and the like. In one embodiment, the base station 114b and the WTRUs 102c, 102d may implement a radio technology such as IEEE 802.11 to establish a wireless local area network (WLAN). In an embodiment, the base station 114b and the WTRUs 102c, 102d may implement a radio technology such as IEEE 802.15 to establish a wireless personal area network (WPAN). In yet another embodiment, the base station 114b and the WTRUs 102c, 102d may utilize a cellular-based RAT (e.g., WCDMA, CDMA2000, GSM, LTE, LTE-A, LTE-A Pro, NR etc.) to establish a picocell or femtocell. As shown in FIG. 1A, the base station 114b may have a direct connection to the Internet 110. Thus, the base station 114b may not be required to access the Internet 110 via the CN 106/115.

[0054] The RAN 104/113 may be in communication with the CN 106/115, which may be any type of network configured to provide voice, data, applications, and/or voice over internet protocol (VoIP) services to one or more of the WTRUs 102a, 102b, 102c, 102d. The data may have varying quality of service (QoS) requirements, such as differing throughput requirements, latency requirements, error tolerance requirements, reliability requirements, data throughput requirements, mobility requirements, and the like. The CN 106/115 may provide call control, billing services, mobile location-based services, pre-paid calling, Internet connectivity, video distribution, etc., and/or perform high-level security functions, such as user authentication. Although not shown in FIG. 1A, it will be appreciated that the RAN 104/113 and/or the CN 106/115 may be in direct or indirect communication with other RANs that employ the same RAT as the RAN 104/113 or a different RAT. For example, in addition to being connected to the RAN 104/113, which may be utilizing a NR radio technology, the CN 106/115 may also be in communication with another RAN (not shown) employing a GSM, UMTS, CDMA 2000, WiMAX, E-UTRA, or WiFi radio technology.

[0055] The CN 106/115 may also serve as a gateway for the WTRUs 102a, 102b, 102c, 102d to access the PSTN 108, the Internet 110, and/or the other networks 112. The PSTN 108 may include circuit-switched telephone networks that provide plain old telephone service (POTS). The Internet 110 may include a global system of interconnected computer networks and devices that use common communication protocols, such as the transmission control protocol (TCP), user datagram protocol (UDP) and/or the internet protocol (IP) in the TCP/IP internet protocol suite. The networks 112 may include wired and/or wireless communications networks owned and/or operated by other service providers. For example, the networks 112 may include another CN connected to one or more RANs, which may employ the same RAT as the RAN 104/113 or a different RAT.

[0056] 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.

[0057] FIG. 1B is a system diagram illustrating an example WTRU 102. As shown in FIG. 1B, the WTRU 102 may include a processor 118, a transceiver 120, a transmit/receive element 122, a speaker/microphone 124, a keypad 126, a display/touchpad 128, non-removable memory 130, removable memory 132, a power source 134, a global positioning system (GPS) chipset 136, and/or other peripherals 138, among others. It will be appreciated that the WTRU 102 may include any sub-combination of the foregoing elements while remaining consistent with an embodiment.

[0058] The processor 118 may be a general purpose processor, a special purpose processor, a conventional processor, a digital signal processor (DSP), a plurality of microprocessors, one or more microprocessors in association with a DSP core, a controller, a microcontroller, Application Specific Integrated Circuits (ASICs), Field Programmable Gate Arrays (FPGAs) circuits, any other type of integrated circuit (IC), a state machine, and the like. The processor 118 may perform signal coding, data processing, power control, input/output processing, and/or any other functionality that enables the WTRU 102 to operate in a wireless environment. The processor 118 may be coupled to the transceiver 120, which may be coupled to the transmit/receive element 122. While FIG. 1B depicts the processor 118 and the transceiver 120 as separate components, it will be appreciated that the processor 118 and the transceiver 120 may be integrated together in an electronic package or chip.

[0059] The transmit/receive element 122 may be configured to transmit signals to, or receive signals from, a base station (e.g., the base station 114a) over the air interface 116. For example, in one embodiment, the transmit/receive element 122 may be an antenna configured to transmit and/or receive RF signals. In an embodiment, the transmit/receive element 122 may be an emitter/detector configured to transmit and/or receive IR, UV, or visible light signals, for example. In yet another embodiment, the transmit/receive element 122 may be configured to transmit and/or receive both RF and light signals. It will be appreciated that the transmit/receive element 122 may be configured to transmit and/or receive any combination of wireless signals.

[0060] Although the transmit/receive element 122 is depicted in FIG. 1B as a single element, the WTRU 102 may include any number of transmit/receive elements 122. More specifically, the WTRU 102 may employ MIMO technology. Thus, in one embodiment, the WTRU 102 may include two or more transmit/receive elements 122 (e.g., multiple antennas) for transmitting and receiving wireless signals over the air interface 116.

[0061] 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.

[0062] 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).

[0063] 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.

[0064] 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.

[0065] The processor 118 may further be coupled to other peripherals 138, which may include one or more software and/or hardware modules that provide additional features, functionality and/or wired or wireless connectivity. For example, the peripherals 138 may include an accelerometer, an e-compass, a satellite transceiver, a digital camera (for photographs and/or video), a universal serial bus (USB) port, a vibration device, a television transceiver, a hands free headset, a Bluetooth® module, a frequency modulated (FM) radio

unit, a digital music player, a media player, a video game player module, an Internet browser, a Virtual Reality and/or Augmented Reality (VR/AR) device, an activity tracker, and the like. The peripherals 138 may include one or more sensors, the sensors may be one or more of a gyroscope, an accelerometer, a hall effect sensor, a magnetometer, an orientation sensor, a proximity sensor, a temperature sensor, a time sensor; a geolocation sensor; an altimeter, a light sensor, a touch sensor, a magnetometer, a barometer, a gesture sensor, a biometric sensor, and/or a humidity sensor.

[0066] The WTRU 102 may include a full duplex radio for which transmission and reception of some or all of the signals (e.g., associated with particular subframes for both the UL (e.g., for transmission) and downlink (e.g., for reception) may be concurrent and/or simultaneous. The full duplex radio may include an interference management unit to reduce and or substantially eliminate self-interference via either hardware (e.g., a choke) or signal processing via a processor (e.g., a separate processor (not shown) or via processor 118). In an embodiment, the WTRU 102 may include a half-duplex radio for which transmission and reception of some or all of the signals (e.g., associated with particular subframes for either the UL (e.g., for transmission) or the downlink (e.g., for reception)).

[0067] FIG. 1C is a system diagram illustrating the RAN 104 and the CN 106 according to an embodiment. As noted above, the RAN 104 may employ an E-UTRA radio technology to communicate with the WTRUs 102a, 102b, 102c over the air interface 116. The RAN 104 may also be in communication with the CN 106.

[0068] The RAN 104 may include eNode-Bs 160a, 160b, 160c, though it will be appreciated that the RAN 104 may include any number of eNode-Bs while remaining consistent with an embodiment. The eNode-Bs 160a, 160b, 160c may each include one or more transceivers for communicating with the WTRUs 102a, 102b, 102c over the air interface 116. In one embodiment, the eNode-Bs 160a, 160b, 160c may implement MIMO technology. Thus, the eNode-B 160a, for example, may use multiple antennas to transmit wireless signals to, and/or receive wireless signals from, the WTRU 102a.

[0069] Each of the eNode-Bs 160a, 160b, 160c may be associated with a particular cell (not shown) and may be configured to handle radio resource management decisions, handover decisions, scheduling of users in the UL and/or DL, and the like. As shown in FIG. 1C, the eNode-Bs 160a, 160b, 160c may communicate with one another over an X2 interface.

[0070] The CN 106 shown in FIG. 1C may include a mobility management entity (MME) 162, a serving gateway (SGW) 164, and a packet data network (PDN) gateway (or PGW) 166. While each of the foregoing elements are depicted as part of the CN 106, it will be appreciated that any of these elements may be owned and/or operated by an entity other than the CN operator.

[0071] The MME 162 may be connected to each of the eNode-Bs 162a, 162b, 162c in the RAN 104 via an S1 interface and may serve as a control node. For example, the MME 162 may be responsible for authenticating users of the WTRUs 102a, 102b, 102c, bearer activation/deactivation, selecting a particular serving gateway during an initial attach of the WTRUs 102a, 102b, 102c, and the like. The MME 162 may provide a control plane function for switching between the RAN 104 and other RANs (not shown) that employ other radio technologies, such as GSM and/or WCDMA.

[0072] 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.

[0073] 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.

[0074] 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.

[0075] 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.

[0076] In representative embodiments, the other network 112 may be a WLAN.

[0077] A WLAN in Infrastructure Basic Service Set (BSS) mode may have an Access Point (AP) for the BSS and one or more stations (STAs) associated with the AP. The AP may have an access or an interface to a Distribution System (DS) or another type of wired/wireless network that carries traffic in to and/or out of the BSS. Traffic to STAs that originates from outside the BSS may arrive through the AP and may be delivered to the STAs. Traffic originating from STAs to destinations outside the BSS may be sent to the AP to be delivered to respective destinations. Traffic between STAs within the BSS may be sent through the AP, for example,

where the source STA may send traffic to the AP and the AP may deliver the traffic to the destination STA. The traffic between STAs within a BSS may be considered and/or referred to as peer-to-peer traffic. The peer-to-peer traffic may be sent between (e.g., directly between) the source and destination STAs with a direct link setup (DLS). In certain representative embodiments, the DLS may use an 802.11e DLS or an 802.11z tunneled DLS (TDLS). A WLAN using an Independent BSS (IBSS) mode may not have an AP, and the STAs (e.g., all of the STAs) within or using the IBSS may communicate directly with each other. The IBSS mode of communication may sometimes be referred to herein as an “ad-hoc” mode of communication.

[0078] When using the 802.11ac infrastructure mode of operation or a similar mode of operations, the AP may transmit a beacon on a fixed channel, such as a primary channel. The primary channel may be a fixed width (e.g., 20 MHz wide bandwidth) or a dynamically set width via signaling. The primary channel may be the operating channel of the BSS and may be used by the STAs to establish a connection with the AP. In certain representative embodiments, Carrier Sense Multiple Access with Collision Avoidance (CSMA/CA) may be implemented, for example in 802.11 systems. For CSMA/CA, the STAs (e.g., every STA), including the AP, may sense the primary channel. If the primary channel is sensed/detected and/or determined to be busy by a particular STA, the particular STA may back off. One STA (e.g., only one station) may transmit at any given time in a given BSS.

[0079] 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.

[0080] Very High Throughput (VHT) STAs may support 20MHz, 40 MHz, 80 MHz, and/or 160 MHz wide channels. The 40 MHz, and/or 80 MHz, channels may be formed by combining contiguous 20 MHz channels. A 160 MHz channel may be formed by combining 8 contiguous 20 MHz channels, or by combining two non-contiguous 80 MHz channels, which may be referred to as an 80+80 configuration. For the 80+80 configuration, the data, after channel encoding, may be passed through a segment parser that may divide the data into two streams. Inverse Fast Fourier Transform (IFFT) processing, and time domain processing, may be done on each stream separately. The streams may be mapped on to the two 80 MHz channels, and the data may be transmitted by a transmitting STA. At the receiver of the receiving STA, the above described operation for the 80+80 configuration may be reversed, and the combined data may be sent to the Medium Access Control (MAC).

[0081] Sub 1 GHz modes of operation are supported by 802.11af and 802.11ah. The channel operating bandwidths, and carriers, are reduced in 802.11af and 802.11ah relative to those used in 802.11n, and

802.11ac, 802.11af supports 5 MHz, 10 MHz and 20 MHz bandwidths in the TV White Space (TVWS) spectrum, and 802.11ah supports 1 MHz, 2 MHz, 4 MHz, 8 MHz, and 16 MHz bandwidths using non-TVWS spectrum. According to a representative embodiment, 802.11ah may support Meter Type Control/Machine-Type Communications, such as MTC devices in a macro coverage area. MTC devices may have certain capabilities, for example, limited capabilities including support for (e.g., only support for) certain and/or limited bandwidths. The MTC devices may include a battery with a battery life above a threshold (e.g., to maintain a very long battery life).

[0082] 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.

[0083] 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.

[0084] 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.

[0085] The RAN 113 may include gNBs 180a, 180b, 180c, though it will be appreciated that the RAN 113 may include any number of gNBs while remaining consistent with an embodiment. The gNBs 180a, 180b, 180c may each include one or more transceivers for communicating with the WTRUs 102a, 102b, 102c over the air interface 116. In one embodiment, the gNBs 180a, 180b, 180c may implement MIMO technology. For example, gNBs 180a, 180b may utilize beamforming to transmit signals to and/or receive signals from the gNBs 180a,

180b, 180c. Thus, the gNB 180a, for example, may use multiple antennas to transmit wireless signals to, and/or receive wireless signals from, the WTRU 102a. In an embodiment, the gNBs 180a, 180b, 180c may implement carrier aggregation technology. For example, the gNB 180a may transmit multiple component carriers to the WTRU 102a (not shown). A subset of these component carriers may be on unlicensed spectrum while the remaining component carriers may be on licensed spectrum. In an embodiment, the gNBs 180a, 180b, 180c may implement Coordinated Multi-Point (CoMP) technology. For example, WTRU 102a may receive coordinated transmissions from gNB 180a and gNB 180b (and/or gNB 180c).

[0086] The WTRUs 102a, 102b, 102c may communicate with gNBs 180a, 180b, 180c using transmissions associated with a scalable numerology. For example, the OFDM symbol spacing and/or OFDM subcarrier spacing may vary for different transmissions, different cells, and/or different portions of the wireless transmission spectrum. The WTRUs 102a, 102b, 102c may communicate with gNBs 180a, 180b, 180c using subframe or transmission time intervals (TTIs) of various or scalable lengths (e.g., containing varying number of OFDM symbols and/or lasting varying lengths of absolute time).

[0087] 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.

[0088] Each of the gNBs 180a, 180b, 180c may be associated with a particular cell (not shown) and may be configured to handle radio resource management decisions, handover decisions, scheduling of users in the UL and/or DL, support of network slicing, dual connectivity, interworking between NR and E-UTRA, routing of user plane data towards User Plane Function (UPF) 184a, 184b, routing of control plane information towards

Access and Mobility Management Function (AMF) 182a, 182b and the like. As shown in FIG. 1D, the gNBs 180a, 180b, 180c may communicate with one another over an Xn interface.

[0089] The CN 115 shown in FIG. 1D may include at least one AMF 182a, 182b, at least one UPF 184a, 184b, at least one Session Management Function (SMF) 183a, 183b, and possibly a Data Network (DN) 185a, 185b. While each of the foregoing elements are depicted as part of the CN 115, it will be appreciated that any of these elements may be owned and/or operated by an entity other than the CN operator.

[0090] The AMF 182a, 182b may be connected to one or more of the gNBs 180a, 180b, 180c in the RAN 113 via an N2 interface and may serve as a control node. For example, the AMF 182a, 182b may be responsible for authenticating users of the WTRUs 102a, 102b, 102c, support for network slicing (e.g., handling of different PDU sessions with different requirements), selecting a particular SMF 183a, 183b, management of the registration area, termination of NAS signaling, mobility management, and the like. Network slicing may be used by the AMF 182a, 182b in order to customize CN support for WTRUs 102a, 102b, 102c based on the types of services being utilized WTRUs 102a, 102b, 102c. For example, different network slices may be established for different use cases such as services relying on ultra-reliable low latency (URLLC) access, services relying on enhanced massive mobile broadband (eMBB) access, services for machine type communication (MTC) access, and/or the like. The AMF 162 may provide a control plane function for switching between the RAN 113 and other RANs (not shown) that employ other radio technologies, such as LTE, LTE-A, LTE-A Pro, and/or non-3GPP access technologies such as WiFi.

[0091] 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.

[0092] The UPF 184a, 184b may be connected to one or more of the gNBs 180a, 180b, 180c in the RAN 113 via an N3 interface, which may provide the WTRUs 102a, 102b, 102c with access to packet-switched networks, such as the Internet 110, to facilitate communications between the WTRUs 102a, 102b, 102c and IP-enabled devices. The UPF 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.

[0093] The CN 115 may facilitate communications with other networks. For example, the CN 115 may include, or may communicate with, an IP gateway (e.g., an IP multimedia subsystem (IMS) server) that serves as an interface between the CN 115 and the PSTN 108. In addition, the CN 115 may provide the WTRUs 102a, 102b, 102c with access to the other networks 112, which may include other wired and/or wireless networks that are owned and/or operated by other service providers. In one embodiment, the WTRUs 102a, 102b, 102c may be connected to a local Data Network (DN) 185a, 185b through the UPF 184a, 184b via the N3 interface to the UPF 184a, 184b and an N6 interface between the UPF 184a, 184b and the DN 185a, 185b.

[0094] In view of Figures 1A-1D, and the corresponding description of Figures 1A-1D, one or more, or all, of the functions described herein with regard to one or more of: WTRU 102a-d, Base Station 114a-b, eNode-B 160a-c, MME 162, SGW 164, PGW 166, gNB 180a-c, AMF 182a-b, UPF 184a-b, SMF 183a-b, DN 185a-b, and/or any other device(s) described herein, may be performed by one or more emulation devices (not shown). The emulation devices may be one or more devices configured to emulate one or more, or all, of the functions described herein. For example, the emulation devices may be used to test other devices and/or to simulate network and/or WTRU functions.

[0095] 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 performing testing using over-the-air wireless communications.

[0096] 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.

[0097] This application describes a variety of aspects, including tools, features, examples, models, approaches, etc. Many of these aspects are described with specificity and, at least to show the individual

characteristics, are often described in a manner that may sound limiting. However, this is for purposes of clarity in description, and does not limit the application or scope of those aspects. Indeed, all of the different aspects may be combined and interchanged to provide further aspects. Moreover, the aspects may be combined and interchanged with aspects described in earlier filings as well.

[0098] The aspects described and contemplated in this application may be implemented in many different forms. FIGs. 5-27 described herein may provide some examples, but other examples are contemplated. The discussion of FIGs. 5-27 does not limit the breadth of the implementations. At least one of the aspects generally relates to video encoding and decoding, and at least one other aspect generally relates to transmitting a bitstream generated or encoded. These and other aspects may be implemented as a method, an apparatus, a computer readable storage medium having stored thereon instructions for encoding or decoding video data according to any of the methods described, and/or a computer readable storage medium having stored thereon a bitstream generated according to any of the methods described.

[0099] In the present application, the terms “reconstructed” and “decoded” may be used interchangeably, the terms “pixel” and “sample” may be used interchangeably, the terms “image,” “picture” and “frame” may be used interchangeably.

[0100] Various methods are described herein, and each of the methods comprises one or more steps or actions for achieving the described method. Unless a specific order of steps or actions is required for proper operation of the method, the order and/or use of specific steps and/or actions may be modified or combined. Additionally, terms such as “first,” “second,” etc., may be used in various examples to modify an element, component, step, operation, etc., such as, for example, a “first decoding” and a “second decoding.” Use of such terms does not imply an ordering to the modified operations unless specifically required. So, in this example, the first decoding need not be performed before the second decoding, and may occur, for example, before, during, or in an overlapping time period with the second decoding.

[0101] Various methods and other aspects described in this application may be used to modify modules, for example, decoding modules, of a video encoder 200 and decoder 300 as shown in FIG. 2 and FIG. 3. Moreover, the subject matter disclosed herein may be applied, for example, to any type, format or version of video coding, whether described in a standard or a recommendation, whether pre-existing or future-developed, and extensions of any such standards and recommendations. Unless indicated otherwise, or technically precluded, the aspects described in this application may be used individually or in combination.

[0102] Various numeric values are used in examples described the present application, such as 1, 2, 4, 7, 8, 16, 32, 64, etc. These and other specific values are for purposes of describing examples and the aspects described are not limited to these specific values.

[0103] FIG. 2 is a diagram showing an example video encoder 200. Variations of example encoder 200 are contemplated, but the encoder 200 is described below for purposes of clarity without describing all expected variations.

[0104] Before being encoded, the video sequence may go through pre-encoding processing (201), for example, applying a color transform to the input color picture (e.g., conversion from RGB 4:4:4 to YCbCr 4:2:0), or performing a remapping of the input picture components in order to get a signal distribution more resilient to compression (for instance using a histogram equalization of one of the color components). Metadata (e.g., which may include film grain parameters determined by pre-processing as described herein) may be associated with the pre-processing and attached to the bitstream.

[0105] In the encoder 200, a picture is encoded by the encoder elements as described below. The picture to be encoded is partitioned (202) and processed in units of, for example, coding units (CUs). Each unit is encoded using, for example, either an intra or inter mode. When a unit is encoded in an intra mode, it performs intra prediction (260). In an inter mode, motion estimation (275) and compensation (270) are performed. The encoder decides (205) which one of the intra mode or inter mode to use for encoding the unit, and indicates the intra/inter decision by, for example, a prediction mode flag. Prediction residuals are calculated, for example, by subtracting (210) the predicted block from the original image block.

[0106] The prediction residuals are then transformed (225) and quantized (230). The quantized transform coefficients, as well as motion vectors and other syntax elements, are entropy coded (245) to output a bitstream. The encoder can skip the transform and apply quantization directly to the non-transformed residual signal. The encoder can bypass both transform and quantization, i.e., the residual is coded directly without the application of the transform or quantization processes.

[0107] The encoder decodes an encoded block to provide a reference for further predictions. The quantized transform coefficients are de-quantized (240) and inverse transformed (250) to decode prediction residuals. Combining (255) the decoded prediction residuals and the predicted block, an image block is reconstructed. In-loop filters (265) are applied to the reconstructed picture to perform, for example, deblocking/SAO (Sample Adaptive Offset) filtering to reduce encoding artifacts. The filtered image is stored at a reference picture buffer (280).

[0108] FIG. 3 is a diagram showing an example of a video decoder. In example decoder 300, a bitstream is decoded by the decoder elements as described below. Video decoder 300 generally performs a decoding pass reciprocal to the encoding pass as described in FIG. 2. The encoder 200 also generally performs video decoding as part of encoding video data.

[0109] In particular, the input of the decoder includes a video bitstream, which may be generated by video encoder 200. The bitstream is first entropy decoded (330) to obtain transform coefficients, motion vectors, and other coded information. The picture partition information indicates how the picture is partitioned. The decoder may therefore divide (335) the picture according to the decoded picture partitioning information. The transform coefficients are de-quantized (340) and inverse transformed (350) to decode the prediction residuals. Combining (355) the decoded prediction residuals and the predicted block, an image block is reconstructed. The predicted block may be obtained (370) from intra prediction (360) or motion-compensated prediction (i.e., inter prediction) (375). In-loop filters (365) are applied to the reconstructed image. The filtered image is stored at a reference picture buffer (380).

[0110] The decoded picture can further go through post-decoding processing (385), for example, an inverse color transform (e.g., conversion from YCbCr 4:2:0 to RGB 4:4:4) or an inverse remapping performing the inverse of the remapping process performed in the pre-encoding processing (201). The post-decoding processing can use metadata derived in the pre-encoding processing and signaled in the bitstream. In an example, the decoded images (e.g., after application of the in-loop filters (365) and/or after post-decoding processing (385), if post-decoding processing is used) may be sent to a display device for rendering to a user.

[0111] FIG. 4 is a diagram showing an example of a system in which various aspects and examples described herein may be implemented. System 400 may be embodied as a device including the various components described below and is configured to perform one or more of the aspects described in this document. Examples of such devices, include, but are not limited to, various electronic devices such as personal computers, laptop computers, smartphones, tablet computers, digital multimedia set top boxes, digital television receivers, personal video recording systems, connected home appliances, and servers. Elements of system 400, singly or in combination, may be embodied in a single integrated circuit (IC), multiple ICs, and/or discrete components. For example, in at least one example, the processing and encoder/decoder elements of system 400 are distributed across multiple ICs and/or discrete components. In various examples, the system 400 is communicatively coupled to one or more other systems, or other electronic devices, via, for example, a communications bus or through dedicated input and/or output ports. In various examples, the system 400 is configured to implement one or more of the aspects described in this document.

[0112] The system 400 includes at least one processor 410 configured to execute instructions loaded therein for implementing, for example, the various aspects described in this document. Processor 410 can include embedded memory, input output interface, and various other circuitries as known in the art. The system 400 includes at least one memory 420 (e.g., a volatile memory device, and/or a non-volatile memory device). System 400 includes a storage device 440, which can include non-volatile memory and/or volatile memory, including, but not limited to, Electrically Erasable Programmable Read-Only Memory (EEPROM), Read-Only Memory (ROM), Programmable Read-Only Memory (PROM), Random Access Memory (RAM), Dynamic Random Access Memory (DRAM), Static Random Access Memory (SRAM), flash, magnetic disk drive, and/or optical disk drive. The storage device 440 can include an internal storage device, an attached storage device (including detachable and non-detachable storage devices), and/or a network accessible storage device, as non-limiting examples.

[0113] System 400 includes an encoder/decoder module 430 configured, for example, to process data to provide an encoded video or decoded video, and the encoder/decoder module 430 can include its own processor and memory. The encoder/decoder module 430 represents module(s) that may be included in a device to perform the encoding and/or decoding functions. As is known, a device can include one or both of the encoding and decoding modules. Additionally, encoder/decoder module 430 may be implemented as a separate element of system 400 or may be incorporated within processor 410 as a combination of hardware and software as known to those skilled in the art.

[0114] Program code to be loaded onto processor 410 or encoder/decoder 430 to perform the various aspects described in this document may be stored in storage device 440 and subsequently loaded onto memory 420 for execution by processor 410. In accordance with various examples, one or more of processor 410, memory 420, storage device 440, and encoder/decoder module 430 can store one or more of various items during the performance of the processes described in this document. Such stored items can include, but are not limited to, the input video, the decoded video or portions of the decoded video, the bitstream, matrices, variables, and intermediate or final results from the processing of equations, formulas, operations, and operational logic.

[0115] In some examples, memory inside of the processor 410 and/or the encoder/decoder module 430 is used to store instructions and to provide working memory for processing that is needed during encoding or decoding. In other examples, however, a memory external to the processing device (for example, the processing device may be either the processor 410 or the encoder/decoder module 430) is used for one or more of these functions. The external memory may be the memory 420 and/or the storage device 440, for

example, a dynamic volatile memory and/or a non-volatile flash memory. In several examples, an external non-volatile flash memory is used to store the operating system of, for example, a television. In at least one example, a fast external dynamic volatile memory such as a RAM is used as working memory for video encoding and decoding operations.

[0116] The input to the elements of system 400 may be provided through various input devices as indicated in block 445. Such input devices include, but are not limited to, (i) a radio frequency (RF) portion that receives an RF signal transmitted, for example, over the air by a broadcaster, (ii) a Component (COMP) input terminal (or a set of COMP input terminals), (iii) a Universal Serial Bus (USB) input terminal, and/or (iv) a High Definition Multimedia Interface (HDMI) input terminal. Other examples, not shown in FIG. 4, include composite video.

[0117] In various examples, the input devices of block 445 have associated respective input processing elements as known in the art. For example, the RF portion may be associated with elements suitable for (i) selecting a desired frequency (also referred to as selecting a signal, or band-limiting a signal to a band of frequencies), (ii) down-converting the selected signal, (iii) band-limiting again to a narrower band of frequencies to select (for example) a signal frequency band which may be referred to as a channel in certain examples, (iv) demodulating the down-converted and band-limited signal, (v) performing error correction, and/or (vi) demultiplexing to select the desired stream of data packets. The RF portion of various examples includes one or more elements to perform these functions, for example, frequency selectors, signal selectors, band-limiters, channel selectors, filters, downconverters, demodulators, error correctors, and demultiplexers. The RF portion can include a tuner that performs various of these functions, including, for example, down-converting the received signal to a lower frequency (for example, an intermediate frequency or a near-baseband frequency) or to baseband. In one set-top box example, the RF portion and its associated input processing element receives an RF signal transmitted over a wired (for example, cable) medium, and performs frequency selection by filtering, down-converting, and filtering again to a desired frequency band. Various examples rearrange the order of the above-described (and other) elements, remove some of these elements, and/or add other elements performing similar or different functions. Adding elements can include inserting elements in between existing elements, such as, for example, inserting amplifiers and an analog-to-digital converter. In various examples, the RF portion includes an antenna.

[0118] The USB and/or HDMI terminals can include respective interface processors for connecting system 400 to other electronic devices across USB and/or HDMI connections. It is to be understood that various aspects of input processing, for example, Reed-Solomon error correction, may be implemented, for example, within a separate input processing IC or within processor 410 as necessary. Similarly, aspects of USB or HDMI

interface processing may be implemented within separate interface ICs or within processor 410 as necessary. The demodulated, error corrected, and demultiplexed stream is provided to various processing elements, including, for example, processor 410, and encoder/decoder 430 operating in combination with the memory and storage elements to process the data stream as necessary for presentation on an output device.

[0119] Various elements of system 400 may be provided within an integrated housing. Within the integrated housing, the various elements may be interconnected and transmit data therebetween using suitable connection arrangement 425, for example, an internal bus as known in the art, including the Inter-IC (I2C) bus, wiring, and printed circuit boards.

[0120] The system 400 includes communication interface 450 that enables communication with other devices via communication channel 460. The communication interface 450 can include, but is not limited to, a transceiver configured to transmit and to receive data over communication channel 460. The communication interface 450 can include, but is not limited to, a modem or network card and the communication channel 460 may be implemented, for example, within a wired and/or a wireless medium.

[0121] Data is streamed, or otherwise provided, to the system 400, in various examples, using a wireless network such as a Wi-Fi network, for example IEEE 802.11 (IEEE refers to the Institute of Electrical and Electronics Engineers). The Wi-Fi signal of these examples is received over the communications channel 460 and the communications interface 450 which are adapted for Wi-Fi communications. The communications channel 460 of these examples is typically connected to an access point or router that provides access to external networks including the Internet for allowing streaming applications and other over-the-top communications. Other examples provide streamed data to the system 400 using a set-top box that delivers the data over the HDMI connection of the input block 445. Still other examples provide streamed data to the system 400 using the RF connection of the input block 445. As indicated above, various examples provide data in a non-streaming manner. Additionally, various examples use wireless networks other than Wi-Fi, for example a cellular network or a Bluetooth® network.

[0122] The system 400 can provide an output signal to various output devices, including a display 475, speakers 485, and other peripheral devices 495. The display 475 of various examples includes one or more of, for example, a touchscreen display, an organic light-emitting diode (OLED) display, a curved display, and/or a foldable display. The display 475 may be for a television, a tablet, a laptop, a cell phone (mobile phone), or another device. The display 475 can also be integrated with other components (for example, as in a smart phone), or separate (for example, an external monitor for a laptop). The other peripheral devices 495 include, in various examples, one or more of a stand-alone digital video disc (or digital versatile disc) (DVD, for both

terms), a disk player, a stereo system, and/or a lighting system. Various examples use one or more peripheral devices 495 that provide a function based on the output of the system 400. For example, a disk player performs the function of playing the output of the system 400.

[0123] In various examples, control signals are communicated between the system 400 and the display 475, speakers 485, or other peripheral devices 495 using signaling such as AV.Link, Consumer Electronics Control (CEC), or other communications protocols that enable device-to-device control with or without user intervention. The output devices may be communicatively coupled to system 400 via dedicated connections through respective interfaces 470, 480, and 490. Alternatively, the output devices may be connected to system 400 using the communications channel 460 via the communications interface 450. The display 475 and speakers 485 may be integrated in a single unit with the other components of system 400 in an electronic device such as, for example, a television. In various examples, the display interface 470 includes a display driver, such as, for example, a timing controller (T Con) chip.

[0124] The display 475 and speakers 485 can alternatively be separate from one or more of the other components, for example, if the RF portion of input 445 is part of a separate set-top box. In various examples in which the display 475 and speakers 485 are external components, the output signal may be provided via dedicated output connections, including, for example, HDMI ports, USB ports, or COMP outputs.

[0125] The examples may be carried out by computer software implemented by the processor 410 or by hardware, or by a combination of hardware and software. As a non-limiting example, the examples may be implemented by one or more integrated circuits. The memory 420 may be of any type appropriate to the technical environment and may be implemented using any appropriate data storage technology, such as optical memory devices, magnetic memory devices, semiconductor-based memory devices, fixed memory, and removable memory, as non-limiting examples. The processor 410 may be of any type appropriate to the technical environment, and can encompass one or more of microprocessors, general purpose computers, special purpose computers, and processors based on a multi-core architecture, as non-limiting examples.

[0126] Various implementations involve decoding. "Decoding," as used in this application, can encompass all or part of the processes performed, for example, on a received encoded sequence in order to produce a final output suitable for display. In various examples, such processes include one or more of the processes typically performed by a decoder, for example, entropy decoding, inverse quantization, inverse transformation, and differential decoding. In various examples, such processes also, or alternatively, include processes performed by a decoder of various implementations described in this application, for example, receiving encoded intermediate data and encoded tensor metadata from a second device; decoding the encoded

intermediate data and the encoded tensor metadata; reconstructing a tensor based on the intermediate data and the tensor metadata, etc.

[0127] As further examples, in one example “decoding” refers only to entropy decoding, in another example “decoding” refers only to differential decoding, and in another example “decoding” refers to a combination of entropy decoding and differential decoding. Whether the phrase “decoding process” is intended to refer specifically to a subset of operations or generally to the broader decoding process will be clear based on the context of the specific descriptions and is believed to be well understood by those skilled in the art.

[0128] Various implementations involve encoding. In an analogous way to the above discussion about “decoding,” “encoding” as used in this application can encompass all or part of the processes performed, for example, on an input video sequence in order to produce an encoded bitstream. In various examples, such processes include one or more of the processes typically performed by an encoder, for example, partitioning, differential encoding, transformation, quantization, and entropy encoding. In various examples, such processes also, or alternatively, include processes performed by an encoder of various implementations described in this application, for example, determining tensor metadata associated with intermediate data; encoding the intermediate data and the tensor metadata associated with the intermediate data; sending the encoded intermediate data and the encoded tensor metadata to a second device, etc.

[0129] As further examples, in one example “encoding” refers only to entropy encoding, in another example “encoding” refers only to differential encoding, and in another example “encoding” refers to a combination of differential encoding and entropy encoding. Whether the phrase “encoding process” is intended to refer specifically to a subset of operations or generally to the broader encoding process will be clear based on the context of the specific descriptions and is believed to be well understood by those skilled in the art.

[0130] Note that syntax elements as used herein are descriptive terms. As such, they do not preclude the use of other syntax element names.

[0131] When a figure is presented as a flow diagram, it should be understood that it also provides a block diagram of a corresponding apparatus. Similarly, when a figure is presented as a block diagram, it should be understood that it also provides a flow diagram of a corresponding method/process.

[0132] The implementations and aspects described herein may be implemented in, for example, a method or a process, an apparatus, a software program, a data stream, or a signal. Even if only discussed in the context of a single form of implementation (for example, discussed only as a method), the implementation of features discussed can also be implemented in other forms (for example, an apparatus or program). An apparatus may be implemented in, for example, appropriate hardware, software, and firmware. The methods

may be implemented in, for example, a processor, which refers to processing devices in general, including, for example, a computer, a microprocessor, an integrated circuit, or a programmable logic device. Processors also include communication devices, such as, for example, computers, cell phones, portable/personal digital assistants ("PDAs"), and other devices that facilitate communication of information between end-users.

[0133] Reference to "one example" or "an example" or "one implementation" or "an implementation," as well as other variations thereof, means that a particular feature, structure, characteristic, and so forth described in connection with the example is included in at least one example. Thus, the appearances of the phrase "in one example" or "in an example" or "in one implementation" or "in an implementation," as well as any other variations, appearing in various places throughout this application are not necessarily all referring to the same example.

[0134] Additionally, this application may refer to "determining" various pieces of information. Determining the information can include one or more of, for example, estimating the information, calculating the information, predicting the information, or retrieving the information from memory. Obtaining may include receiving, retrieving, constructing, generating, and/or determining.

[0135] Further, this application may refer to "accessing" various pieces of information. Accessing the information can include one or more of, for example, receiving the information, retrieving the information (for example, from memory), storing the information, moving the information, copying the information, calculating the information, determining the information, predicting the information, or estimating the information.

[0136] Additionally, this application may refer to "receiving" various pieces of information. Receiving is, as with "accessing," intended to be a broad term. Receiving the information can include one or more of, for example, accessing the information, or retrieving the information (for example, from memory). Further, "receiving" is typically involved, in one way or another, during operations such as, for example, storing the information, processing the information, transmitting the information, moving the information, copying the information, erasing the information, calculating the information, determining the information, predicting the information, or estimating the information.

[0137] It is to be appreciated that the use of any of the following "/", "and/or," and "at least one of," for example, in the cases of "A/B," "A and/or B," and "at least one of A and B" is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of both options (A and B). As a further example, in the cases of "A, B, and/or C" and "at least one of A, B, and C," such phrasing is intended to encompass the selection of the first listed option (A) only, or the selection of the second listed option (B) only, or the selection of the third listed option (C) only, or the selection of the first and the second listed options (A and B) only, or the selection of the first and third listed options (A and C) only,

or the selection of the second and third listed options (B and C) only, or the selection of all three options (A and B and C). This may be extended, as is clear to one of ordinary skill in this and related arts, for as many items as are listed.

[0138] Also, as used herein, the word “signal” refers to, among other things, indicating something to a corresponding decoder. Encoder signals may include, for example, number of intensity intervals, number of model values, grain parameters, grain identification, scaling factor, etc. In this way, in an example the same parameter is used at both the encoder side and the decoder side. Thus, for example, an encoder can transmit (explicit signaling) a particular parameter to the decoder so that the decoder can use the same particular parameter. Conversely, if the decoder already has the particular parameter as well as others, then signaling may be used without transmitting (implicit signaling) to simply allow the decoder to know and select the particular parameter. By avoiding transmission of any actual functions, a bit savings is realized in various examples. It is to be appreciated that signaling may be accomplished in a variety of ways. For example, one or more syntax elements, flags, and so forth are used to signal information to a corresponding decoder in various examples. While the preceding relates to the verb form of the word “signal,” the word “signal” can also be used herein as a noun.

[0139] As will be evident to one of ordinary skill in the art, implementations may produce a variety of signals formatted to carry information that may be, for example, stored or transmitted. The information can include, for example, instructions for performing a method, or data produced by one of the described implementations. For example, a signal may be formatted to carry the bitstream of a described example. Such a signal may be formatted, for example, as an electromagnetic wave (for example, using a radio frequency portion of spectrum) or as a baseband signal. The formatting may include, for example, encoding a data stream and modulating a carrier with the encoded data stream. The information that the signal carries may be, for example, analog or digital information. The signal may be transmitted over a variety of different wired or wireless links, as is known. The signal may be stored on, or accessed or received from, a processor-readable medium.

[0140] Many examples are described herein. Features of examples may be provided alone or in any combination, across various claim categories and types. Further, examples may include one or more of the features, devices, or aspects described herein, alone or in any combination, across various claim categories and types. For example, features described herein may be implemented in a bitstream or signal that includes information generated as described herein. The information may allow a decoder to decode a bitstream, the encoder, bitstream, and/or decoder according to any of the embodiments described. For example, features described herein may be implemented by creating and/or transmitting and/or receiving and/or decoding a

bitstream or signal. For example, features described herein may be implemented a method, process, apparatus, medium storing instructions, medium storing data, or signal. For example, features described herein may be implemented by a TV, set-top box, cell phone, tablet, or other electronic device that performs decoding. The TV, set-top box, cell phone, tablet, or other electronic device may display (e.g., using a monitor, screen, or other type of display) a resulting image (e.g., an image from residual reconstruction of the video bitstream). The TV, set-top box, cell phone, tablet, or other electronic device may receive a signal including an encoded image and perform decoding.

[0141] FIGs. 5 and 6 illustrate example architectures for split inferences of a model (e.g., an artificial intelligence (AI) model) composed of n layers (e.g., layers $1 \dots n$) between the network and a WTRU. A first inference may process a first part of the AI model (e.g., layers $1 \dots k$). A second inference may process a second part of the AI model (e.g., layers $k+1 \dots n$).

[0142] The architectures in FIGs. 5 and 6 illustrate the delivery and access functions of intermediate data between the WTRU and the network for different scenarios (e.g., two different scenarios). For example, if the media data source originates from the WTRU, the first part of the inference may be performed in the WTRU. In this case, the second part may be performed in the network. The resulting output data may be sent back to the WTRU. If the media data source comes from the network, or from the network via the WTRU, the first part of the AI model may be performed on the network side. In this case, the second part of the AI model may be performed in the WTRU.

[0143] FIG. 5 illustrates a split inference between the WTRU and network, with the media data source originating from the WTRU. FIG. 6 illustrates a split inference between the WTRU and network, with the media data source originating from the network.

[0144] Among the different functions, the intermediate data delivery function may apply a serialization technique (e.g., to encode the intermediate data output). An intermediate data access function may apply a corresponding deserialization technique (e.g., to obtain the intermediate data). A binary serialization may transform a data structure into a series of bytes for the intermediate data.

[0145] The intermediate data resulting from the first part of the AI model (e.g., coming from the WTRU, to the network) may include multi-dimensional arrays or tensor(s). The transmitted tensor(s) may feed a second inference (e.g., taking place in the network, respectively in the WTRU on another device).

[0146] A tensor shape may be a tuple of positive integers. The size of the tuple may represent the dimension of the tensor. A value (e.g., each value) may represent the size in a dimension (e.g., each

dimension). For example, a tensor of shape (1, 64, 112, 112) may describe a four-dimension array with 1, 64, 112, and 112 elements in each dimension, respectively.

[0147] The output intermediate data dimension may be lost. The tensor shape may be different depending on the split point. If the data is encoded to a binary bitstream, the tensor dimension may be lost.

[0148] Efficient delivery may be desired. A device may receive and process the intermediate data bitstream. The speed at which the device processes the intermediate data bit stream may depend on a given frame rate (e.g., the speed may be fast if a processing high frame rate input is used).

[0149] Metadata may be transmitted. The metadata may include the tensor shape(s) of the intermediate data. This may enable the reconstruction of the intermediate tensor(s) from the decoded bitstream (e.g., with the proper original intermediate data shape). The metadata may be transmitted in a band alongside the bitstream of the intermediate data. The metadata may be transmitted out of band. For example, the metadata may be carried in a specific control channel or through the same data channel as the intermediate data.

[0150] Feature(s) associated with tensor shape metadata for encoded intermediate data are provided herein.

[0151] A first device may process a first part of the split AI model. The first device may encode the output intermediate data. The first device may encode the tensor shape as metadata (e.g., alongside the bitstream of the intermediate data). The first device may transmit the combined encoded bitstream to a second device.

[0152] The second device may receive the encoded bitstream. The second device may decode the intermediate data and the metadata. The second device may extract the tensor shape from the metadata. The second device may reconstruct the intermediate data bitstream to its original tensor shape (e.g., before feeding the intermediate data to the second part of the AI model).

[0153] The tensor shape may be encoded using one or more encoding techniques or algorithms.

[0154] An encoding type may indicate which encoding technique/algorithm or tensor shape encoding structure was used to encode the intermediate data and/or the metadata. The encoding type may be a Boolean (e.g., if there are only two possible encoding types).

[0155] The encoded shape may include a total length indication that indicates the total length of the encoded structure (e.g., 4, as illustrated in FIG. 7). The encoded shape may include a payload with a list including a sequence of unsigned integers (e.g., 1, 64, 64, 112 in FIG. 7). The encoded shape may include a list of unsigned integers [dimension number, value dimension 1, value dimension 2, ...] (e.g., 4, 1, 64, 62, 112 for a 4 four-dimension array with 1, 64, 64, and 112 elements in the respective dimensions).

[0156] FIG. 7 illustrates an example encoded tensor shape. In FIG. 7, the encoding type “X” means that the tensor shape is encoded with bytes. The encoding type “Y” means that the tensor shape is encoded with strings. In the examples shown in FIG. 7, the tensor shape is [1,64,64,64] (e.g., as indicated by the payload portion of the encoded shape).

[0157] The encoding technique or algorithm indication may be transmitted from the first device to the second device. For example, the encoding technique or algorithm indication may be transmitted as metadata alongside encoded intermediate data.

[0158] In some examples, the encoded tensor may be transmitted alongside the intermediate bitstream (e.g., every time that the intermediate data is transmitted). In some examples, the encoded tensor may be transmitted if the shape changes (e.g., only when the shape changes). For example, some neural network models (e.g., the Resnet family of deep neural network models) may have (e.g., start with) several layers (e.g., conv2d, BatchNorm2d_2, Relu_3). For example, a first layer may provide intermediate data with a shape (e.g., the same shape S1). Another layer (e.g., the next layer, for example, maxPoll_4) may provide intermediate data with a different shape (e.g., S2). In some examples, an encoded shape S1 may be transmitted with the conv2d layer. In some examples, no encoded shape may be transmitted with (BatchNorm2d_2, Relu_3). In some examples, the encoded shape S2 may be transmitted before processing the maxPoll_4 layer.

[0159] FIG. 8 illustrates an example change mode for encoding a tensor shape. As illustrated in FIG. 8, an encoding type “C” may indicate that the tensor shape is provided (e.g., only provided) if the tensor shape changes. This may save bandwidth and increase computing efficiency. The device receiving the intermediate data may keep track of the current shape of the tensor.

[0160] In some examples, the metadata may include tensor information (e.g., detailed tensor information). For example, the metadata may include a tensor structure identifier. The tensor structure identifier may identify the tensor structure (e.g., the exact tensor structure) in use (e.g., including the version of the tensor structure). A tensor structure identification table may be shared and known by the devices (e.g., both the transmitting and receiving devices). The tensor structure identification table may be transmitted during a setup stage (e.g., from sending control data between the devices). Table 1 illustrates an example of a tensor table with identifiers and associated descriptions.

Identifier	Description
1	pytorch 2.0
2	tensor flow v2.13.0
3	numpy v1.25

Table 1. Example tensor table

[0161] The metadata may include a tensor name. The tensor name may identify the tensor structure (e.g., the exact tensor structure) in use based on a string representation. For example, the tensor name may be "Pytorch.Tensor 2.0", "tensor flow v2.13.0", or "numpy v1 .25".

[0162] In an example, the first device may transmit tensor shape metadata out of band. The tensor shape metadata may be carried in a control channel (e.g., a specific control channel). The tensor shape metadata may be carried through the data channel (e.g., the same data channel) as the intermediate data. The metadata may not be synchronized and carried with intermediate data.

[0163] FIG. 9 illustrates an example split inference between a WTRU and a network. FIG. 9 illustrates an example in which the media data source (e.g., with tensor shape metadata) is in the WTRU. In this case, the WTRU may provide tensor shape metadata to the network.

[0164] FIG. 10 illustrates an example split inference between a WTRU and a network in which the media data source (e.g., with tensor shape metadata) is in the network. FIG. 10 illustrates an example in which the network provides the tensor shape metadata to the WTRU. FIG. 11 illustrates examples of the intermediate data access and delivery functions.

[0165] The intermediate data delivery function may be in the first device. The first device may receive output intermediate data (e.g., resulting from inference of the first part of the AI model (e.g., a head part). The first device may encode the tensor shape as metadata (e.g., according to a tensor shape structure or algorithm). The encoding algorithm may be a binary structure, a binary serialization, or a textual serialization (e.g., available for the small data size). The first device may encode the intermediate data as a bitstream (e.g., according to a bitstream structure). For example, the intermediate data may be encoded as a bitstream of float or integer (e.g., int). Encoding a tensor may involve reshaping and/or flattening the input multi-dimensional tensor into a one-dimension tensor. The tensor may be encoded (e.g., may then be encoded) according to a binary structure (e.g., a predefined binary structure). The binary serialization may be used for large blocks of data (e.g., such as intermediate data, which may be several megabytes in size).

[0166] A binary serialization (e.g., a single binary serialization) may comprise both data and metadata. The serialization may be based on a compound structure (e.g., a predefined compound structure) for both data and metadata.

[0167] The first device may send the encoded bitstream (e.g., data and metadata) to the second device. In some examples, data and metadata bitstreams may be combined. In some examples, metadata may be separated and sent independently (e.g., out of band).

[0168] The structure may be known to the first device and the second device (e.g., on both sender and receiver sides). The structure may be distributed or shared during a configuration stage. The structure may be transmitted from one device to the other (e.g., at least before starting to encode and transmitting the encoded data and metadata).

[0169] The intermediate data access function may be in the second device (e.g., the receiver device). The second device may receive a data bitstream and metadata. The metadata may include at least the tensor shape. The second device may decode the tensor shape metadata (e.g., according to the tensor shape structure or algorithm) to obtain the tensor shape. The second device may decode the intermediate data bitstream (e.g., according to a bitstream structure) to obtain the series of intermediate data bytes. The second device may reconstruct the original tensor from the series of intermediate data bytes. The reconstruction may involve converting binary data to a tensor, (e.g., to reshape the tensor into the indicated tensor shape). The second device may provide the reshaped (e.g., original) intermediate data to the inference processing the second part of the AI model (e.g., the last part of the AI model, sometimes referred to as the tail part of the AI model).

[0170] The size of the tensor shape may be used to compute/reconstruct/decode the received intermediate data. Sending the tensor shape alongside the intermediate data bitstream may avoid providing and/or updating tensor static shapes configurations for the trained models (e.g., all of the trained models to be used). One or more trained model (e.g., with the same name) may include modified layer shapes. In this case, the inference may not raise an error.

[0171] Although some features described herein may be performed by a video codec, one of ordinary skill in the art will appreciate that these features are not limited to the use of a video codec.

[0172] Although features and elements are described above in particular combinations, one of ordinary skill in the art will appreciate that each feature or element can be used alone or in any combination with the other features and elements. In addition, the methods described herein may be implemented in a computer program, software, or firmware incorporated in a computer-readable medium for execution by a computer or processor. Examples of computer-readable 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.

CLAIMS

What is claimed:

1. A first device, comprising:
a processor configured to:
 - determine tensor metadata associated with intermediate data;
 - encode the intermediate data and the tensor metadata associated with the intermediate data; and
 - send the encoded intermediate data and the encoded tensor metadata to a second device.
2. The first device of claim 1, wherein the first device is associated with a first portion of an artificial intelligence (AI) model, wherein the first portion is an initial portion of the AI model, the second device is associated with a second portion of the AI model, wherein the second portion is a latter portion of the AI model, and the processor is further configured to receive the intermediate data as an output of the first portion of the AI model.
3. The first device of claim 1 or 2, wherein the intermediate data is first intermediate data, the tensor metadata is first tensor metadata indicative of a first tensor shape, and the processor is further configured to:
 - determine a second tensor shape associated with second intermediate data, wherein the second tensor shape is different from the first tensor shape;
 - encode the second intermediate data and second tensor metadata, wherein the second tensor metadata is indicative of the second tensor shape; and
 - send the encoded second intermediate data and the encoded second tensor metadata to the second device.
4. The first device of claim 1 or 2, wherein the intermediate data is first intermediate data, the tensor metadata is indicative of a tensor shape, and the processor is further configured to:
 - determine that second intermediate data is associated with the tensor shape;
 - encode the second intermediate data and an indication that the tensor metadata has not been included because the tensor shape has not changed; and
 - send the encoded second intermediate data and the encoded indication to the second device.

5. The first device of one of claims 1-4, wherein the tensor metadata is indicative of at least one of: a tensor shape, tensor structure information, a data type, or an encoding algorithm.
6. The first device of any one of claims 1-5, wherein the processor being configured to encode the intermediate data comprises the processor being configured to:
- transform a multi-dimensional tensor into a one-dimensional tensor; and
 - encode the one-dimensional tensor in accordance with a binary structure.
7. The first device of any one of claims 1-6, wherein the processor being configured to send the encoded intermediate data and the encoded tensor metadata to the second device comprises the processor being configured to:
- send the encoded intermediate data in a first band associated with a data channel; and
 - send the encoded tensor metadata in a second band associated with a control channel.
8. The first device of any one of claims 1-7, wherein the processor being configured to encode the intermediate data and the tensor metadata associated with the intermediate data comprises the processor being configured to generate an encoded bitstream, wherein the encoded bitstream comprises the encoded intermediate data and the encoded tensor metadata, and the processor being configured to send the encoded intermediate data and the encoded tensor metadata to the second device comprises the processor being configured to send the encoded bitstream to the second device.
9. The first device of any one of claims 1-8, wherein the first device is a wireless transmit/receive unit (WTRU) or a network entity.
10. A first device, comprising:
- a processor configured to:
 - receive encoded intermediate data and encoded tensor metadata from a second device;
 - decode the encoded intermediate data and the encoded tensor metadata; and
 - reconstruct a tensor based on the intermediate data and the tensor metadata.

11. The first device of claim 10, wherein the second device is associated with a first portion of an artificial intelligence (AI) model, wherein the first portion is an initial portion of the AI model, the first device is associated with a second portion of the AI model, wherein the second portion is a latter portion of the AI model, and the processor is further configured to input the reconstructed tensor into the second portion of the AI model.

12. The first device of claim 10 or 11, wherein the intermediate data is first intermediate data, the tensor metadata is first tensor metadata indicative of a first tensor shape, the tensor is a first tensor, and the processor is further configured to:

- receive encoded second intermediate data and encoded second tensor metadata from the second device;

- decode the encoded second intermediate data and the encoded second tensor metadata, wherein the second tensor metadata is indicative of a second tensor shape; and

- reconstruct a second tensor based on the second intermediate data and the second tensor metadata.

13. The first device of claim 10 or 11, wherein the intermediate data is first intermediate data, the tensor metadata is indicative of a tensor shape, and the processor is further configured to:

- receive encoded second intermediate data and an encoded indication that the tensor metadata has not been included because the tensor shape has not changed;

- decode the encoded second intermediate data and the indication; and

- determine that the second intermediate data is associated with the tensor shape based on the indication.

14. The first device of one of claims 10-13, wherein the tensor metadata is indicative of at least one of: a tensor shape, tensor structure information, a data type, or an encoding algorithm.

15. The first device of any one of claims 10-14, wherein the processor being configured to decode the encoded intermediate data comprises the processor being configured to:

- transform a one-dimensional tensor into a multi-dimensional tensor; and

- decode the multi-dimensional tensor in accordance with a binary structure.

16. The first device of any one of claims 10-15, wherein the processor being configured to receive the encoded intermediate data and the encoded tensor metadata from the second device comprises the processor being configured to:

- receive the encoded intermediate data in a first band associated with a data channel; and
- receive the encoded tensor metadata in a second band associated with a control channel.

17. The first device of any one of claims 10-16, wherein the processor being configured to receive the encoded intermediate data and the encoded tensor metadata from the second device comprises the processor being configured to receive an encoded bitstream from the second device, wherein the encoded bitstream comprises the encoded intermediate data and the encoded tensor metadata, and the processor being configured to decode the encoded intermediate data and the encoded tensor metadata comprises the processor being configured to decode the encoded bitstream.

18. The first device of any one of claim 10-17, wherein the processor being configured to decode the encoded intermediate data and the encoded tensor metadata comprises the processor being configured to decode one or more of:

- the encoded tensor metadata according to a tensor shape structure or algorithm to obtain the tensor shape; or
- the encoded intermediate data according to a bitstream structure to obtain a series of intermediate data bytes.

19. The first device of any one of claims 10-18, wherein the processor being configured to reconstruct the tensor based on the intermediate data and the tensor metadata comprises the processor being configured to:

- convert binary data, in the intermediate data, to tensor data; and
- reshape the tensor to an original tensor shape based on the tensor data and the tensor metadata.

20. The first device of any one of claims 10-19, wherein the first device is a wireless transmit/receive unit (WTRU) or a network entity.

21. A method, performed by a first device, comprising:

- determining tensor metadata associated with intermediate data;

encoding the intermediate data and the tensor metadata associated with the intermediate data; and
sending the encoded intermediate data and the encoded tensor metadata to a second device.

22. The method of claim 21, wherein the first device is associated with a first portion of an artificial intelligence (AI) model, wherein the first portion is an initial portion of the AI model, the second device is associated with a second portion of the AI model, wherein the second portion is a latter portion of the AI model, and the method further comprises receiving the intermediate data as an output of the first portion of the AI model.

23. The method of claim 21 or 22, wherein the intermediate data is first intermediate data, the tensor metadata is first tensor metadata indicative of a first tensor shape, and the method further comprises:

determining a second tensor shape associated with second intermediate data, wherein the second tensor shape is different from the first tensor shape;

encoding the second intermediate data and second tensor metadata, wherein the second tensor metadata is indicative of the second tensor shape; and

sending the encoded second intermediate data and the encoded second tensor metadata to the second device.

24. The method of claim 21 or 22, wherein the intermediate data is first intermediate data, the tensor metadata is indicative of a tensor shape, and the method further comprises:

determining that second intermediate data is associated with the tensor shape;

encoding the second intermediate data and an indication that the tensor metadata has not been included because the tensor shape has not changed; and

sending the encoded second intermediate data and the encoded indication to the second device.

25. The method of one of claims 21-24, wherein the tensor metadata is indicative of at least one of: a tensor shape, tensor structure information, a data type, or an encoding algorithm.

26. The method of any one of claims 21-25, wherein encoding the intermediate data comprises:

transforming a multi-dimensional tensor into a one-dimensional tensor; and

encoding the one-dimensional tensor in accordance with a binary structure.

27. The method of any one of claims 21-26, wherein sending the encoded intermediate data and the encoded tensor metadata to the second device comprises:

- sending the encoded intermediate data in a first band associated with a data channel; and
- sending the encoded tensor metadata in a second band associated with a control channel.

28. The method of any one of claims 21-27, wherein encoding the intermediate data and the tensor metadata associated with the intermediate data comprises generating an encoded bitstream, wherein the encoded bitstream comprises the encoded intermediate data and the encoded tensor metadata, and sending the encoded intermediate data and the encoded tensor metadata to the second device comprises sending the encoded bitstream to the second device.

29. The method of any one of claims 21-28, wherein the first device is a wireless transmit/receive unit (WTRU) or a network entity.

30. A method, performed by a first device, comprising:

- receiving encoded intermediate data and encoded tensor metadata from a second device;
- decoding the encoded intermediate data and the encoded tensor metadata; and
- reconstructing a tensor based on the intermediate data and the tensor metadata.

31. The method of claim 30, wherein the second device is associated with a first portion of an artificial intelligence (AI) model, wherein the first portion is an initial portion of the AI model, the first device is associated with a second portion of the AI model, wherein the second portion is a latter portion of the AI model, and the method further comprises inputting the reconstructed tensor into the second portion of the AI model.

32. The method of claim 30 or 31, wherein the intermediate data is first intermediate data, the tensor metadata is first tensor metadata indicative of a first tensor shape, the tensor is a first tensor, and the method further comprises:

- receiving encoded second intermediate data and encoded second tensor metadata from the second device;
- decoding the encoded second intermediate data and the encoded second tensor metadata, wherein the second tensor metadata is indicative of a second tensor shape; and

reconstructing a second tensor based on the second intermediate data and the second tensor metadata.

33. The method of claim 30 or 31, wherein the intermediate data is first intermediate data, the tensor metadata is indicative of a tensor shape, and the method further comprises:

receiving encoded second intermediate data and an encoded indication that the tensor metadata has not been included because the tensor shape has not changed;

decoding the encoded second intermediate data and the indication; and

determining that the second intermediate data is associated with the tensor shape based on the indication.

34. The method of one of claims 30-33, wherein the tensor metadata is indicative of at least one of: a tensor shape, tensor structure information, a data type, or an encoding algorithm.

35. The method of any one of claims 30-34, wherein decoding the encoded intermediate data comprises:

transforming a one-dimensional tensor into a multi-dimensional tensor; and

decoding the multi-dimensional tensor in accordance with a binary structure.

36. The method of any one of claims 30-35, wherein receiving the encoded intermediate data and the encoded tensor metadata from the second device comprises:

receiving the encoded intermediate data in a first band associated with a data channel; and

receiving the encoded tensor metadata in a second band associated with a control channel.

37. The method of any one of claims 30-36, wherein receiving the encoded intermediate data and the encoded tensor metadata from the second device comprises receiving an encoded bitstream from the second device, wherein the encoded bitstream comprises the encoded intermediate data and the encoded tensor metadata, and decoding the encoded intermediate data and the encoded tensor metadata comprises decoding the encoded bitstream.

38. The method of any one of claim 30-37, wherein decoding the encoded intermediate data and the encoded tensor metadata comprises decoding one or more of:

the encoded tensor metadata according to a tensor shape structure or algorithm to obtain the tensor shape; or

the encoded intermediate data according to a bitstream structure to obtain a series of intermediate data bytes.

39. The method of any one of claims 30-38, wherein reconstructing the tensor based on the intermediate data and the tensor metadata comprises:

converting binary data, in the intermediate data, to tensor data; and

reshaping the tensor to an original tensor shape based on the tensor data and the tensor metadata.

40. The method of any one of claims 30-39, wherein the first device is a wireless transmit/receive unit (WTRU) or a network entity.

100

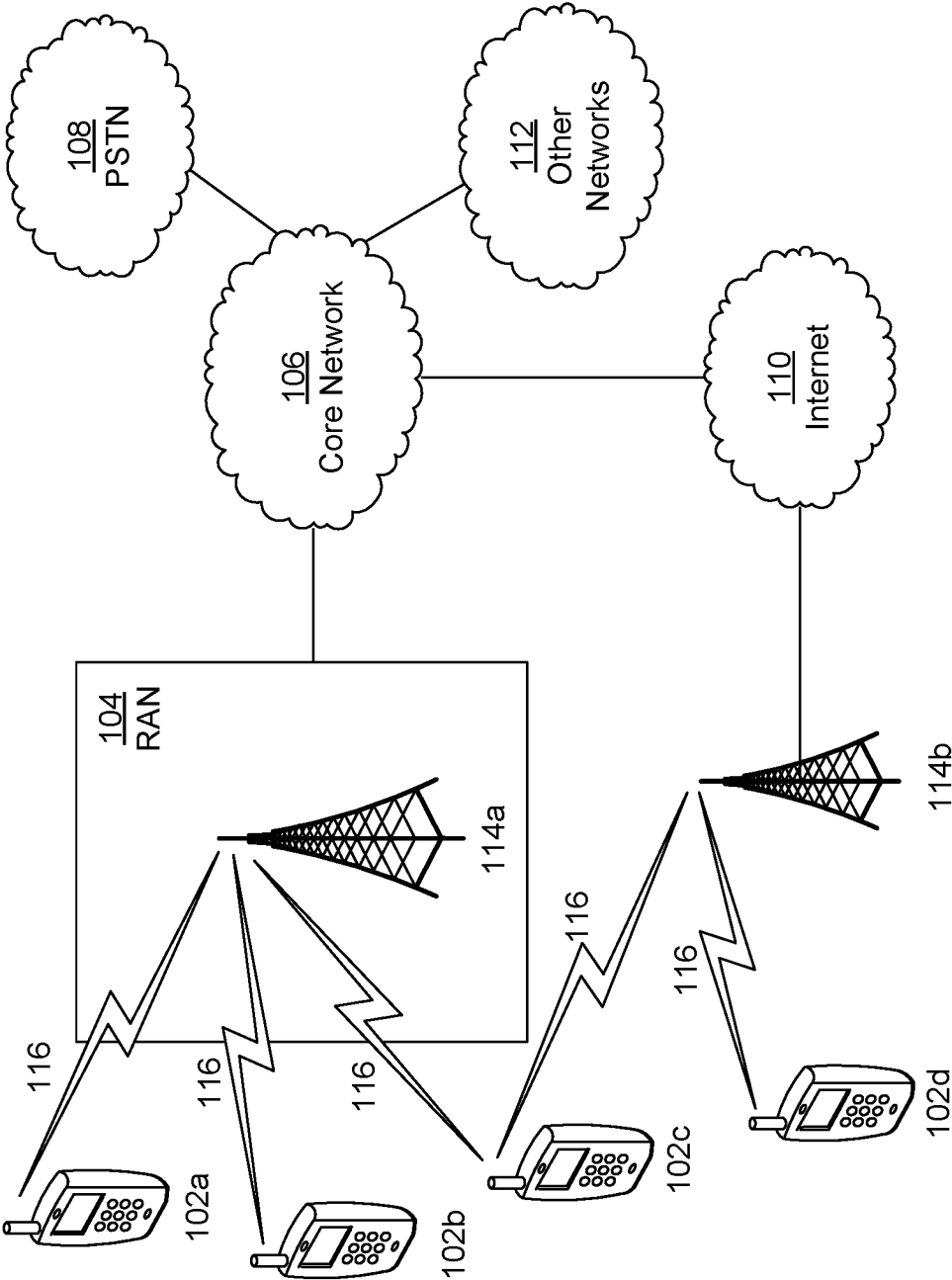


FIG. 1A

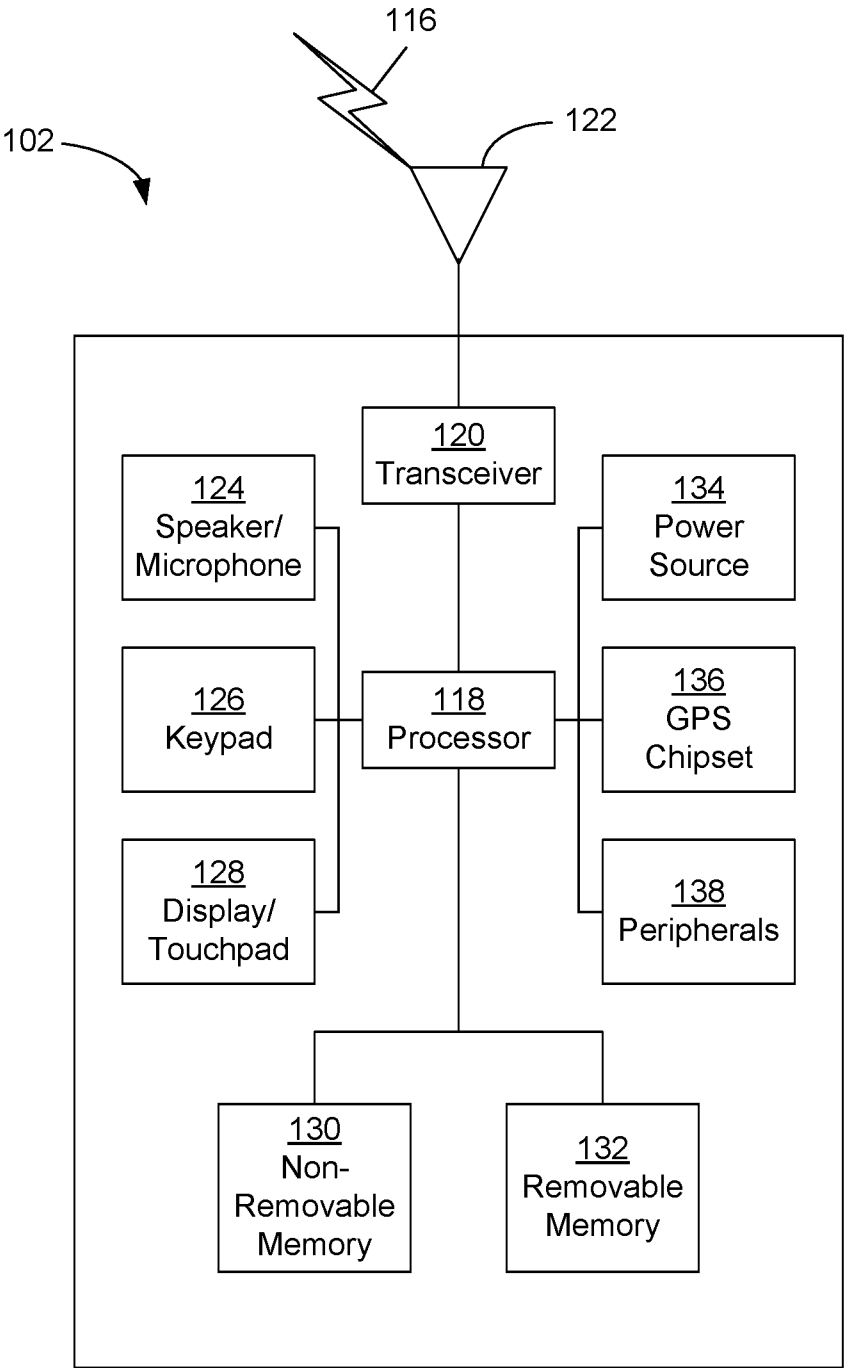


FIG. 1B

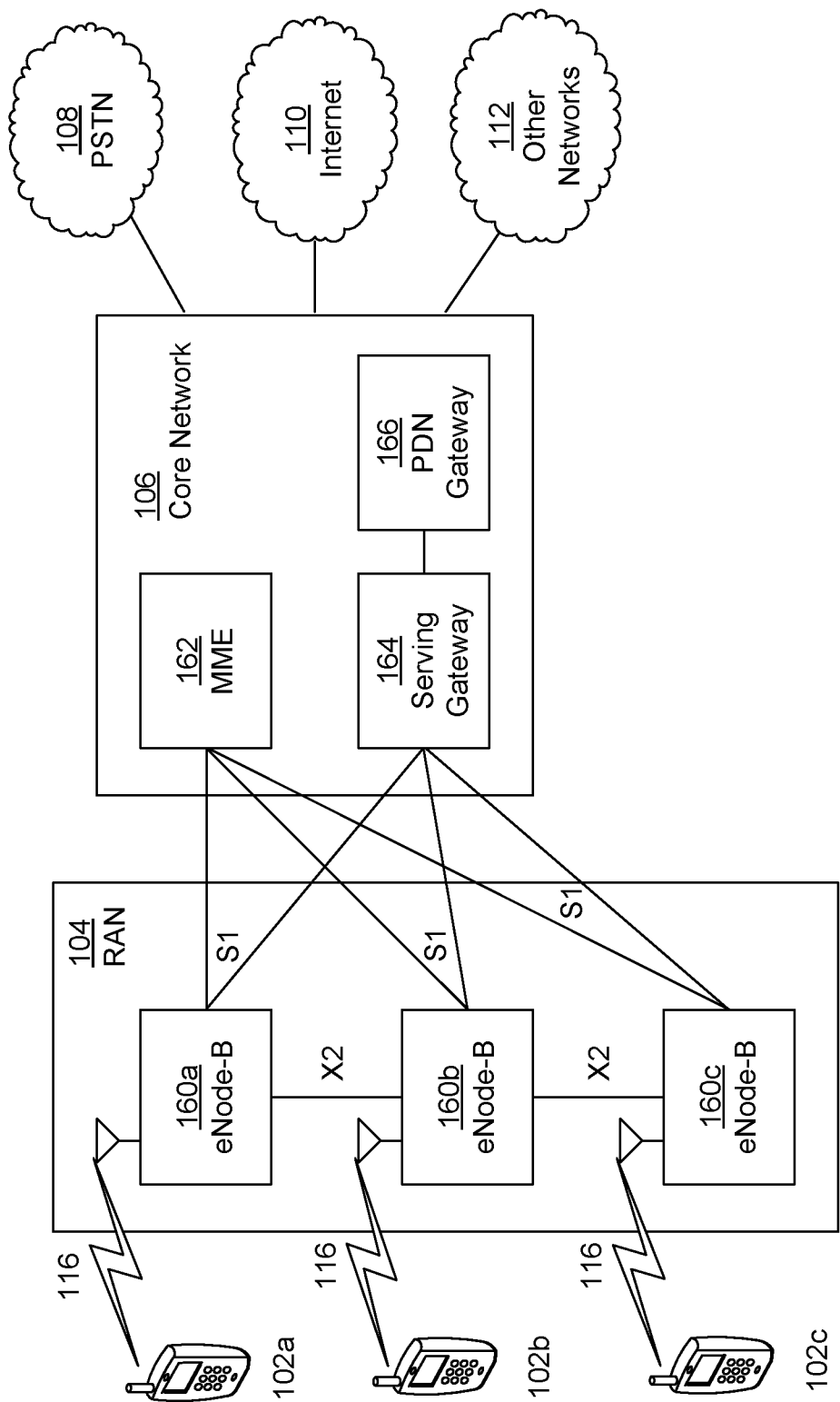


FIG. 1C

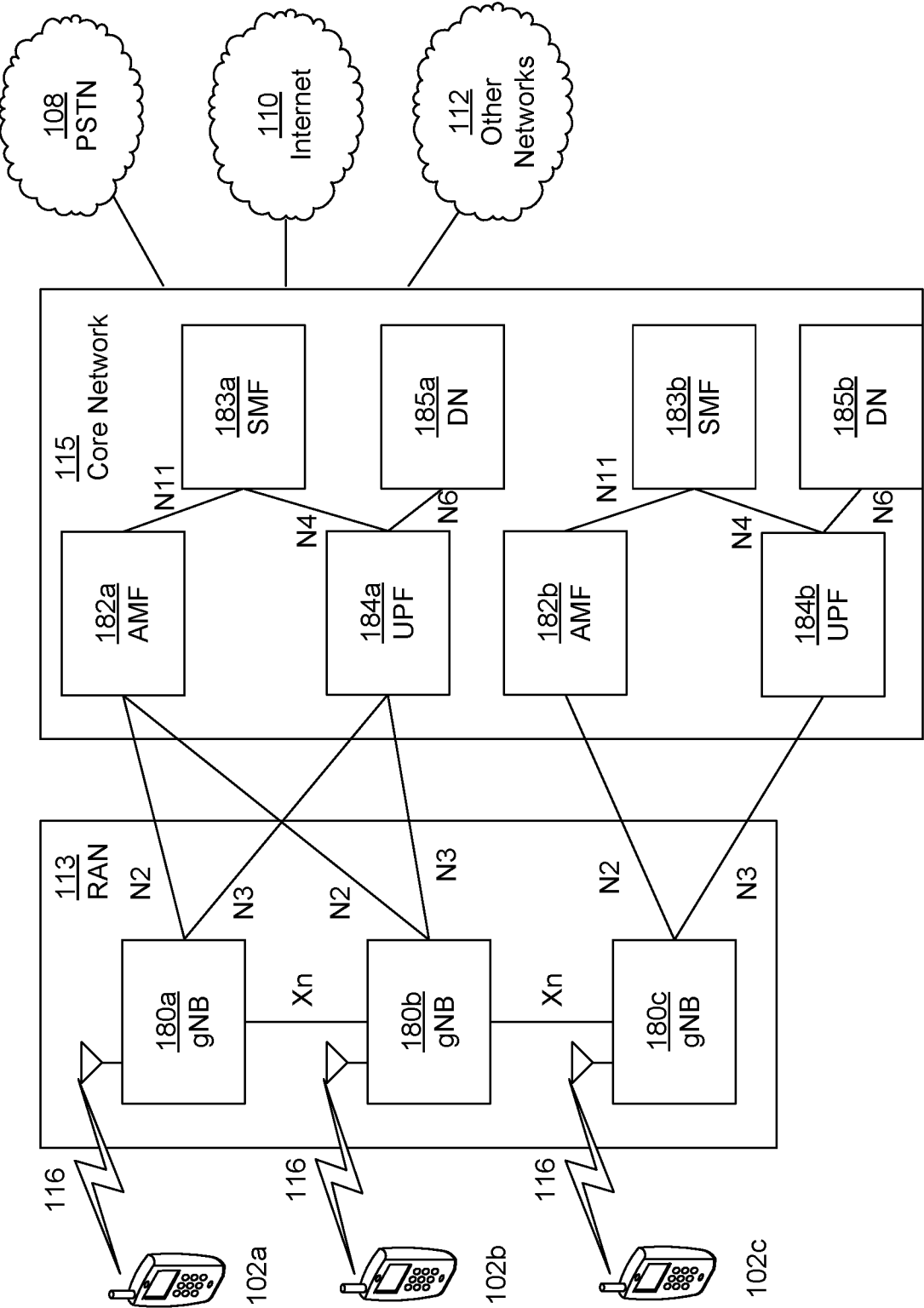


FIG. 1D

200

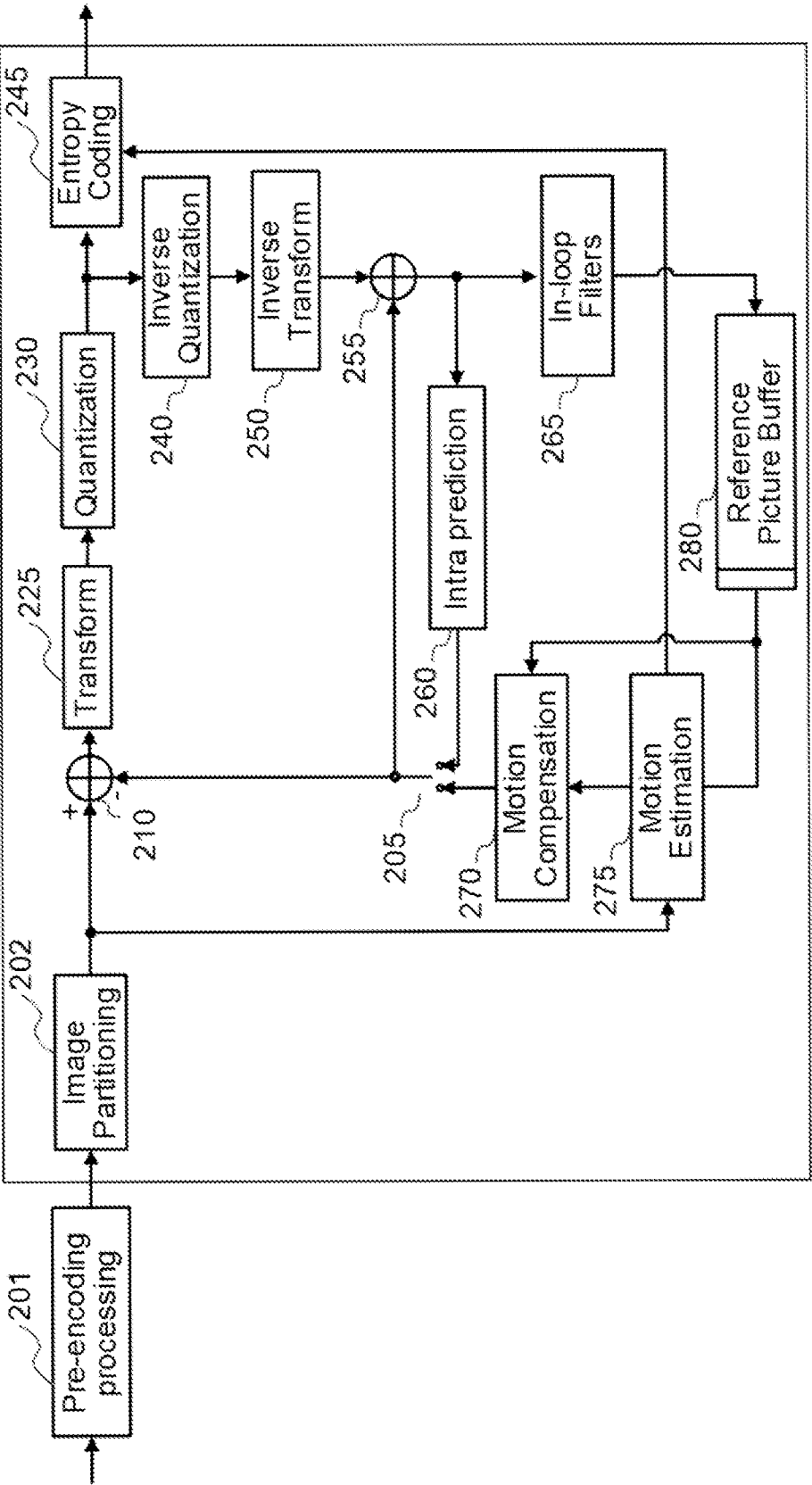


FIG. 2

300

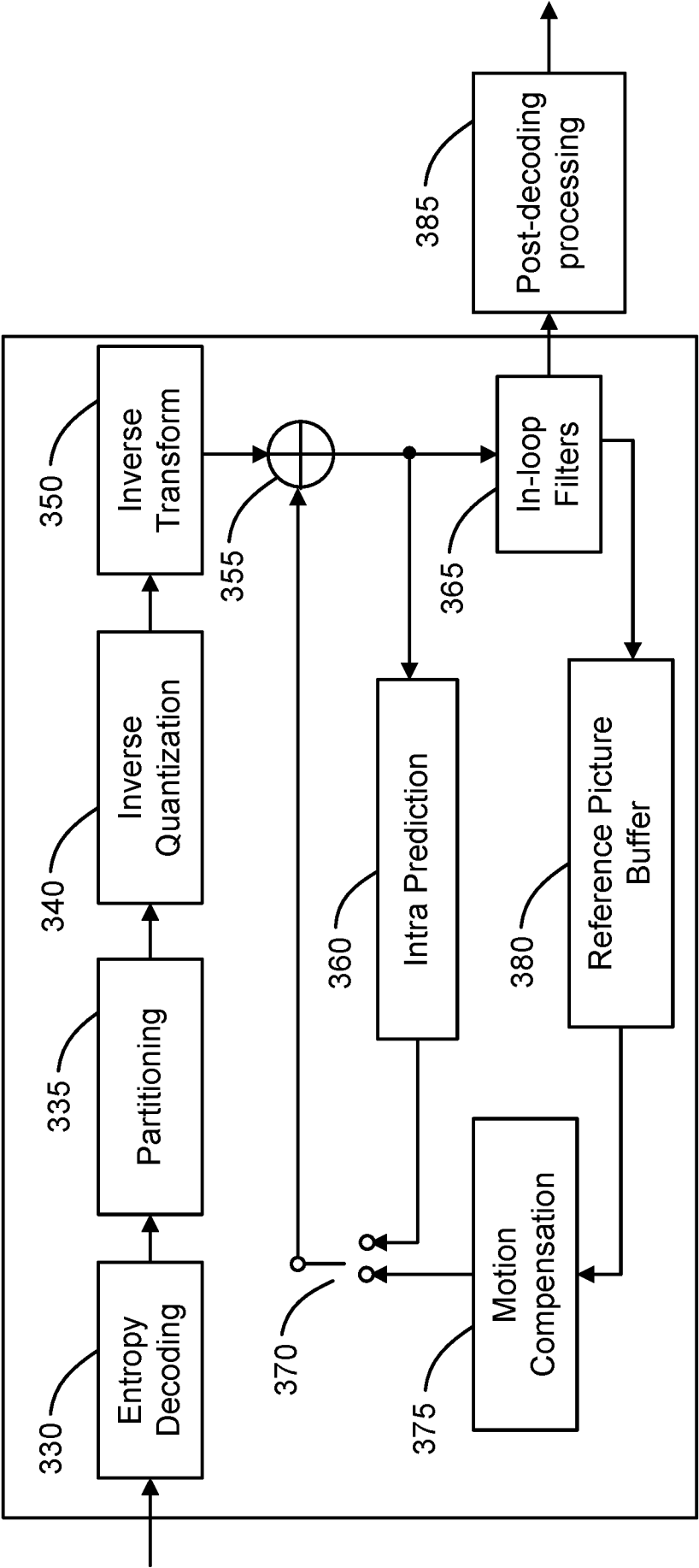


FIG. 3

400

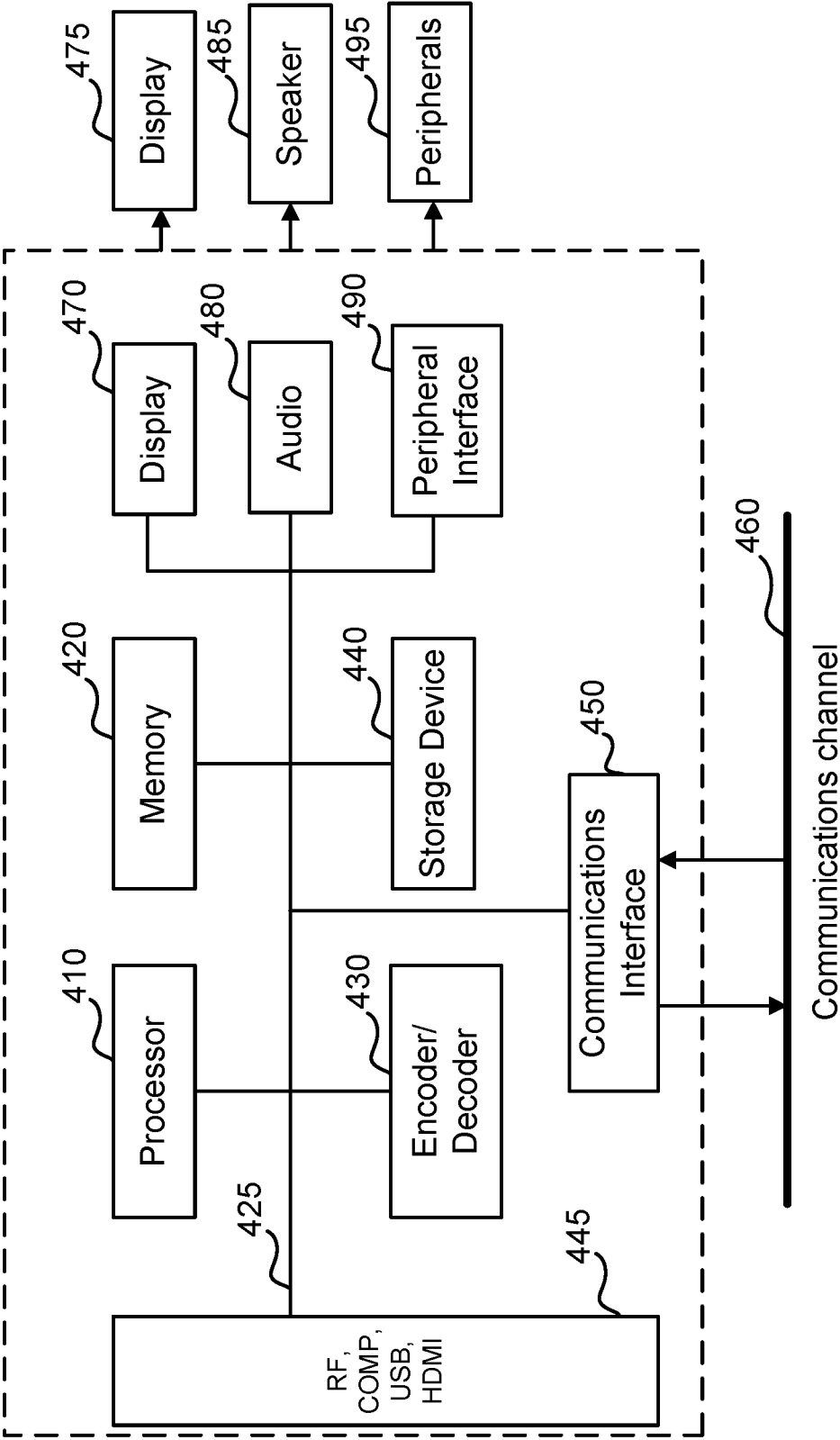


FIG. 4

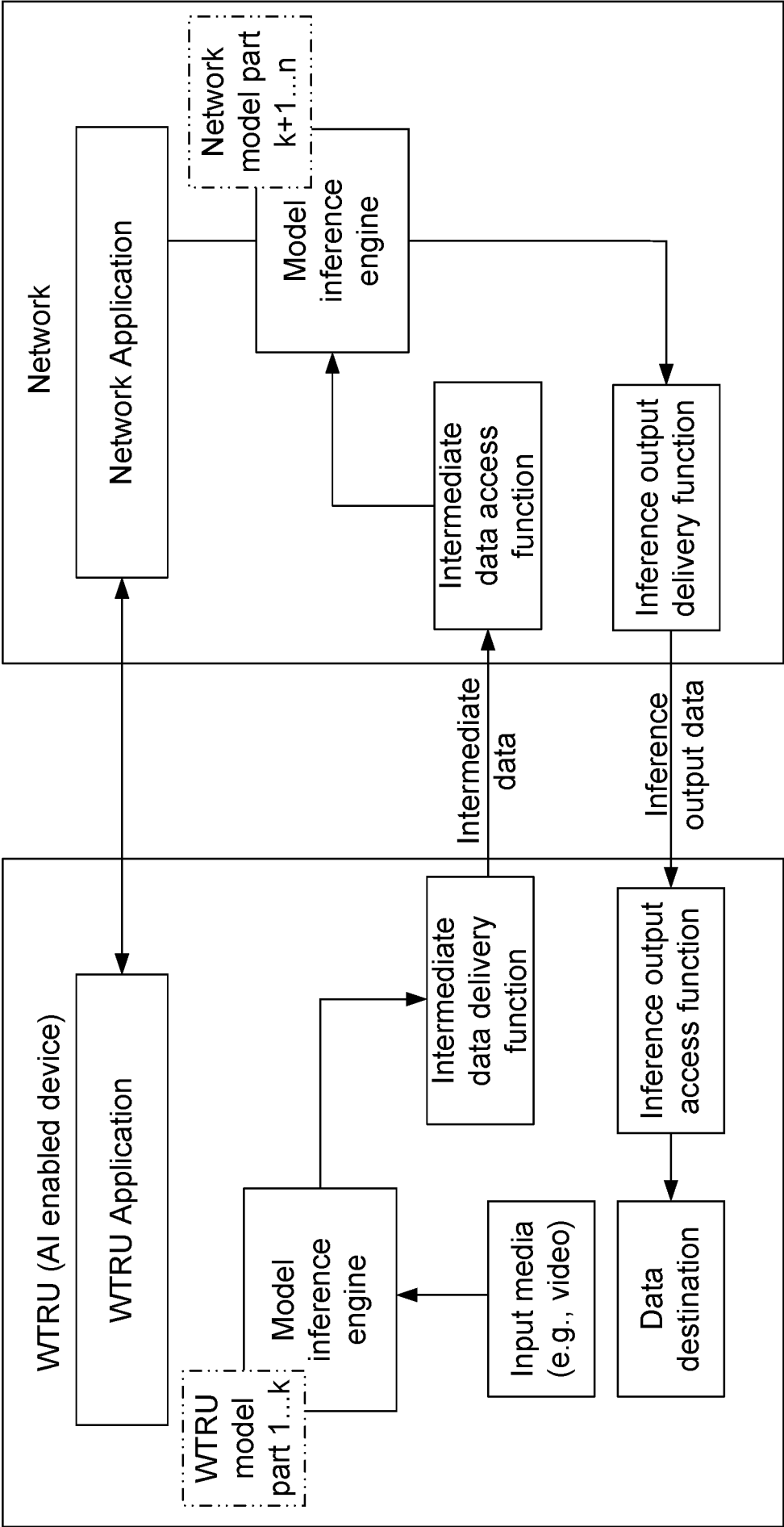


FIG. 5

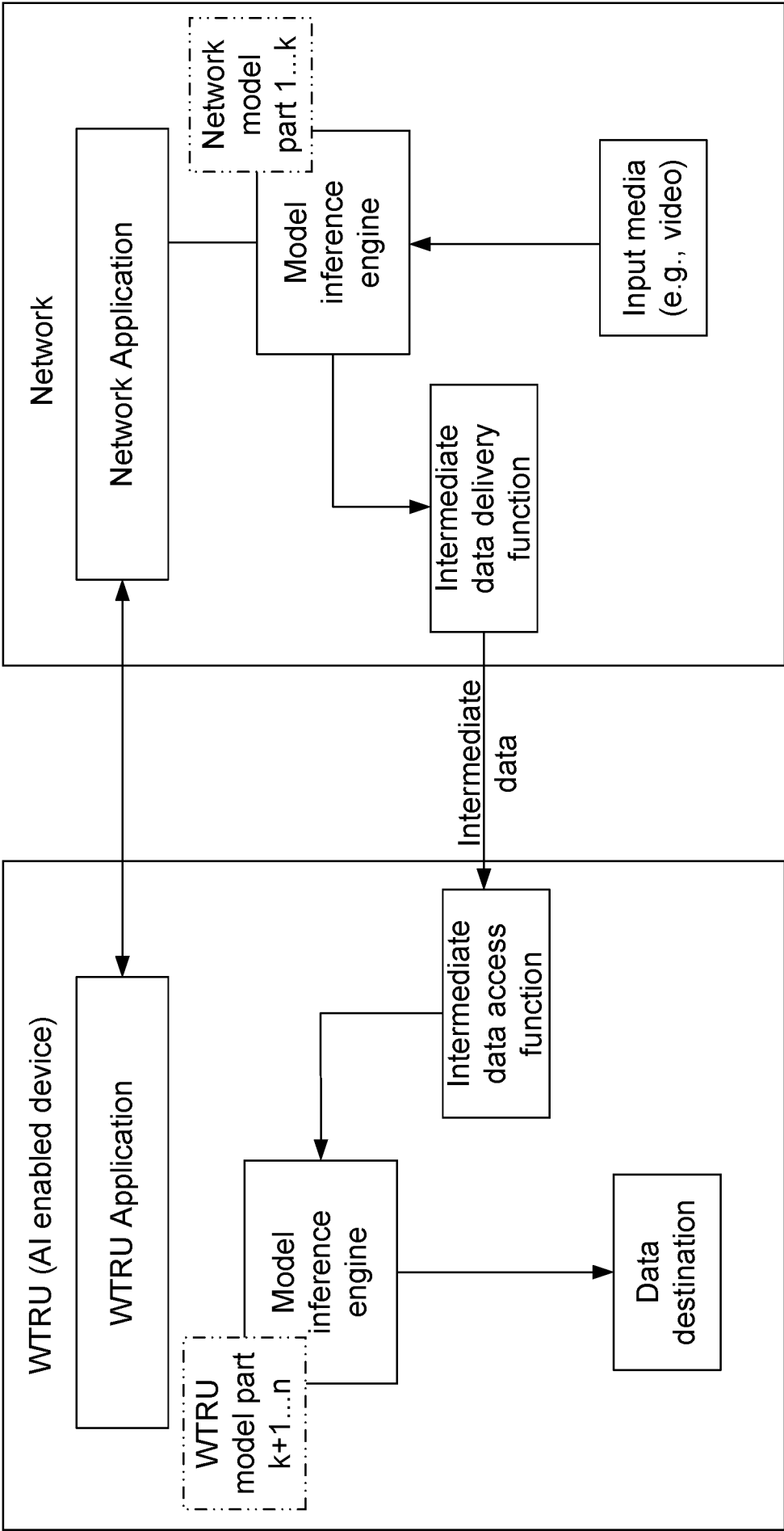


FIG. 6

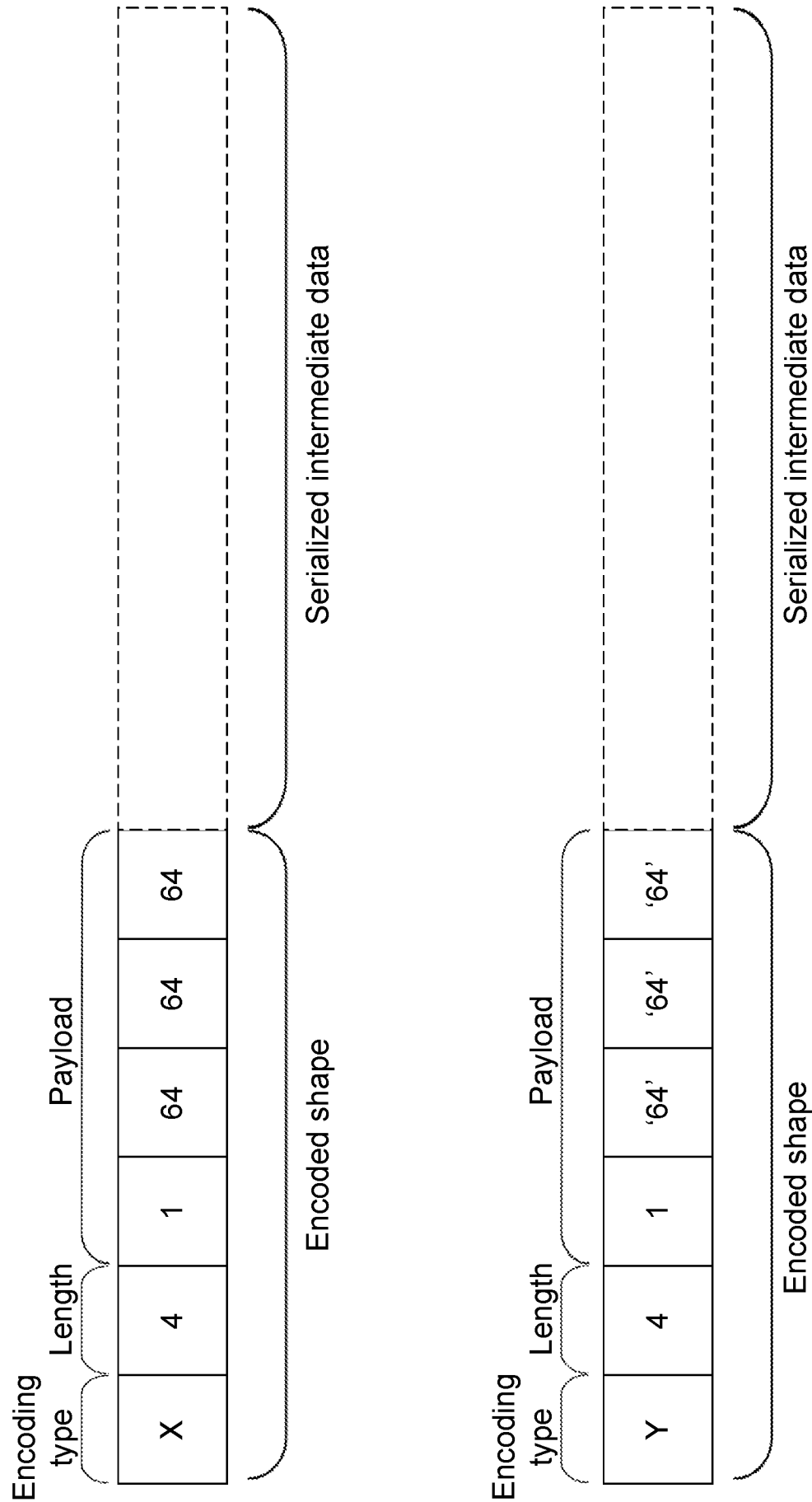


FIG. 7

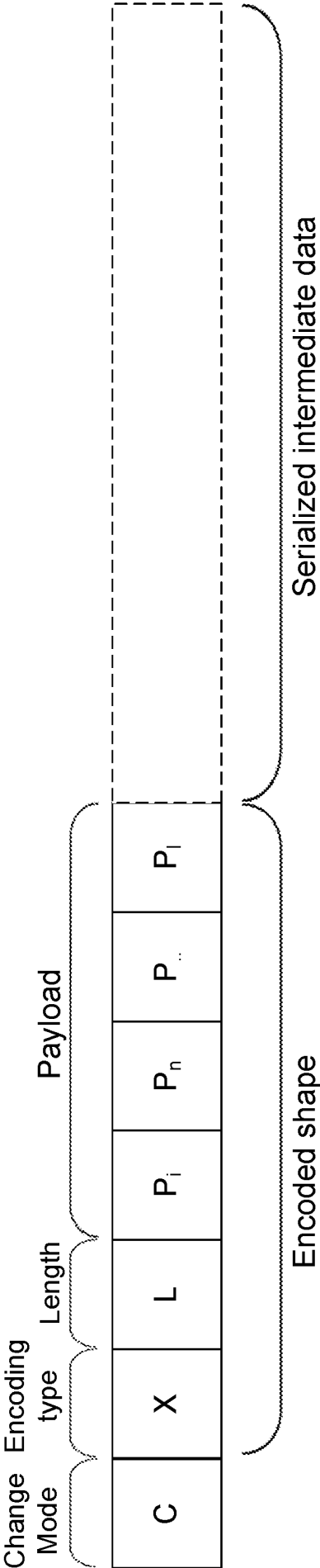


FIG. 8

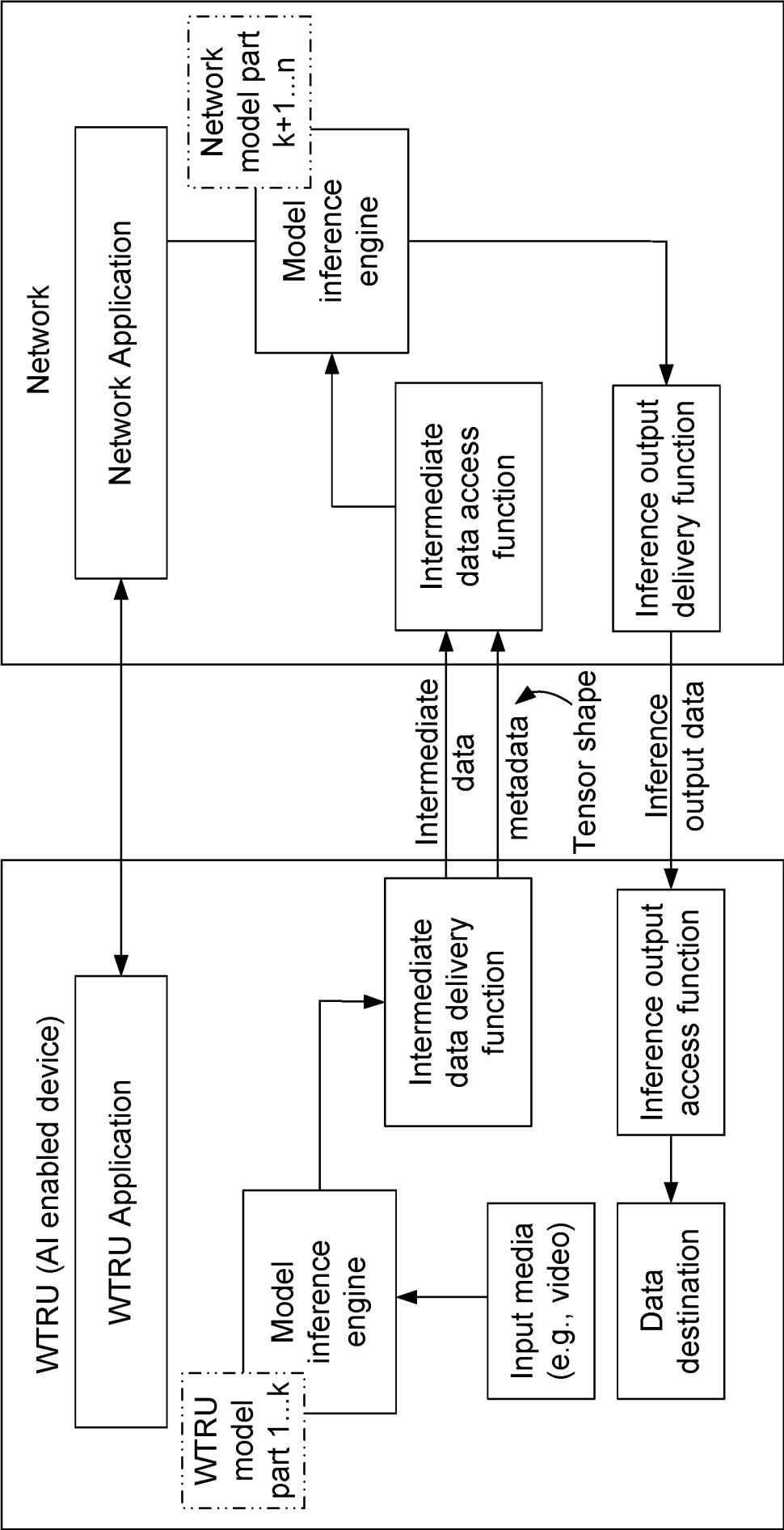


FIG. 9

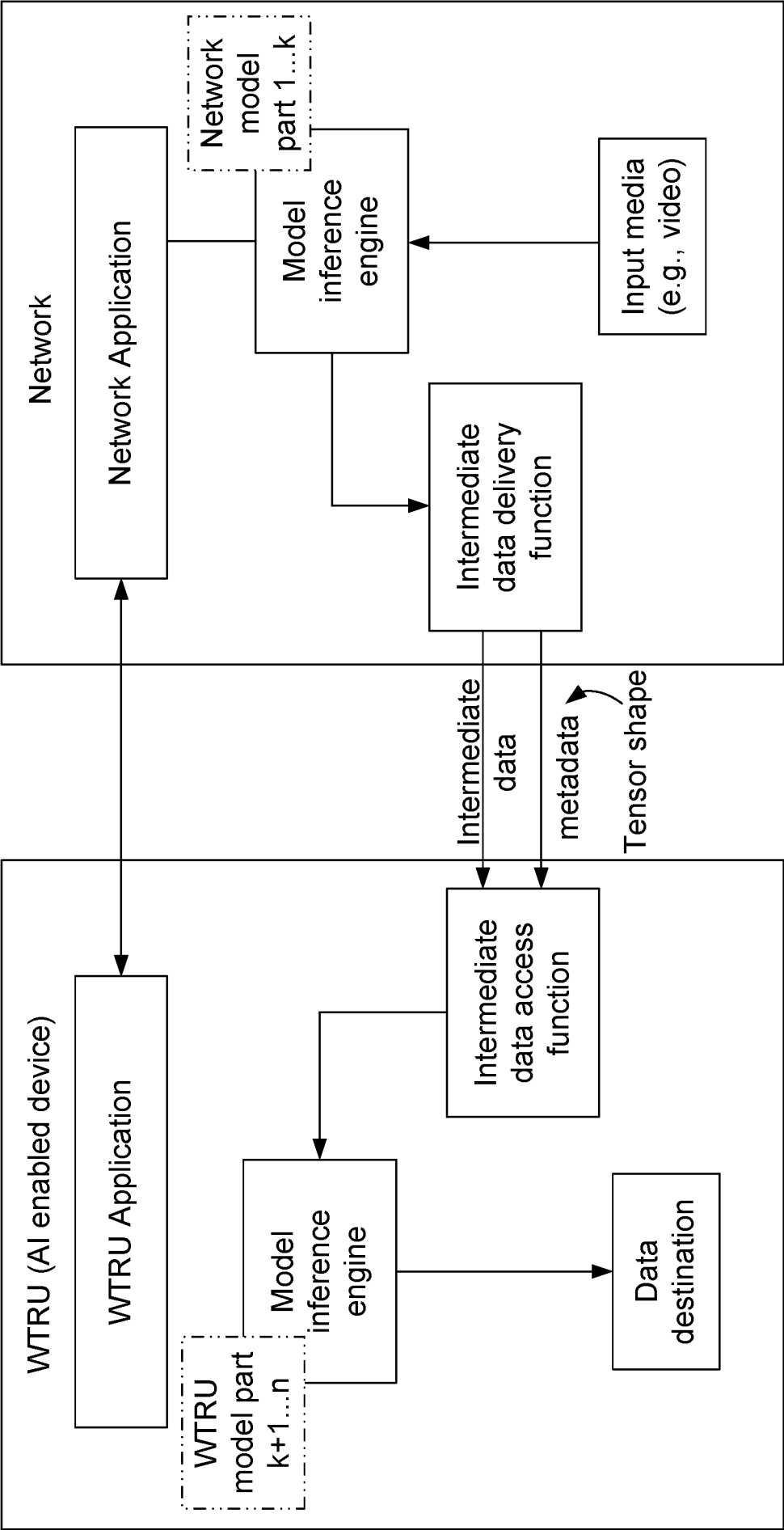


FIG. 10

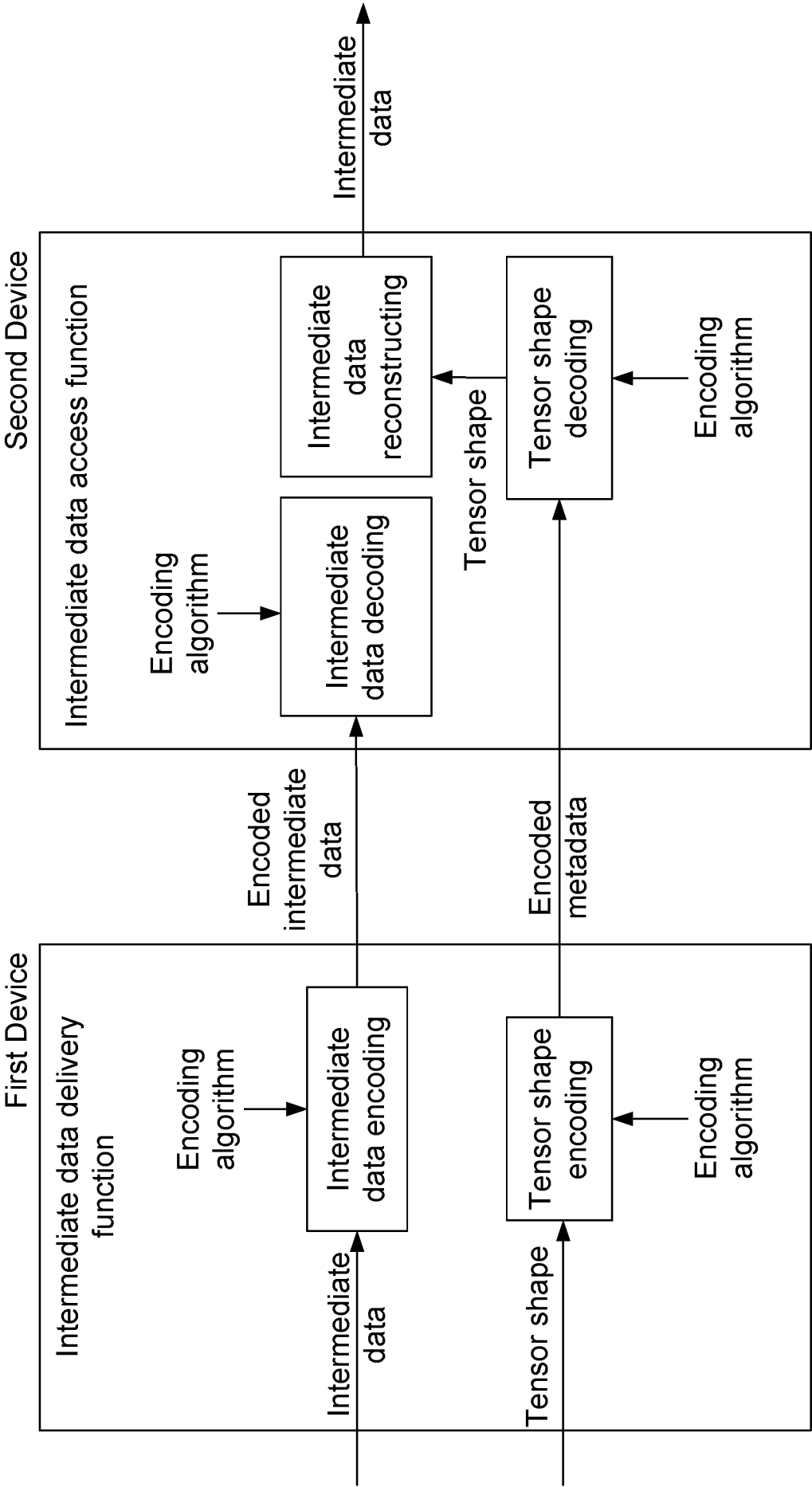


FIG. 11

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2024/037509

A. CLASSIFICATION OF SUBJECT MATTER

INV. H04N19/70 G06N3/045

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)

H04N 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	WO 2023/039627 A1 (CANON KK [JP]; CANON AUSTRALIA PTY LTD [AU]) 23 March 2023 (2023-03-23) paragraphs [0010] - [0021] paragraphs [0052] - [0066] paragraphs [0086] - [0096] paragraphs [0139] - [0148] -----	1 - 40
X	WO 2022/167547 A1 (INTERDIGITAL CE PATENT HOLDINGS SAS [FR]) 11 August 2022 (2022-08-11) paragraphs [0003] - [0006] paragraphs [0050] - [0104] ----- - / - -	1 - 40

☒ 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

2 October 2024

Date of mailing of the international search report

15/10/2024

Name and mailing address of the ISA/

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

Authorized officer

André, Thomas

INTERNATIONAL SEARCH REPORT

International application No

PCT/US2024/037509

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT		
Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	MATEEN ULHAQ: "Mobile-Cloud Inference for Collaborative Intelligence", ARXIV.ORG, CORNELL UNIVERSITY LIBRARY, 201 OLIN LIBRARY CORNELL UNIVERSITY ITHACA, NY 14853, 24 June 2023 (2023-06-24), XP091546989, Sections 1, 1.2, 2, 2.2.1, 2.3, 4.1 and 4.1.2 -----	1-40
X	HSIEH CHENG-YEN ET AL: "C3-SL: Circular Convolution-Based Batch-Wise Compression for Communication-Efficient Split Learning", 2022 IEEE 32ND INTERNATIONAL WORKSHOP ON MACHINE LEARNING FOR SIGNAL PROCESSING (MLSP), IEEE, 22 August 2022 (2022-08-22), pages 1-6, XP034229444, DOI: 10.1109/MLSP55214.2022.9943507 [retrieved on 2022-11-17] the whole document -----	1-40
A	Anonymous: "Serialization - Wikipedia", , 17 November 2017 (2017-11-17), XP055919266, Retrieved from the Internet: URL:https://en.wikipedia.org/w/index.php?t itle=Serialization&oldid=810836620 [retrieved on 2022-05-09] the whole document -----	6,15,18, 19,26, 35,38,39

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No

PCT/US2024/037509

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2023039627 A1	23-03-2023	AU 2021232739 A1	30-03-2023
		CN 117980926 A	03-05-2024
		EP 4402617 A1	24-07-2024
		JP 2024532068 A	05-09-2024
		KR 20240035564 A	15-03-2024
		WO 2023039627 A1	23-03-2023

WO 2022167547 A1	11-08-2022	CN 116940946 A	24-10-2023
		EP 4288907 A1	13-12-2023
		JP 2024509670 A	05-03-2024
		US 2024311621 A1	19-09-2024
		WO 2022167547 A1	11-08-2022
