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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
05 October 2023 (05.10.2023)

WIPO | PCT

(10) International Publication Number
WO 2023/191505 A1

(51) International Patent Classification:
H04L 43/02 (2022.01) *H04L 41/5003* (2022.01)
H04L 41/16 (2022.01) *H04L 67/14* (2022.01)
H04L 43/062 (2022.01) *H04W 88/14* (2009.01)

(21) International Application Number:
PCT/KR2023/004195

(22) International Filing Date:
29 March 2023 (29.03.2023)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:

2204491.1	29 March 2022 (29.03.2022)	GB
2212329.3	24 August 2022 (24.08.2022)	GB
2304171.8	22 March 2023 (22.03.2023)	GB

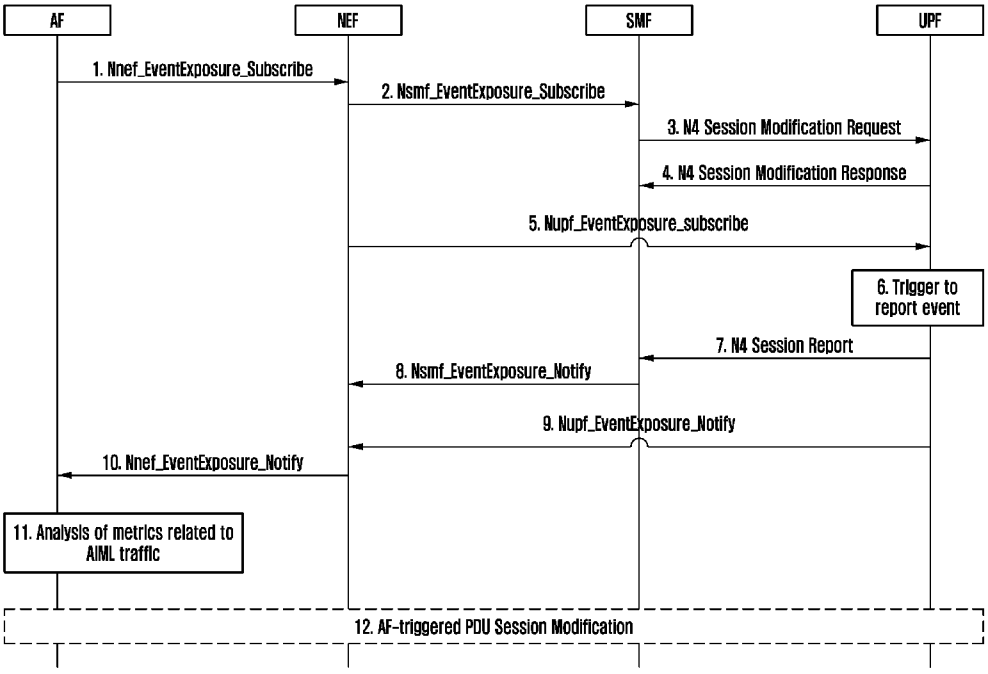
(71) Applicant: **SAMSUNG ELECTRONICS CO., LTD.** [KR/KR]; 129, Samsung-ro, Yeongtong-gu, Suwon-si, Gyeonggi-do 16677 (KR).

(72) Inventor: **ESTEVEZ, David Gutierrez**; Samsung Electronics (UK) Limited Samsung Electronics, Research Institute, Communications House, South Street Staines Middlesex TW18 4QE (GB).

(74) Agent: **YOON & LEE INTERNATIONAL PATENT & LAW FIRM**; 3rd Fl, Ace Highend Tower-5, 226, Gasan Digital 1-ro, Geumcheon-gu, Seoul 08502 (KR).

(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,

(54) Title: MONITORING FOR APPLICATION AI/ML-BASED SERVICES AND OPERATIONS



(57) Abstract: A 5G or 6G communication system for supporting a higher data transmission rate. A method for monitoring events in a network comprises: subscribing, by an Application Function (AF), to monitoring the events, wherein the AF supports one or more AIML-based services and/or operations, and the events relate to the performance of the AIML-based services and/or operations; in response to the subscribing, monitoring, by a User Plane Function (UPF), the events; and when an event is detected, reporting, by the UPF to a Network Exposure Function (NEF), the event. The monitoring events may comprise one or more of: monitoring session inactivity of a PDU session; monitoring traffic volume of a PDU session; monitoring PCF events; per-5QI monitoring; and monitoring of edge resources. The monitoring may comprise QoS monitoring and may be performed for multiple UEs.

WO 2023/191505 A1

[Continued on next page]

MA, MD, MG, MK, MN, MU, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

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

Published:

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

Description

Title of Invention: MONITORING FOR APPLICATION AI/ ML-BASED SERVICES AND OPERATIONS

Technical Field

[1] Certain examples of the present disclosure provide techniques relating to Artificial Intelligence (AI) and/or Machine Leaning (ML), in particular techniques relating to monitoring application AI/ML operation. For example, certain examples of the present disclosure provide methods, apparatus and systems in 3rd Generation Partnership Project (3GPP) 5th Generation (5G) System (5GS) for monitoring one or more features (e.g., network resource utilization and/or Quality of Service (QoS)) for application AI and/or ML operation.

Background Art

[2] At the beginning of the development of 5G mobile communication technologies, in order to support services and to satisfy performance requirements in connection with enhanced Mobile BroadBand (eMBB), Ultra Reliable Low Latency Communications (URLLC), and massive Machine-Type Communications (mMTC), there has been ongoing standardization regarding beamforming and massive MIMO for mitigating radio-wave path loss and increasing radio-wave transmission distances in mmWave, supporting numerologies (for example, operating multiple subcarrier spacings) for efficiently utilizing mmWave resources and dynamic operation of slot formats, initial access technologies for supporting multi-beam transmission and broadbands, definition and operation of BWP (BandWidth Part), new channel coding methods such as a LDPC (Low Density Parity Check) code for large amount of data transmission and a polar code for highly reliable transmission of control information, L2 pre-processing, and network slicing for providing a dedicated network specialized to a specific service.

[3] Currently, there are ongoing discussions regarding improvement and performance enhancement of initial 5G mobile communication technologies in view of services to be supported by 5G mobile communication technologies, and there has been physical layer standardization regarding technologies such as V2X (Vehicle-to-everything) for aiding driving determination by autonomous vehicles based on information regarding positions and states of vehicles transmitted by the vehicles and for enhancing user convenience, NR-U (New Radio Unlicensed) aimed at system operations conforming to various regulation-related requirements in unlicensed bands, NR UE Power Saving, Non-Terrestrial Network (NTN) which is UE-satellite direct communication for providing coverage in an area in which communication with terrestrial networks is unavailable, and positioning.

[4] Moreover, there has been ongoing standardization in air interface architecture/ protocol regarding technologies such as Industrial Internet of Things (IIoT) for supporting new services through interworking and convergence with other industries, IAB (Integrated Access and Backhaul) for providing a node for network service area expansion by supporting a wireless backhaul link and an access link in an integrated manner, mobility enhancement including conditional handover and DAPS (Dual Active Protocol Stack) handover, and two-step random access for simplifying random access procedures (2-step RACH for NR). There also has been ongoing standardization in system architecture/service regarding a 5G baseline architecture (for example, service based architecture or service based interface) for combining Network Functions Virtualization (NFV) and Software-Defined Networking (SDN) technologies, and Mobile Edge Computing (MEC) for receiving services based on UE positions.

[5] As 5G mobile communication systems are commercialized, connected devices that have been exponentially increasing will be connected to communication networks, and it is accordingly expected that enhanced functions and performances of 5G mobile communication systems and integrated operations of connected devices will be necessary. To this end, new research is scheduled in connection with eXtended Reality (XR) for efficiently supporting AR (Augmented Reality), VR (Virtual Reality), MR (Mixed Reality) and the like, 5G performance improvement and complexity reduction by utilizing Artificial Intelligence (AI) and Machine Learning (ML), AI service support, metaverse service support, and drone communication.

[6] Furthermore, such development of 5G mobile communication systems will serve as a basis for developing not only new waveforms for providing coverage in terahertz bands of 6G mobile communication technologies, multi-antenna transmission technologies such as Full Dimensional MIMO (FD-MIMO), array antennas and large-scale antennas, metamaterial-based lenses and antennas for improving coverage of terahertz band signals, high-dimensional space multiplexing technology using OAM (Orbital Angular Momentum), and RIS (Reconfigurable Intelligent Surface), but also full-duplex technology for increasing frequency efficiency of 6G mobile communication technologies and improving system networks, AI-based communication technology for implementing system optimization by utilizing satellites and AI (Artificial Intelligence) from the design stage and internalizing end-to-end AI support functions, and next-generation distributed computing technology for implementing services at levels of complexity exceeding the limit of UE operation capability by utilizing ultra-high-performance communication and computing resources.

Disclosure of Invention

Technical Problem

[7] Aspects of the disclosure are to address at least the above-mentioned problems and/or disadvantages and to provide at least the advantages described below. Accordingly, an aspect of the disclosure is to provide a method for detecting and recovering a beam failure resulting from a channel state change signals in a wireless communication system.

[8] Additional aspects will be set forth in part in the description, which follows and, in part, will be apparent from the description, or may be learned by practice of the presented embodiments.

[9] According to an embodiment of the disclosure monitoring for application AI/ ML-based services and operations are needed.

Solution to Problem

[10] In accordance with an aspect of the disclosure, a method performed by a network exposure function (NEF) for monitoring an event in a network is provided. The method comprises receiving, from an application function (AF) which supports one or more artificial intelligence/machine leaning (AI/ML)-based services, first information indicating a subscription for monitoring an event related to the AI/ML-based services, transmitting, to a session management function (SMF), second information for obtaining information on the event related to an inactivity time, transmitting, to a user plane function (UPF), third information for obtaining information on the event related to a traffic volume, receiving at least one of the information on the event related to the inactivity time or the information to the event related to the traffic volume and transmitting, to the AF, one or more parameter obtained based on the at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume.

[11] In accordance with another aspect of the disclosure, a method performed by a user plane function (UPF) for monitoring an event in a network is provided. The method comprises receiving, from a session management function (SMF), first information for obtaining information on the event related to an inactivity time, receiving, from a network exposure function (NEF), second information for obtaining information on the event related to a traffic volume, monitoring at least one event based on the first information and the second information, and reporting at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume, in case that a trigger happens to report the at least one event.

[12] In accordance with another aspect of the disclosure, a method performed by an application function (AF) for monitoring an event in a network is provided. The method comprises transmitting, to a network exposure function (NEF), first information indicating a subscription for monitoring an event related to artificial intelligence/machine

leaning (AI/ML), receiving, from the NEF, one or more parameter obtained based on at least one of information on the event related to an inactivity time or information on the event related to a traffic volume and analyzing the one or more parameter for AI/ML traffic, wherein the AF supports one or more AI/ML-based services.

[13] In accordance with another aspect of the disclosure, a network exposure function (NEF) for monitoring an event in a network is provided. The NEF comprises a transceiver and a controller configured to receive, from an application function (AF) which supports one or more artificial intelligence/machine leaning (AI/ML)-based services, first information indicating a subscription for monitoring an event related to the AI/ML-based services, to transmit, to a session management function (SMF), second information for obtaining information on the event related to an inactivity time, to transmit, to a user plane function (UPF), third information for obtaining information on the event related to a traffic volume, to receive at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume, and to transmit, to the AF, one or more parameter obtained based on the at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume.

[14] In accordance with another aspect of the disclosure, a user plane function (UPF) for monitoring an event in a network is provided. The UPF comprises a transceiver and a controller configured to receive from a session management function (SMF), first information for obtaining information on the event related to an inactivity time, to receive from a network exposure function (NEF), second information for obtaining information on the event related to a traffic volume, to monitor at least one event based on the first information and the second information, and to report at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume, in case that a trigger happens to report the at least one event.

[15] In accordance with another aspect of the disclosure, an application function (AF) for monitoring an event in a network is provided. The AF comprises a transceiver; and a controller configured to transmit, to a network exposure function (NEF), first information indicating a subscription for monitoring an event related to artificial intelligence/machine leaning (AI/ML) to receive from the NEF, one or more parameter obtained based on at least one of information on the event related to an inactivity time or information on the event related to a traffic volume, and to analyze the one or more parameter for AI/ML traffic wherein the AF supports one or more AI/ML-based services.

Advantageous Effects of Invention

[16] Advantages, and salient features of the invention will become apparent to those skilled in the art from the following detailed description, which, taken in conjunction with the annexed drawings, discloses exemplary embodiments of the invention.

Brief Description of Drawings

[17] For a more complete understanding of the present disclosure and its advantages, reference is now made to the following description taken in conjunction with the accompanying drawings, in which like reference numerals represent like parts:

[18] Figure 1 illustrates an example call flow according to an example of the present disclosure;

[19] Figure 2 illustrates an example call flow according to another example of the present disclosure; and

[20] Figure 3 illustrates a block diagram of an exemplary network entity that may be used in certain examples of the present disclosure.

Mode for the Invention

[21] It is an aim of certain examples of the present disclosure to address, solve and/or mitigate, at least partly, at least one of the problems and/or disadvantages associated with the related art, for example at least one of the problems and/or disadvantages described herein. It is an aim of certain examples of the present disclosure to provide at least one advantage over the related art, for example at least one of the advantages described herein.

[22] The present disclosure is defined in the independent claims. Advantageous features are defined in the dependent claims.

[23] Embodiments or examples disclosed in the description and/or figures falling outside the scope of the claims are to be understood as examples useful for understanding the present disclosure.

[24] Other aspects, advantages and salient features of the disclosure will become apparent to those skilled in the art from the following detailed description taken in conjunction with the accompanying drawings.

[25] Before undertaking the DETAILED DESCRIPTION below, it may be advantageous to set forth definitions of certain words and phrases used throughout this patent document: the terms "include" and "comprise," as well as derivatives thereof, mean inclusion without limitation; the term "or," is inclusive, meaning and/or; the phrases "associated with" and "associated therewith," as well as derivatives thereof, may mean to include, be included within, interconnect with, contain, be contained within, connect to or with, couple to or with, be communicable with, cooperate with, interleave, juxtapose, be proximate to, be bound to or with, have, have a property of, or the like; and the term "controller" means any device, system or part thereof that controls at least

one operation, such a device may be implemented in hardware, firmware or software, or some combination of at least two of the same. It should be noted that the functionality associated with any particular controller may be centralized or distributed, whether locally or remotely.

[26] Moreover, various functions described below can be implemented or supported by one or more computer programs, each of which is formed from computer readable program code and embodied in a computer readable medium. The terms "application" and "program" refer to one or more computer programs, software components, sets of instructions, procedures, functions, objects, classes, instances, related data, or a portion thereof adapted for implementation in a suitable computer readable program code. The phrase "computer readable program code" includes any type of computer code, including source code, object code, and executable code. The phrase "computer readable medium" includes any type of medium capable of being accessed by a computer, such as read only memory (ROM), random access memory (RAM), a hard disk drive, a compact disc (CD), a digital video disc (DVD), or any other type of memory. A "non-transitory" computer readable medium excludes wired, wireless, optical, or other communication links that transport transitory electrical or other signals. A non-transitory computer readable medium includes media where data can be permanently stored and media where data can be stored and later overwritten, such as a rewritable optical disc or an erasable memory device.

[27] Definitions for certain words and phrases are provided throughout this patent document, those of ordinary skill in the art should understand that in many, if not most instances, such definitions apply to prior, as well as future uses of such defined words and phrases.

[28] FIGS. 1 through 3, discussed below, and the various embodiments used to describe the principles of the present disclosure in this patent document are by way of illustration only and should not be construed in any way to limit the scope of the disclosure. Those skilled in the art will understand that the principles of the present disclosure may be implemented in any suitably arranged system or device.

[29] Herein, the following documents are referenced:

[30] [1] 3GPP TS 22.261 (e.g., V18.5.0)

[31] [2] 3GPP TS 23.501 (e.g., V17.3.0)

[32] [3] 3GPP TS 23.502 (e.g., V17.3.0)

[33] [4] 3GPP TS 23.288 (e.g., V17.2.0)

[34] AI/ML is being used in a range of application domains across industry sectors. In mobile communications systems, conventional algorithms (e.g., speech recognition, image recognition, video processing) in mobile devices (e.g., smartphones, automotive, robots) are being increasingly replaced with AI/ML models to enable various ap-

plications.

[35] The 5G system can support various types of AI/ML operations, in including the following three defined in [1]:

[36] AI/ML operation splitting between AI/ML endpoints

[37] The AI/ML operation/model may be split into multiple parts, for example according to the current task and environment. The intention is to offload the computation-intensive, energy-intensive parts to network endpoints, and to leave the privacy-sensitive and delay-sensitive parts at the end device. The device executes the operation/model up to a specific part/layer and then sends the intermediate data to the network endpoint. The network endpoint executes the remaining parts/layers and feeds the inference results back to the device.

[38] AI/ML model/data distribution and sharing over 5G system

[39] Multi-functional mobile terminals may need to switch an AI/ML model, for example in response to task and environment variations. An assumption of adaptive model selection is that the models to be selected are available for the mobile device. However, since AI/ML models are becoming increasingly diverse, and with the limited storage resource in a UE, not all candidate AI/ML models may be pre-loaded on-board. Online model distribution (i.e., new model downloading) may be needed, in which an AI/ML model can be distributed from a Network (NW) endpoint to the devices when they need it to adapt to the changed AI/ML tasks and environments. For this purpose, the model performance at the UE may need to be monitored constantly.

[40] Distributed/Federated Learning over 5G system

[41] A cloud server may train a global model by aggregating local models partially trained by each of a number of end devices (e.g., UEs). Within each training iteration, a UE performs the training based on a model downloaded from the AI server using local training data. Then the UE reports the interim training results to the cloud server, for example via 5G UL channels. The server aggregates the interim training results from the UEs and updates the global model. The updated global model is then distributed back to the UEs and the UEs can perform the training for the next iteration.

[42] Different levels of interactions are expected between UE and AF as AI/ML endpoints, for example based on [1], to exchange AI/ML models, intermediate data, local training data, inference results and/or model performance as Application AI/ML traffic. However support for the transmission of Application AI/ML traffic, for example over 5GS, between AI/ML endpoints (e.g., UE and AF) as described above is not currently defined in the existing 5GC data transfer/traffic routing mechanisms.

[43] 3GPP approved a study item for Rel-18 in SA WG2 focused on 5G System Support for AI/ML-based Services. As part of this study, an issue (i.e., problem statement) was approved with the scope of monitoring the network resource utilization for support of

Application AI/ML operations. Another issue was also approved with the scope of Quality of Service (QoS) and policy enhancements for AI/ML-based services. As part of this issue, it was agreed to study QoS monitoring aspects to support the operation of the 3rd party AI/ML operation.

[44] NWDAF represents an (operator-managed) network analytics logical function providing (slice specific) network data analytics to NFs and/or AFs. A NF or AF may subscribe to network analytics provided by NWDAF. NWDAF collects data from NFs, AFs and/or OAM and derives network analytics. NWDAF provides suitable network analytics to subscribed NFs and/or AFs, for example based on triggering events.

[45] The following is stated in 3GPP TS 23.501 V17.2.0, Clause 6.2.18:

[46] The Network Data Analytics Function (NWDAF) includes one or more of the following functionalities:

[47] - Support data collection from NFs and AFs;

[48] - Support data collection from OAM;

[49] - NWDAF service registration and metadata exposure to NFs and AFs;

[50] - Support analytics information provisioning to NFs and AFs;

[51] - Support Machine Learning (ML) model training and provisioning to NWDAFs (containing Analytics logical function).The details of the NWDAF functionality are defined in TS 23.288 [86].

[52] The following is stated in 3GPP TS 23.288 V17.2.0, Clause 4.1:

[53] The NWDAF (Network Data Analytics Function) is part of the architecture specified in TS 23.501 [2] and uses the mechanisms and interfaces specified for 5GC in TS 23.501 [2] and OAM services (see clause 6.2.3.1).

[54] The NWDAF interacts with different entities for different purposes:

[55] - Data collection based on subscription to events provided by the AMF, SMF, PCF, UDM, AF (directly or via NEF), and OAM;

[56] - [Optionally] Analytics and Data collection using the DCCF (Data Collection Coordination Function);

[57] - Retrieval of information from data repositories (e.g., UDR via UDM for subscriber-related information);

[58] - [Optionally] Storage and retrieval of information from the ADRF (Analytics Data Repository Function);

[59] - [Optionally] Analytics and Data collection from the MFAF (Messaging Framework Adaptor Function);

[60] - Retrieval of information about NFs (e.g., from NRF for NF-related information);

[61] - On demand provision of analytics to consumers, as specified in clause 6.

[62] - Provision of bulked data to consumers, as specified in clause 6.

[63] A single instance or multiple instances of NWDAF may be deployed in a PLMN. If

multiple NWDAF instances are deployed, the architecture supports deploying the NWDAF as a central NF, as a collection of distributed NFs, or as a combination of both. If multiple NWDAF instances are deployed, an NWDAF can act as an aggregate point (i.e., Aggregator NWDAF) and collect analytics information from other NWDAFs, which may have different Serving Areas, to produce the aggregated analytics (per Analytics ID), possibly with Analytics generated by itself.

[64] NOTE 1: When multiple NWDAFs exist, not all of them need to be able to provide the same type of analytics results, i.e., some of them can be specialized in providing certain types of analytics. An Analytics ID information element is used to identify the type of supported analytics that NWDAF can generate.

[65] NOTE 2: NWDAF instance(s) can be collocated with a 5GS NF.

[66] The above information is presented as background information only to assist with an understanding of the present disclosure. No determination has been made, and no assertion is made, as to whether any of the above might be applicable as prior art with regard to the present disclosure.

[67] The following description of examples of the present disclosure, with reference to the accompanying drawings, is provided to assist in a comprehensive understanding of the present disclosure, as defined by the claims. The description includes various specific details to assist in that understanding but these are to be regarded as merely exemplary. Accordingly, those of ordinary skill in the art will recognize that various changes and modifications of the examples described herein can be made without departing from the scope of the disclosure.

[68] The same or similar components may be designated by the same or similar reference numerals, although they may be illustrated in different drawings.

[69] Detailed descriptions of techniques, structures, constructions, functions or processes known in the art may be omitted for clarity and conciseness, and to avoid obscuring the subject matter of the present disclosure.

[70] The terms and words used herein are not limited to the bibliographical or standard meanings, but, are merely used to enable a clear and consistent understanding of the disclosure.

[71] Throughout the description and claims of this specification, the words "comprise", "include" and "contain" and variations of the words, for example "comprising" and "comprises", means "including but not limited to", and is not intended to (and does not) exclude other features, elements, components, integers, steps, processes, operations, functions, characteristics, properties and/or groups thereof.

[72] Throughout the description and claims of this specification, the singular form, for example "a", "an" and "the", encompasses the plural unless the context otherwise requires. For example, reference to "an object" includes reference to one or more of

such objects.

[73] Throughout the description and claims of this specification, language in the general form of "X for Y" (where Y is some action, process, operation, function, activity or step and X is some means for carrying out that action, process, operation, function, activity or step) encompasses means X adapted, configured or arranged specifically, but not necessarily exclusively, to do Y.

[74] Features, elements, components, integers, steps, processes, operations, functions, characteristics, properties and/or groups thereof described or disclosed in conjunction with a particular aspect, embodiment, example or claim are to be understood to be applicable to any other aspect, embodiment, example or claim described herein unless incompatible therewith.

[75] Certain examples of the present disclosure provide techniques relating to Artificial Intelligence (AI) and/or Machine Leaning (ML), in particular techniques relating to monitoring application AI/ML operation. For example, certain examples of the present disclosure provide methods, apparatus and systems in 3rd Generation Partnership Project (3GPP) 5th Generation (5G) System (5GS) for monitoring one or more features (e.g., network resource utilisation and/or Quality of Service (QoS)) for application AI and/or ML operation.

[76] However, the skilled person will appreciate that the present disclosure is not limited to these examples, and may be applied in any suitable system or standard, for example one or more existing and/or future generation wireless communication systems or standards, including any existing or future releases of the same standards specification, for example 3GPP 5G.

[77] The following examples are applicable to, and use terminology associated with, 3GPP 5G. However, the skilled person will appreciate that the techniques disclosed herein are not limited to 3GPP 5G. For example, the functionality of the various network entities and other features disclosed herein may be applied to corresponding or equivalent entities or features in other communication systems or standards. Corresponding or equivalent entities or features may be regarded as entities or features that perform the same or similar role, function or purpose within the network.

[78] The skilled person will also appreciate that the transmission of information between network entities is not limited to the specific form, type or order of messages described in relation to the examples disclosed herein.

[79] A particular network entity may be implemented as a network element on a dedicated hardware, as a software instance running on a dedicated hardware, and/or as a virtualized function instantiated on an appropriate platform, e.g., on a cloud infrastructure.

[80] The skilled person will appreciate that the present disclosure is not limited to the

specific examples disclosed herein. For example:

[81] The techniques disclosed herein are not limited to 3GPP 5G.

[82] One or more entities in the examples disclosed herein may be replaced with one or more alternative entities performing equivalent or corresponding functions, processes or operations.

[83] One or more of the messages in the examples disclosed herein may be replaced with one or more alternative messages, signals or other type of information carriers that communicate equivalent or corresponding information.

[84] One or more further entities and/or messages may be added to the examples disclosed herein.

[85] One or more non-essential entities and/or messages may be omitted in certain examples.

[86] The functions, processes or operations of a particular entity in one example may be divided between two or more separate entities in an alternative example.

[87] The functions, processes or operations of two or more separate entities in one example may be performed by a single entity in an alternative example.

[88] Information carried by a particular message in one example may be carried by two or more separate messages in an alternative example.

[89] Information carried by two or more separate messages in one example may be carried by a single message in an alternative example.

[90] The order in which operations are performed and/or the order in which messages are transmitted may be modified, if possible, in alternative examples.

[91] Certain examples of the present disclosure may be provided in the form of an apparatus/device/network entity configured to perform one or more defined network functions and/or a method therefor. Certain examples of the present disclosure may be provided in the form of a system (e.g., network or wireless communication system) comprising one or more such apparatuses/devices/network entities, and/or a method therefor.

[92] In the present disclosure, a UE may refer to one or both of Mobile Termination (MT) and Terminal Equipment (TE). MT may offer common mobile network functions, for example one or more of radio transmission and handover, speech encoding and decoding, error detection and correction, signalling and access to a SIM. An IMEI code, or any other suitable type of identity, may attached to the MT. TE may offer any suitable services to the user via MT functions. However, it may not contain any network functions itself.

[93] AI/ML Application may be part of TE using the services offered by MT in order to support AI/ML operation, whereas AI/ML Application Client may be part of MT. Alternatively, part of AI/ML Application client may be in TE and a part of AI/ML ap-

plication client may be in MT.

[94] The procedures disclosed herein may refer to various network functions/entities. The functions and definitions of certain network functions/entities, for example those indicated below, are known to the skilled person, and are defined, for example, in at least [2] and [3]:

[95] AI/ML Application Function: AI/ML AF

[96] Network Exposure Function: NEF

[97] Unified Data Repository: UDR

[98] Policy Control Function: PCF

[99] Session Management Function: SMF

[100] User Plane Function: UPF

[101] Access and Mobility management Function: AMF

[102] User Equipment: UE

[103] However, as noted above, the skilled person will appreciate that the present disclosure is not limited to the definitions given in [2] and [3], and that equivalent functions/entities may be used.

[104] As noted above, as part of the study item for Rel-18 in SA WG2, an issue was approved with the scope of monitoring the network resource utilization for support of Application AI/ML operations. Another issue was also approved with the scope of Quality of Service (QoS) and policy enhancements for AI/ML-based services. As part of this issue, it was agreed to study QoS monitoring aspects to support the operation of the 3rd party AI/ML operation.

[105] The following requirements related to the above context have been approved for the 5G system:

[106] Based on operator policy, the 5G system shall be able to provide means to allow an authorized third-party to monitor the resource utilisation of the network service that is associated with the third-party.

[107] NOTE 1: Resource utilization in the preceding requirement refers to measurements relevant to the UE's performance such as the data throughput provided to the UE.

[108] Based on operator policy, the 5G system shall be able to provide an indication about a planned change of bitrate, latency, or reliability for a QoS flow to an authorized 3rd party so that the 3rd party AI/ML application is able to adjust the application layer behaviour if time allows. The indication shall provide the anticipated time and location of the change, as well as the target QoS parameters.

[109] Based on operator policy, 5G system shall be able to provide means to predict and expose predicted network condition changes (i.e., bitrate, latency, reliability) per UE, to an authorized third party.

[110] Subject to user consent, operator policy and regulatory constraints, the 5G system

shall be able to support a mechanism to expose monitoring and status information of an AI-ML session to a 3rd party AI/ML application.

[111] NOTE 2: Such mechanism is needed for AI/ML application to determine an in-time transfer of AI/ML model.

[112] Certain examples of the present disclosure provide one or more techniques for monitoring features and capabilities to support AI/ML-based services and operations, for example over the 5G System. In certain examples, application AI/ML operations may be understood as including model splitting, model sharing, federated/distributed learning, etc. However, the skilled person will appreciate that the present disclosure is not limited to these examples.

[113] Certain examples of the present disclosure provide a method for monitoring events in a network, the method comprising: subscribing, by an Application Function (AF), to monitoring the events, wherein the AF supports one or more AIML-based services and/or operations, and the events relate to the performance of the AIML-based services and/or operations; in response to the subscribing, monitoring, by a User Plane Function (UPF), the events; and when an event is detected, reporting, by the UPF to a Network Exposure Function (NEF), the event.

[114] In certain examples, the method may further comprise, reporting, by the NEF to the AF, one or more parameters related to the event.

[115] In certain examples, the event may be reported by the UPF to the NEF directly.

[116] In certain examples, reporting, by the UPF to the NEF, the event may comprise: reporting, by the UPF to a Session Management Function (SMF), the event; and reporting, by the SMF to the NEF, the event.

[117] In certain examples, the monitoring events may comprise one or more of: monitoring session inactivity of a PDU session; monitoring traffic volume of a PDU session; monitoring PCF events; per-5QI monitoring; and monitoring of edge resources.

[118] In certain examples, the monitoring may comprise QoS monitoring and may be performed for multiple UEs.

[119] Certain examples of the present disclosure provide a method for monitoring traffic volume of a PDU session in a network, the method comprising: subscribing, by an Application Function (AF), to monitoring events for traffic volume; in response to the subscribing, monitoring, by a User Plane Function (UPF), traffic volume of the PDU session; when a traffic volume event is detected, reporting, by the UPF to a Network Exposure Function (NEF), a parameter including information on the traffic volume; reporting, by the NEF to the AF, one or more parameters related to the event.

[120] Certain examples of the present disclosure provide a network (or wireless communication system) comprising one or more network entities (e.g., AF, UPF, NEF and/or SMF) configured to operate according to a method of any example, aspect, em-

bodiment and/or claim disclosed herein.

[121] Certain examples of the present disclosure provide a computer program comprising instructions which, when the program is executed by a computer or processor, cause the computer or processor to carry out a method according to any example, aspect, embodiment and/or claim disclosed herein.

[122] Certain examples of the present disclosure provide a computer or processor-readable data carrier having stored thereon a computer program according to any example, aspect, embodiment and/or claim disclosed herein.

[123] Certain examples of the present disclosure may provide one or more of the following monitoring features and capabilities, for example as part of the 5G System:

[124] Support for monitoring of QoS parameters relevant to the performance of AI/ML-based services at the UE and the application AI/ML operation. Non-limiting examples of such parameters may include one or more of packet delay, traffic/data volume, and any other suitable parameter(s).

[125] Support for monitoring of N4 session related parameters relevant to the application AI/ML operations. Non-limiting examples of such parameters include one or more of session inactivity timer, and any other suitable parameter(s).

[126] Support for Application Function (AF) subscription to a series of NWDAF analytics relevant to the application AI/ML operation and performance. Non-limiting examples of such analytics include one or more of DN performance, UE communication, QoS sustainability, for example as defined in 3GPP TS 23.288 (Architecture enhancements for 5G System (5GS) to support network data analytics services), and any other suitable analytics.

[127] Support for new Policy Control Function (PCF) events enabling the PCF reporting of the information described above related to AI/ML-based services (e.g., to the NEF or AF) and/or AI/ML operation.

[128] Support for per-5QI (5G QoS Identifier) monitoring mechanisms for AI/ML-based services and operations. A non-limiting example includes separate monitoring support of traffic 5QI for collected data for training and traffic 5QI for shared models.

[129] Support for monitoring of edge resources related to the AI/ML-based services and operations.

[130] The skilled person will appreciate that the examples of monitoring features and capabilities described above is not exhaustive. Other monitoring features and capabilities directly or indirectly supporting AI/ML-based services are also possible.

[131] Certain examples of the present disclosure may provide specific capabilities and features enabling monitoring of relevant AI/ML-based services and operations according to one or more of the following examples:

[132] - AF monitoring of UL, DL and/or round trip packet delay measurement: the three

application AI/ML operations defined in TS 22.261 [1] (model split, model distribution, FL) may benefit from AF packet delay monitoring as the AI/ML application server may schedule training operation at UEs according to application needs, and delay measurements can assist the server in determining what the best strategy for such scheduling is. For example, delay measurements of the UE members of an FL group may assist the application server to decide which UEs need to provide model updates in each iteration. Delay information may also be critical for model split during joint inference, when sharing of inference results may be critical for application performance. Current specifications already detail how the AF monitors this parameter for URLLC services.

[133] - AF monitoring of traffic/data volume: AF knowledge on the traffic/data volume may help the application server to decide on the AI/ML operations that may be suitable for the application. For example, it may not be convenient to share or distribute very large models frequently, but smaller size models could be shared frequently. Similarly, there are important implications for model splitting depending on the traffic/data volume that needs to be exchanged between UE and application server, and the application server may need to monitor information on traffic/data volume, for example to decide models' optimal splitting points. In current specifications the UPF is already capable of monitoring this parameter and reporting it to the SMF. Certain examples of the present disclosure enable the AF to trigger the monitoring procedure and the SMF to deliver the report to the AF via the NEF.

[134] - AF monitoring of session inactivity time: when a training operation is to be scheduled by an AI/ML application server, either for a single UE or a group of UEs engaged in FL, awareness of session inactivity time as reported by the UPF to the SMF may be very helpful as it aids the server to determine when a UE or a group of UEs is sharing specific types of AI/ML traffic such as trained models. In particular, the dynamicity of FL groups may greatly benefit from this parameter as UE's may dynamically join or leave the group in a pre-scheduled way (therefore optimizing performance) with assistance of this parameter. In current specifications the UPF is already capable of monitoring this parameter and reporting it to the SMF. Certain examples of the present disclosure enable the AF to trigger the monitoring procedure and the SMF to deliver the report to the AF via the NEF.

[135] In some examples, if the application has monitoring capabilities to accurately determine traffic volume and session inactivity, then 5GS monitoring capabilities for traffic/data volume and session inactivity time may not be required.

[136] - AF subscription to NWDAF analytics (e.g., DN performance, UE communication, QoS sustainability), which may assist the AI/ML application operation and is already supported. Current specifications already detail how the AF subscribes to these

analytics.

[137] Figure 1 illustrates an example call flow according to an example of the present disclosure. The procedure in Figure 1 is based on a combination of certain procedures, for example the procedures in clauses 4.15.16.6, 4.16.5, and 4.4.2.2 of TS 23.502 [3]. However, referring to Figure 1, further modifications are described in the following operations.

[138] In operation 1, the AF subscribes to NWDAF analytics relevant to the performance of the UE(s) using the AI/ML application over (a) PDU Session(s) (e.g., DN performance, UE communication, QoS sustainability).

[139] In operation 2, the AF utilizes the Nnef_AFsessionWithQoS service to indicate a subscription to notifications of QoS monitoring for UE traffic related to AI/ML-based services, including packet delay measurement parameter, for example as described in clause 5.33.3 of TS 23.501 [2] for the case of URLLC services. Other requested monitored resources may include usage report and inactivity timer.

[140] Operations 2b and 3 may be performed, for example according to Figure 4.15.6.6-1 in TS 23.502 [3], when the NEF determines to contact the PCF directly. The requested resource monitoring capabilities are forwarded to the PCF.

[141] In certain examples, operation 4 may be performed instead of operations 1-3 when the AF is trusted by the operator to interact directly with the PCF to request monitoring capabilities for an AF session related to AI/ML-based services.

[142] Operations 5 and 6 may be performed, for example according to clause 4.16.5.2 of TS 23.502 [3], for the case when the PCF determines that the SMF needs updated policy information.

[143] Operations 7-11 may be performed, for example, as steps 4, 5 and 6 in Figure 4.15.6.6-1 of TS 23.502 [3] as applied in the case when QoS monitoring is requested for URLLC services without involvement of TSCTSF. In certain examples, operations 9 and 11 may be performed instead of 7, 8 and 10 when the AF is trusted by the operator.

[144] Operations 12-15 may be performed, for example, according to clause 4.4.2.2 of TS 23.502 [3], where the N4 session report may include UL/DL/round trip packet delay measurement, usage report, PDU session inactivity.

[145] Operations 16 and 17 may be performed, for example, according to clause 4.16.5.1 of TS 23.502 [3], providing the event condition(s) that have been met to the PCF.

[146] Operations 18-19 may be performed, for example, as steps 7-8 in Figure 4.15.6.6-1 of TS 23.502 [3] with the event information reported by the PCF. In certain examples, operation 20 may be performed instead when the AF is trusted by the operator.

[147] In operation 21, the AF analyses the monitored information exposed by the 5GS related to the AI/ML traffic.

[148] In operation 22, if needed, the AF may trigger a modification of the PDU Session after having analysed the monitored data.

[149] The procedure illustrated in Figure 1 may also be applied to QoS monitoring in certain examples as described in the following. The procedure in Figure 1 is based on a combination of certain procedures, for example the procedures in clauses 4.15.6.6, 4.16.5 and 4.4.2.2 of 3GPP TS 23.502 [3]. However, referring to Figure 1, further modifications are described in the following operations.

[150] In operation 1, the AF subscribes to NWDAF analytics relevant to the performance of the UE(s) using the AI/ML application (e.g., DN performance, UE communication, QoS sustainability).

[151] In operation 2, the AF uses the Nnef_AFsessionWithQoS service to indicate a subscription to notifications of QoS monitoring for delay measurements of UE traffic related to AI/ML-based services. Hence, the AF request may indicate a packet delay measurement parameter (UL, DL, and/or round trip), for example as described in clause 5.33.3 of TS 23.501 [2] and clause 5.2.6.9 of TS 23.502 [3] for the case of URLLC services.

[152] Operations 2b and 3 may be performed, for example according to Figure 4.15.6.6-1 in TS 23.502 [3], when the NEF determines to contact the PCF directly. The NEF interacts with the PCF by triggering a Npcf_PolicyAuthorization_Create request for session management policy control to authorize the AF request and optionally subscribe to PCF events for measurement of packet delay.

[153] In certain examples, operation 4 may be performed instead of operations 1-3 when the AF is trusted by the operator to interact directly with the PCF to request monitoring capabilities for an AF session related to AI/ML-based services.

[154] In operations 5 and 6, if the PCF determines that the SMF needs updated policy information (for example as is the case for the monitoring parameters in this example), the PCF issues a Npcf_SMPolicyControl_UpdateNotify request with updated policy information, for example as described in the PCF initiated SM Policy Association Modification procedure in clause 4.16.5.2 of TS 23.502 [3]. The SMF then acknowledges the PCF request with a Npcf_SMPolicyControl_UpdateNotify response.

[155] In operations 7-9, the PCF determines whether the request is authorized and notifies the AF if the request is not authorized via the NEF or directly by issuing a Npcf_PolicyAuthorization_Create response message.

[156] In operations 10 and 11, the NEF or AF may send a Npcf_PolicyAuthorization_Subscribe message to the PCF to subscribe to the notification of the PCF events for measurement of packet delay measurement if not done so previously in operation 3.

[157] In operations 12-15, after receiving the event subscription notification from the PCF,

the SMF issues an N4 Session Modification Request, for example as described in clause 4.4.1.3 of TS 23.502 [3], configuring the triggers for event reporting in the UPF. The reporting triggers configured by the SMF may entail session reports for packet delay measurement. The UPF may report on these events, for example according to clause 4.4.2.2 of TS 23.502 [3].

[158] Operations 16-17 may be performed, for example according to clause 4.16.5.1 of TS 23.502 [3], with the SMF providing the event condition(s) that have been met to the PCF.

[159] Operations 18-19 may be performed, for example, as steps 7-8 in Figure 4.15.6.6-1 of TS 23.502 [3] with the event information reported by the PCF to the AF via the NEF in case of the AF is untrusted. In certain examples, operation 20 may be performed instead when the AF is trusted by the operator.

[160] In operation 21, the AF analyses the monitored information exposed by the 5GS related to the AI/ML traffic.

[161] In operation 22, if needed, the AF may trigger a modification of the PDU Session after having analysed the monitored data (for example, see also clause 4.15.6.6a of TS 23.502 [3], without the need for TSCTSF involvement).

[162] Figure 2 illustrates an example call flow according to another example of the present disclosure. In particular, Figure 2 illustrates a procedure for monitoring of session inactivity time and traffic volume.

[163] In operation 1, the AF uses the Nnef_EventExposure_Subscribe service operation to indicate a subscription to new NEF monitoring events for traffic volume and/or session inactivity time.

[164] In operation 2, the NEF subscribes to the user plane status information SMF event, for example described in clause 5.2.8.3.1 of TS 23.502 [3], providing the (AI/ML) application ID and the SUPI(s) of the UE(s) being monitored as event filters, for example as described in Table 5.2.8.3.1-1 of TS 23.502 [3].

[165] In operations 3 and 4, after receiving the event subscription notification from the NEF, the SMF issues an N4 Session Modification Request, for example as described in clause 4.4.1.3 of TS 23.502 [3], configuring the trigger for event reporting in the UPF of PDU session inactivity.

[166] In operation 5, the NEF may subscribe to event exposure from the UPF to obtain information on data usage of a PDU session. For example, the event may be UserDataUsageMeasures. This event provides information of user data usage of the User PDU Session, for example Volume Measurement and Throughput Measurement.

[167] In operation 6, a trigger happens to report events on session inactivity and/or traffic usage by the UPF.

[168] In operation 7, when a PDU session inactivity event is detected, the UPF reports it to

the SMF, for example according to clause 4.4.2.2 of TS 23.502 [3].

[169] In operation 8, the SMF reports the detection of a user plane status information event to the NEF with information on session inactivity time.

[170] In operation 9, the UPF reports to the NEF the detection of the event in operation 5 with information on data usage of a PDU session.

[171] In operation 10, the NEF reports the monitored session inactivity time and/or traffic volume to the AF via Nnef_EventExposure_Notify service operation.

[172] In operation 11, the AF analyses the monitored information exposed by the 5GS related to the AI/ML traffic.

[173] In operation 12, if needed, the AF may trigger a modification of the PDU Session after having analysed the monitored data.

[174] The skilled person will appreciate that the techniques described herein may be used to support both AI/ML-based services (e.g., robotics, computer vision, etc.) and AI/ML model operations that can be done in the network (e.g., AI/ML operation splitting, AI/ML model/data distribution and sharing, and distributed/federated learning). The techniques described herein may improve the performance of such services and operations. The skilled person will also appreciate that certain examples of the present disclosure may also be applied to non-AI/ML related network aspects.

[175] Figure 3 illustrates a block diagram of an exemplary network entity that may be used in examples of the present disclosure, such as the techniques disclosed in relation to Figures 1 and/or 2. For example, the UE, AF, NEF, PCF, SMF, UPF, NWDAF and/or other NFs may be provided in the form of the network entity illustrated in Figure 3. The skilled person will appreciate that a network entity may be implemented, for example, as a network element on a dedicated hardware, as a software instance running on a dedicated hardware, and/or as a virtualised function instantiated on an appropriate platform, e.g., on a cloud infrastructure.

[176] The entity 300 comprises a processor (or controller) 301, a transmitter 303 and a receiver 305. The receiver 305 is configured for receiving one or more messages from one or more other network entities, for example as described above. The transmitter 303 is configured for transmitting one or more messages to one or more other network entities, for example as described above. The processor 301 is configured for performing one or more operations, for example according to the operations as described above.

[177] The techniques described herein may be implemented using any suitably configured apparatus and/or system. Such an apparatus and/or system may be configured to perform a method according to any aspect, embodiment, example or claim disclosed herein. Such an apparatus may comprise one or more elements, for example one or more of receivers, transmitters, transceivers, processors, controllers, modules, units,

and the like, each element configured to perform one or more corresponding processes, operations and/or method steps for implementing the techniques described herein. For example, an operation/function of X may be performed by a module configured to perform X (or an X-module). The one or more elements may be implemented in the form of hardware, software, or any combination of hardware and software.

[178] It will be appreciated that examples of the present disclosure may be implemented in the form of hardware, software or any combination of hardware and software. Any such software may be stored in the form of volatile or non-volatile storage, for example a storage device like a ROM, whether erasable or rewritable or not, or in the form of memory such as, for example, RAM, memory chips, device or integrated circuits or on an optically or magnetically readable medium such as, for example, a CD, DVD, magnetic disk or magnetic tape or the like.

[179] It will be appreciated that the storage devices and storage media are embodiments of machine-readable storage that are suitable for storing a program or programs comprising instructions that, when executed, implement certain examples of the present disclosure. Accordingly, certain examples provide a program comprising code for implementing a method, apparatus or system according to any example, embodiment, aspect and/or claim disclosed herein, and/or a machine-readable storage storing such a program. Still further, such programs may be conveyed electronically via any medium, for example a communication signal carried over a wired or wireless connection.

[180] While the disclosure has been shown and described with reference to certain examples, it will be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the scope of the disclosure, as defined by the appended claims.

[181] Certain examples of the present disclosure provide one or more techniques as disclosed in the accompanying annex to the description. The skilled person will appreciate that any of these techniques may be applied in combination with any of the techniques described above and illustrated in the Figures.

[182] Acronyms and Definitions

[183] 3GPP 3rd Generation Partnership Project

[184] 5G 5th Generation

[185] 5GC 5G Core

[186] 5GS 5G System

[187] 5QI 5G QoS Identifier

[188] ADRF Analytics Data Repository Function

[189] AF Application Function

[190] AI Artificial Intelligence

[191] AMF Access and Mobility management Function

[192] DCCF Data Collection Coordination Function

[193] DL Downlink

[194] DN Data Network

[195] ID Identity/Identifier

[196] IMEI International Mobile Equipment Identities

[197] MFAF Messaging Framework Adaptor Function

[198] ML Machine Learning

[199] MT Mobile Termination

[200] N4 Interface between Control Plane and User Plane

[201] NEF Network Exposure Function

[202] NF Network Function

[203] NRF Network Repository Function

[204] NW Network

[205] NWDAF Network Data Analytics Function

[206] OAM Operations, Administration and Maintenance

[207] PCF Policy Control Function

[208] PDU Protocol Data Unit

[209] PLMN Public Land Mobile Network

[210] QoS Quality of Service

[211] Rel Release

[212] SIM Subscriber Identity Module

[213] SMF Session Management Function

[214] TE Terminal Equipment

[215] TS Technical Specification

[216] TSCTSF Time Sensitive Communication Time Synchronisation Function

[217] UDM Unified Data Manager

[218] UDR Unified Data Repository

[219] UE User Equipment

[220] UL Uplink

[221] UPF User Plane Function

[222] URLLC Ultra-Reliable Low-Latency Communication

[223] WG Working Group

[224] Although the present disclosure has been described with various embodiments, various changes and modifications may be suggested to one skilled in the art. It is intended that the present disclosure encompass such changes and modifications as fall within the scope of the appended claims.

Claims

[Claim 1] A method performed by a network exposure function (NEF) for monitoring an event in a network, the method comprising:
receiving, from an application function (AF) which supports one or more artificial intelligence/machine leaning (AI/ML)-based services, first information indicating a subscription for monitoring an event related to the AI/ML-based services;
transmitting, to a session management function (SMF), second information for obtaining information on the event related to an inactivity time;
transmitting, to a user plane function (UPF), third information for obtaining information on the event related to a traffic volume;
receiving at least one of the information on the event related to the inactivity time or the information to the event related to the traffic volume; and
transmitting, to the AF, one or more parameter obtained based on the at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume

[Claim 2] The method of claim 1,
wherein the information on the event related to the traffic volume is reported by the UPF to the NEF directly,
wherein the information on the event related to the inactivity time is reported by the UPF to the NEF through the SMF, and
wherein the event is associated with one or more of a session inactivity of a protocol data unit (PDU) session, a traffic volume of a PDU session, policy control function (PCF) events, per-5G quality of service (QoS) identifier (5QI) monitoring, and edge resources.

[Claim 3] A method performed by a user plane function (UPF) for monitoring an event in a network, the method comprising:
receiving, from a session management function (SMF), first information for obtaining information on the event related to an inactivity time;
receiving, from a network exposure function (NEF), second information for obtaining information on the event related to a traffic volume;
monitoring at least one event based on the first information and the second information; and

reporting at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume, in case that a trigger happens to report the at least one event.

[Claim 4] The method of claim 3,
wherein the information on the event related to the traffic volume is reported by the UPF to the NEF directly,
wherein the information on the event related to the inactivity time is reported by the UPF to the NEF through the SMF, and
wherein the event is associated with one or more of a session inactivity of a protocol data unit (PDU) session, a traffic volume of a PDU session, policy control function (PCF) events, per-5G quality of service (QoS) identifier (5QI) monitoring, and edge resources.

[Claim 5] A method performed by an application function (AF) for monitoring an event in a network, the method comprising:
transmitting, to a network exposure function (NEF), first information indicating a subscription for monitoring an event related to artificial intelligence/machine leaning (AI/ML);
receiving, from the NEF, one or more parameter obtained based on at least one of information on the event related to an inactivity time or information on the event related to a traffic volume; and
analyzing the one or more parameter for AI/ML traffic,
wherein the AF supports one or more AI/ML-based services.

[Claim 6] The method of claim 5,
wherein second information for obtaining information on the event related to the inactivity time is transmitted to a user plane function (UPF) through the NEF and a session management function (SMF),
wherein the information on the event related to the inactivity time is reported by the UPF to the NEF through the SMF,
wherein third information for obtaining information on the event related to the traffic volume is transmitted to the UPF through the NEF, and
wherein the information on the event related to the traffic volume is reported by the UPF to the NEF directly.

[Claim 7] The method of claim 5, further comprising:
triggering a modification of a protocol data unit (PDU) session based on analysis of the one or more parameter,
wherein the event is associated with one or more of a session inactivity of a protocol data unit (PDU) session, a traffic volume of a PDU

session, policy control function (PCF) events, per-5G quality of service (QoS) identifier (5QI) monitoring, and edge resources.

[Claim 8] A network exposure function (NEF) for monitoring an event in a network, the NEF comprising:

a transceiver; and

a controller configured to:

receive, from an application function (AF) which supports one or more artificial intelligence/machine leaning (AI/ML)-based services, first information indicating a subscription for monitoring an event related to the AI/ML-based services,

transmit, to a session management function (SMF), second information for obtaining information on the event related to an inactivity time,

transmit, to a user plane function (UPF), third information for obtaining information on the event related to a traffic volume,

receive at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume, and

transmit, to the AF, one or more parameter obtained based on the at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume.

[Claim 9] The NEF of claim 8,

wherein the information on the event related to the traffic volume is reported by the UPF to the NEF directly,

wherein the information on the event related to the inactivity time is reported by the UPF to the NEF through the SMF, and

wherein the event is associated with one or more of a session inactivity of a protocol data unit (PDU) session, a traffic volume of a PDU session, policy control function (PCF) events, per-5G quality of service (QoS) identifier (5QI) monitoring, and edge resources.

[Claim 10] A user plane function (UPF) for monitoring an event in a network, the UPF comprising:

a transceiver; and

a controller configured to:

receive, from a session management function (SMF), first information for obtaining information on the event related to an inactivity time,

receive, from a network exposure function (NEF), second information for obtaining information on the event related to a traffic volume,

monitor at least one event based on the first information and the second

information, and

report at least one of the information on the event related to the inactivity time or the information on the event related to the traffic volume, in case that a trigger happens to report the at least one event.

[Claim 11] The UPF of claim 10,

wherein the information on the event related to the traffic volume is reported by the UPF to the NEF directly,

wherein the information on the event related to the inactivity time is reported by the UPF to the NEF through the SMF, and

wherein the event is associated with one or more of a session inactivity of a protocol data unit (PDU) session, a traffic volume of a PDU session, policy control function (PCF) events, per-5G quality of service (QoS) identifier (5QI) monitoring, and edge resources.

[Claim 12] An application function (AF) for monitoring an event in a network, the AF comprising:

a transceiver; and

a controller configured to:

transmit, to a network exposure function (NEF), first information indicating a subscription for monitoring an event related to artificial intelligence/machine leaning (AI/ML),

receive, from the NEF, one or more parameter obtained based on at least one of information on the event related to an inactivity time or information on the event related to a traffic volume, and

analyze the one or more parameter for AI/ML traffic,

wherein the AF supports one or more AI/ML-based services.

[Claim 13] The AF of claim 12,

wherein second information for obtaining information on the event related to the inactivity time is transmitted to a user plane function (UPF) through the NEF and a session management function (SMF),

wherein the information on the event related to the inactivity time is reported by the UPF to the NEF through the SMF,

wherein third information for obtaining information on the event related to the traffic volume is transmitted to the UPF through the NEF, and

wherein the information on the event related to the traffic volume is reported by the UPF to the NEF directly.

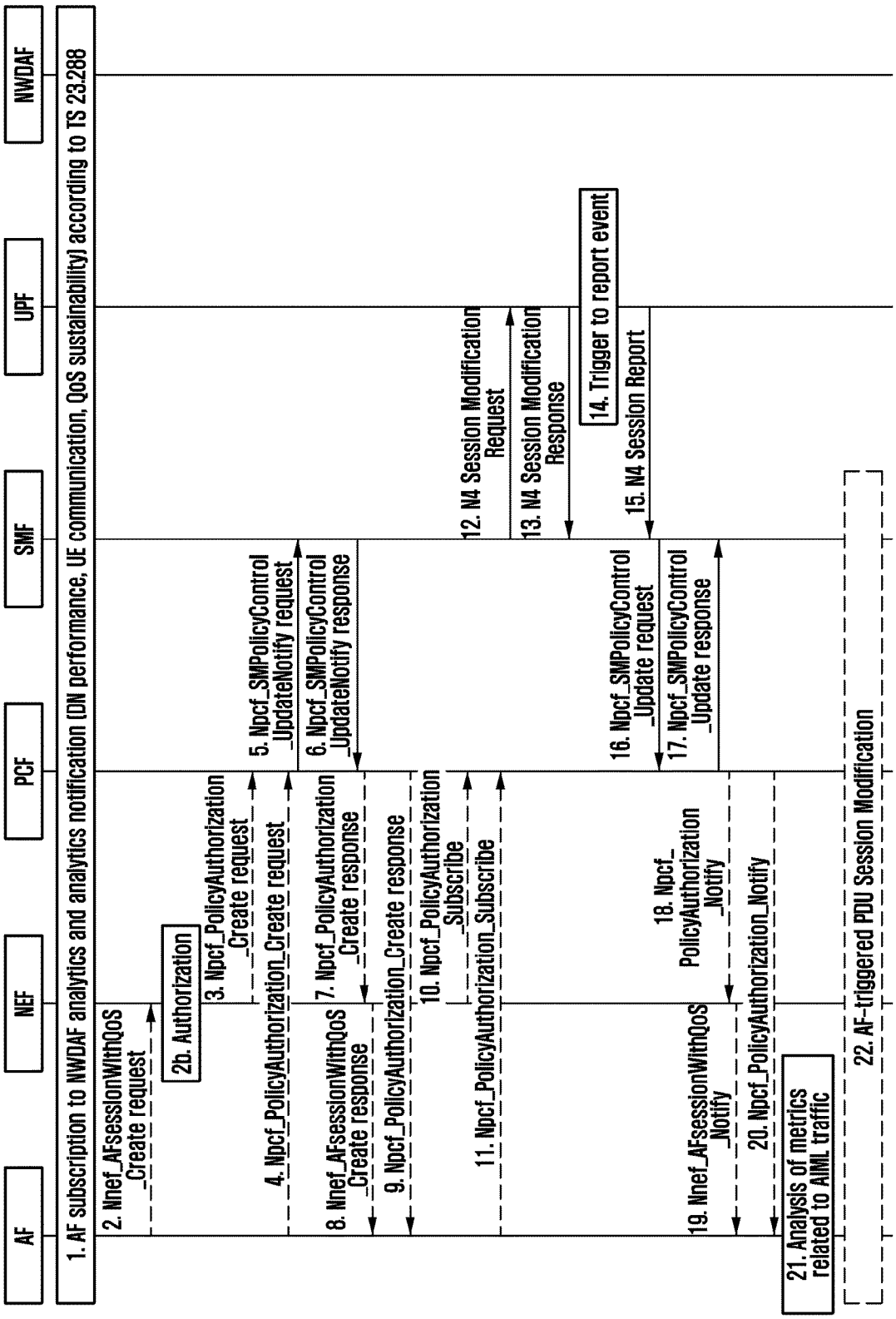
[Claim 14] The AF of claim 12,

wherein the event is associated with one or more of:

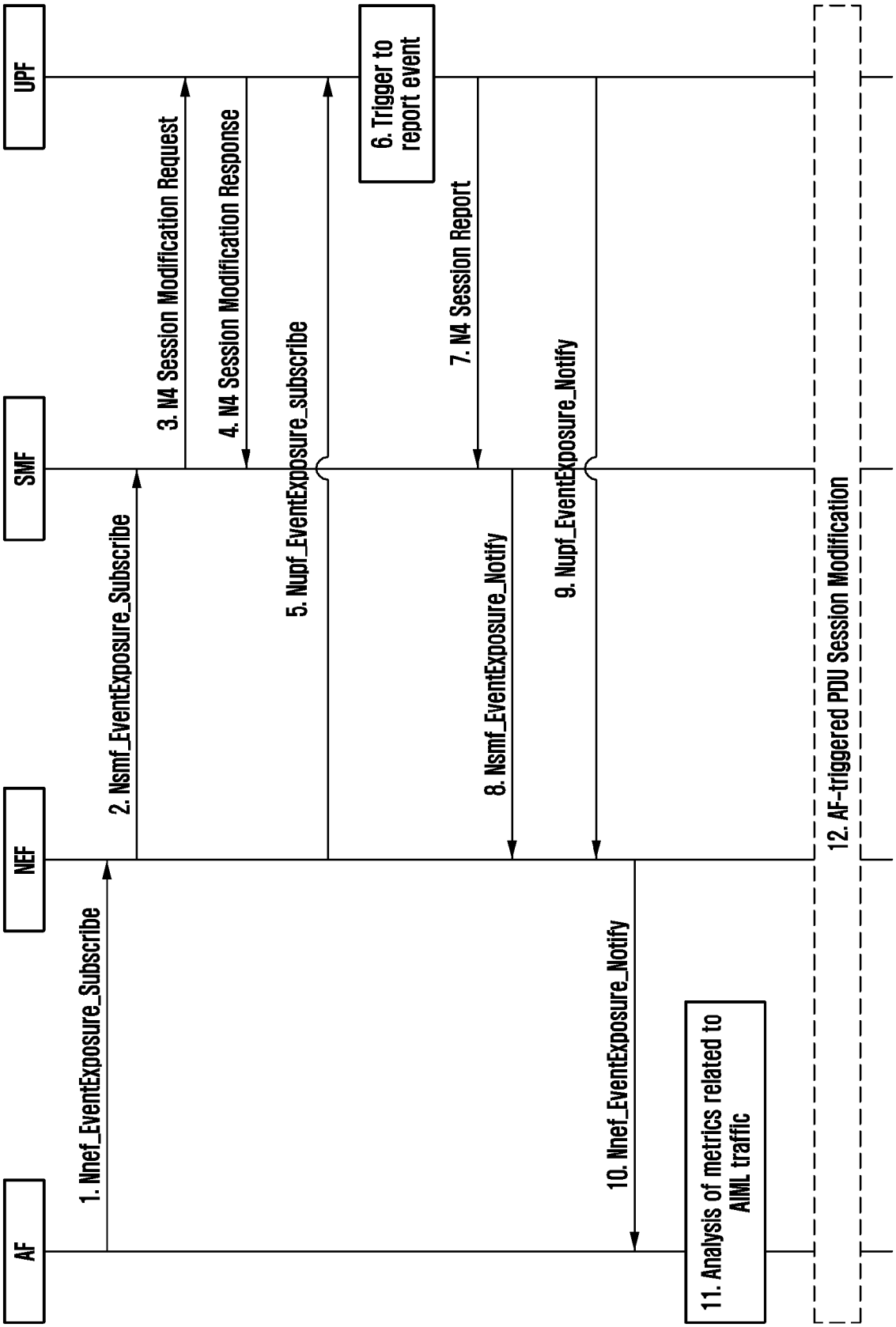
a session inactivity of a protocol data unit (PDU) session, a traffic volume of a PDU session, policy control function (PCF) events, per-5G quality of service (QoS) identifier (5QI) monitoring, and edge resources.

[Claim 15] The AF of claim 12,
wherein the controller is configured to trigger a modification of a protocol data unit (PDU) session based on analysis of the one or more parameter.

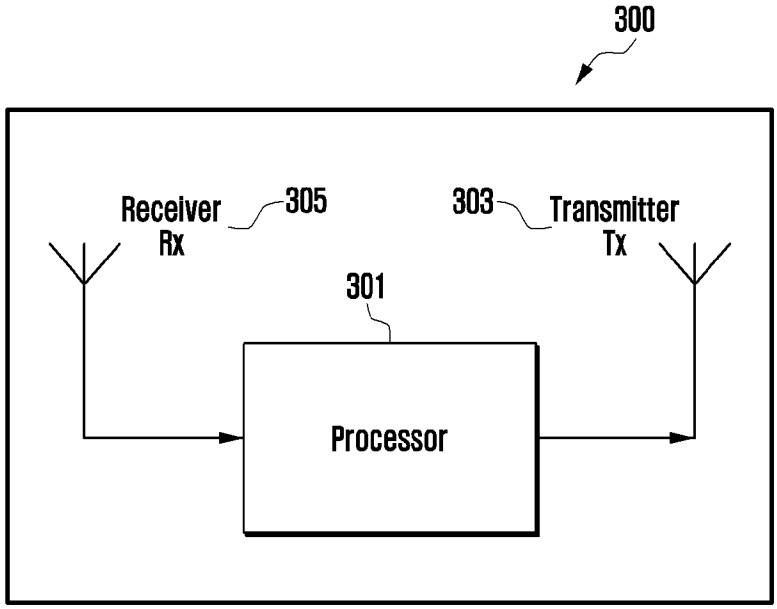
[Fig. 1]


AF
NEF
PCF
SMF
UPF
NWDAF
1. AF subscription to NWDAF analytics and analytics notification (DN performance, UE communication, QoS sustainability) according to TS 23.288
2. Nnef_AFsessionWithQoS _Create request
2b. Authorization
3. Npcf_PolicyAuthorization _Create request
4. Npcf_PolicyAuthorization_Create request
5. Npcf_SMPolicyControl _UpdateNotify request
6. Npcf_SMPolicyControl _UpdateNotify response
7. Npcf_PolicyAuthorization _Create response
8. Nnef_AFsessionWithQoS _Create response
9. Npcf_PolicyAuthorization_Create response
10. Npcf_PolicyAuthorization _Subscribe
11. Npcf_PolicyAuthorization_Subscribe
12. N4 Session Modification Request
13. N4 Session Modification Response
14. Trigger to report event
15. N4 Session Report
16. Npcf_SMPolicyControl _Update request
17. Npcf_SMPolicyControl _Update response
18. Npcf_ PolicyAuthorization _Notify
19. Nnef_AFsessionWithQoS _Notify
20. Npcf_PolicyAuthorization_Notify
21. Analysis of metrics related to AIML traffic
22. AF-triggered PDU Session Modification

[Fig. 2]


UPF
SMF
NEF
AF
1. Nnef_EventExposure_Subscribe
2. Nsmf_EventExposure_Subscribe
3. N4 Session Modification Request
4. N4 Session Modification Response
5. Nupf_EventExposure_subscribe
6. Trigger to report event
7. N4 Session Report
8. Nsmf_EventExposure_Notify
9. Nupf_EventExposure_Notify
10. Nnef_EventExposure_Notify
11. Analysis of metrics related to AIML traffic
12. AF-triggered PDU Session Modification

[Fig. 3]



INTERNATIONAL SEARCH REPORT

International application No.
PCT/KR2023/004195

A. CLASSIFICATION OF SUBJECT MATTER

H04L 43/02(2022.01)i; **H04L 41/16**(2022.01)i; **H04L 43/062**(2022.01)i; **H04L 41/5003**(2022.01)i; **H04L 67/14**(2022.01)i; **H04W 88/14**(2009.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

H04L 43/02(2022.01); H04W 72/02(2009.01); H04W 72/04(2009.01)

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

Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

eKOMPASS(KIPO internal) & Keywords: NEF, AF, AI/ML, SMF, UPF, event, traffic volume, inactivity time

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	`3GPP; TSGSA; Procedures for the 5G System (5GS); Stage 2 (Release 17)`, 3GPP TS 23.502 V17.4.0, 23 March 2022 pages 1-704; table 5.2.8.1-1 and figures 4.15.3.2.3-1, 4.15.6.6-1	3-4,10-11
Y		1-2,5-9,12-15
Y	OPPO, `5GS Assisted AIML Services and Transmissions (FS_5GAIML)`, S2-2103759, 3GPP TSG-SA WG2 Meeting #145E, 10 May 2021 slides 1-16	1-2,5-9,12-15
A	`3GPP; TSGSA; Architecture enhancements for 5G System (5GS) to support network data analytics services (Release 17)`, 3GPP TS 23.288 V17.4.0, 23 March 2022 pages 1-198	1-15
A	`3GPP; TSGCT; 5G System; Session Management Policy Control Service; Stage 3 (Release 17)`, 3GPP TS 29.512 V17.6.0, 21 March 2022 pages 1-243	1-15
A	WO 2018-008980 A1 (LG ELECTRONICS INC.) 11 January 2018 (2018-01-11) paragraphs [0144]-[0221] and figures 6-7	1-15

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

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

Date of the actual completion of the international search	Date of mailing of the international search report
20 June 2023	**21 June 2023**

Name and mailing address of the ISA/KR	Authorized officer
Korean Intellectual Property Office **189 Cheongsa-ro, Seo-gu, Daejeon** **35208, Republic of Korea**	**BYUN, SUNG CHEAL**
Facsimile No. **+82-42-481-8578**	Telephone No. **+82-42-481-8262**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/KR2023/004195

Patent document cited in search report	Publication date (day/month/year)	Patent family member(s)	Publication date (day/month/year)
WO 2018-008980 A1	11 January 2018	US 10856265 B2	01 December 2020
		US 2019-0364541 A1	28 November 2019