SECURITY AND PRIVACY IMPLICATIONS OF 3GPP AI/ML NETWORKING STUDIES FOR 6G

Behcet Sarikaya,
Roland Schott
IETF 119 Side Meeting
March 2024

NOTE WELL

This is a reminder of IETF policies in effect on various topics such as patents or code of conduct. It is only meant to point you in the right direction. Exceptions may apply. The IETF's patent policy and the definition of an IETF "contribution" and "participation" are set forth in BCP 79; please read it carefully.

As a reminder:

- By participating in the IETF, you agree to follow IETF processes and policies.
- If you are aware that any IETF contribution is covered by patents or patent applications that are owned or controlled by you or your sponsor, you must disclose that fact, or not participate in the discussion.
- As a participant in or attendee to any IETF activity you acknowledge that written, audio, video, and photographic records of meetings may be made public.
- Personal information that you provide to IETF will be handled in accordance with the IETF Privacy Statement.
- As a participant or attendee, you agree to work respectfully with other participants; please contact the ombudsteam (https://www.ietf.org/contact/ombudsteam/) if you have questions or concerns about this.

Definitive information is in the documents listed below and other IETF BCPs. For advice, please talk to WG chairs or ADs:

- BCP 9 (Internet Standards Process)
- BCP 25 (Working Group processes)
- BCP 25 (Anti-Harassment Procedures)
- BCP 54 (Code of Conduct)
- BCP 78 (Copyright)
- BCP 79 (Patents, Participation)
- https://www.ietf.org/privacy-policy/(Privacy Policy)

OUTLINE

- Introduction to AI, ML, NN
- Application Areas of AI/ML Networking in Mobile Network
- Architecture
- Security and Privacy Issues

Note: Some slides are adapted from Tricci So presentation at IETF 118 in 6GIP Side Meeting entitled **AI/ML Standardization Status in 3GPP R18**

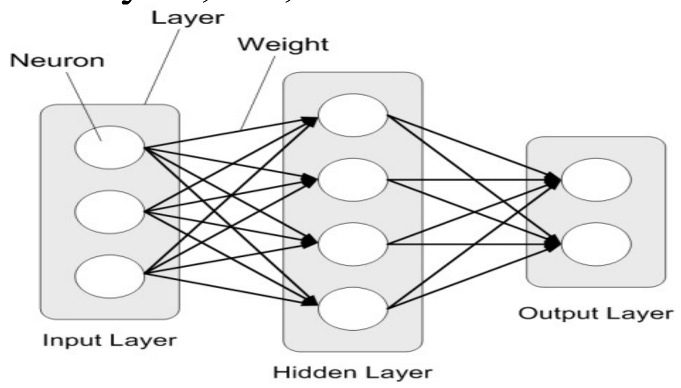
DEFINITIONS

- ***Artificial Intelligence (AI)*** is the science and engineering to build intelligent machines capable of carrying out tasks as humans do

- Within AI is a large subfield called ***machine learning (ML)***, which was defined as the field of study that gives computers the ability to learn without being explicitly programmed

- Within the ML field, there is an area that is often referred to as brain-inspired computation, which is a program aiming to emulate some aspects of how we understand the brain to operate the more popular ML approaches are using ***"neural network"*** as the model. Neural networks (NN) take their inspiration from the notion that a neuron's computation involves a weighted sum of the input values.

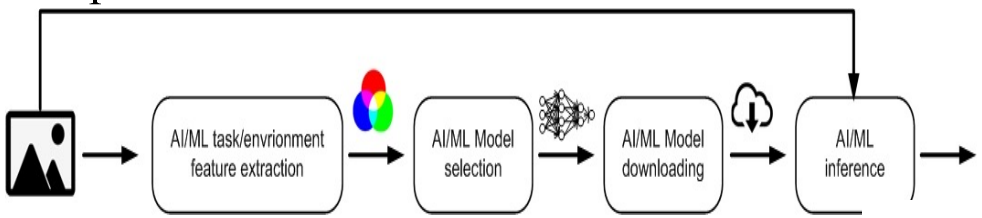
- Neural networks having more than three layers, i.e., more than one hidden layer are called ***deep neural networks (DNN)***.



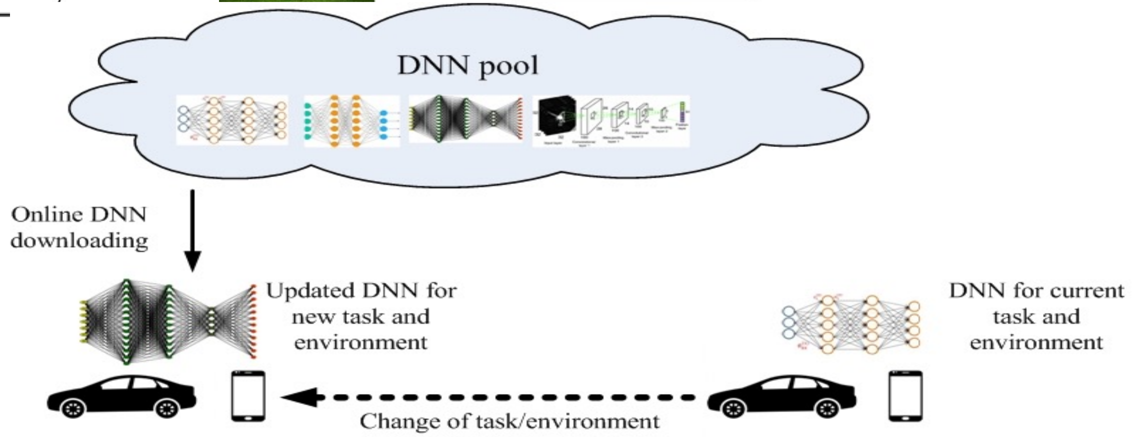
DEFINITIONS

- ***Training*** is a process in which a AI/ML model learns to perform its given tasks, more specifically, by optimizing the value of the weights in the DNN. A DNN is trained by inputting a training set, which are often correctly-labelled training samples. The gradient indicates how the weights should change in order to reduce the loss (the gap between the correct outputs and the outputs computed by the DNN based on its current weights). The training process is repeated iteratively to continuously reduce the overall loss. Until the loss is below a predefined threshold, the DNN with high precision is obtained.

- After a DNN is trained, it can perform its task by computing the output of the network using the weights determined during the training process, which is referred to as ***inference***. In the model inference process, the inputs from the real world are passed through the DNN. Then the prediction for the task is output.



- Model Selection and Downloading an AI/ML model can be distributed from a NW endpoint to the devices when they need it to adapt to the changed AI/ML tasks and environments



WIDELY-USED DNN MODELS AND ALGORITHMS

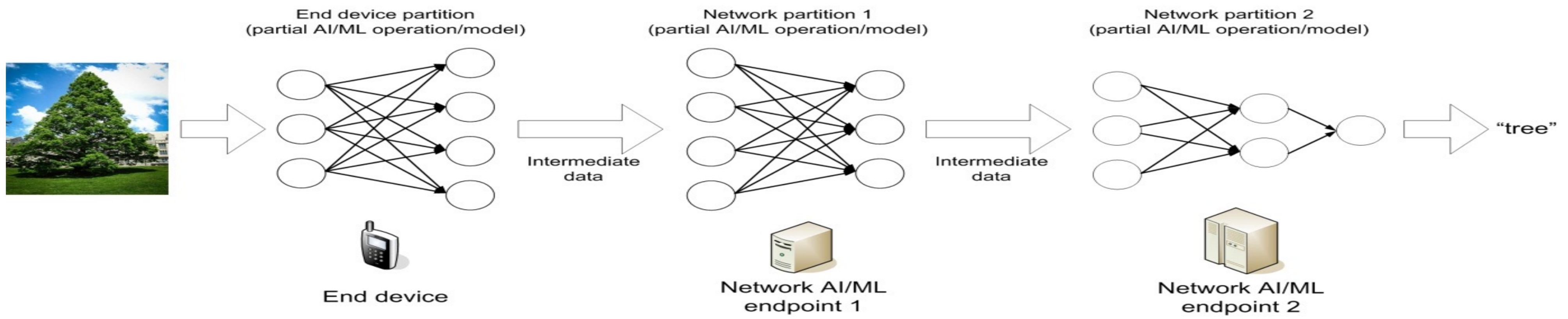
three popular structures of DNNs: multilayer perceptrons (MLPs), convolution neural networks (CNNs), and recurrent neural networks (RNNs).

Multilayer perceptrons (MLP) model is the most basic DNN, which is composed of a series of fully connected layers, hence MLP requires a significant amount of storage and computation

An extremely popular window-based DNN model uses a convolution operation to structure the computation, hence is named as ***convolution neural network (CNN)***. A CNN is composed of multiple convolutional layers. Applying various convolutional filters, CNN models can capture the high-level representation of the input data, making it popular for image classification and speech recognition tasks

Recurrent neural network (RNN) models are another type of DNNs, which use sequential data feeding. The input of RNN consists of the current input and the previous samples, **LLM** (Large Language Models, also called generative AI) are based on RNNs

Split AI/ML inference in many cases, the split AI/ML inference over device and network are required, to enable the AI/ML applications with conflicting requirements which are computation-intensive, energy-intensive as well as privacy-sensitive and delay- sensitive

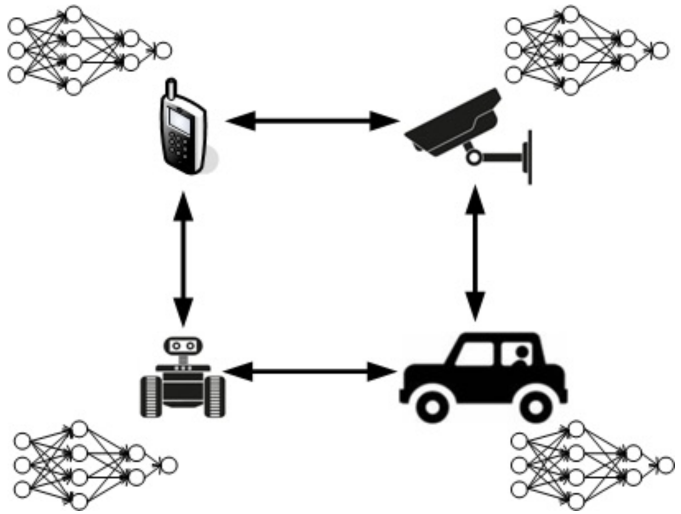


Application area: image recognition

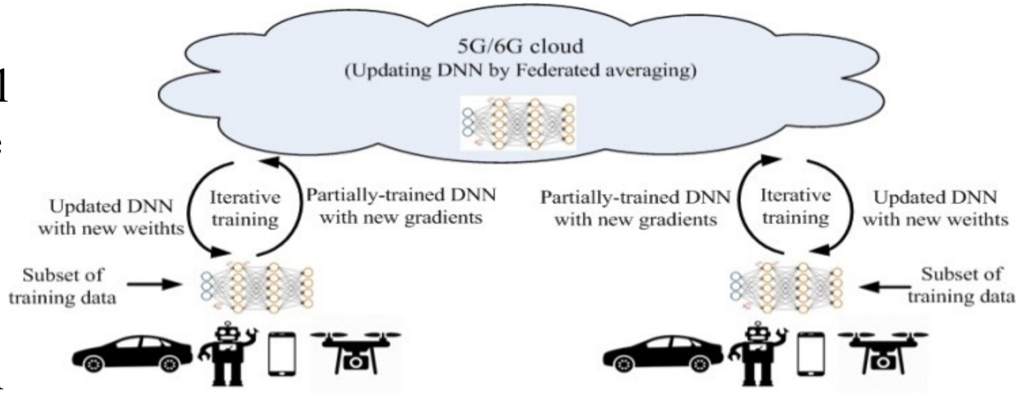
TRAINING

- ➢Similar to the split AI/ML inference, AI/ML model training tasks can also work in a cloud-device coordination manner.

➢Distributed Learning: each computing node trains its own DNN model locally with local data, which preserves private information locally. To obtain the global DNN model by sharing local training improvement, nodes in the network will communicate with each other to exchange the local model updates. In this mode, the global DNN model can be trained without the intervention of the cloud datacenter



Federated Learning In Federated Learning (FL) mode, the cloud server trains a global model by aggregating local models partially-trained by each end devices. The most agreeable Federated Learning algorithm so far is based on the iterative model averaging. Within each training iteration, a UE performs the training based on the model downloaded from the AI server using the local training data. Then the UE reports the interim training results (e.g., gradients for the DNN) to the cloud server via UL channels. The server aggregates the gradients from the UEs, and updates the global model. Next, the updated global model is distributed to the UEs via DL channels. Then the UEs can perform the training for the next iteration



AI APPLICATIONS

Network analytics Based on analytics of UE's location, mobility, download data size and etc, the mobile network could predict that lots UE will download certain amount of data from an AI/ML model server in some location area and inform the AI/ML model that certain UE will probably download certain amount of data. The AI/ML model server could use such information to adjust its prediction

Measured data rate/delay and other traffic analytics information prediction and exposure

E2E data volume transfer time analytics may be used to assist an AF or NEF with AI/ML-based services, e.g. for member UE selection of federated learning

Consumer AF/Application Server determines to adjust service parameters, e.g. service parameters of video for adjustment of may be bit rate, frame rate, codec format, compression parameter, screen size, etc. or service parameters

Network analytics can be done using CNN models

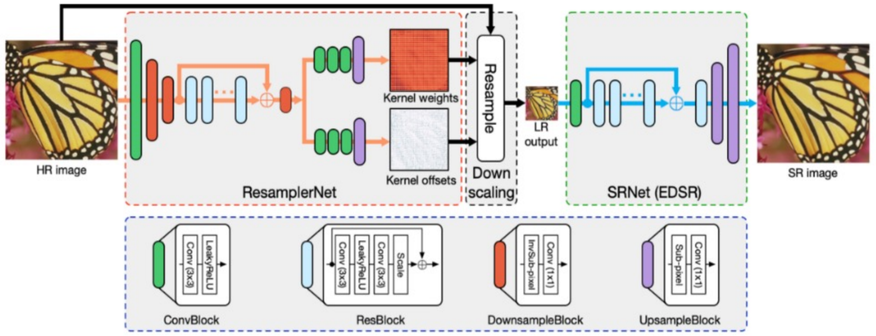
APPLICATION AREAS OF AI/ML NETWORKING IN MOBILE NETWORK

split AI/ML **image recognition** The split AI/ML image recognition algorithms can be analyzed based on the computation and data characteristics of the layers in the CNN

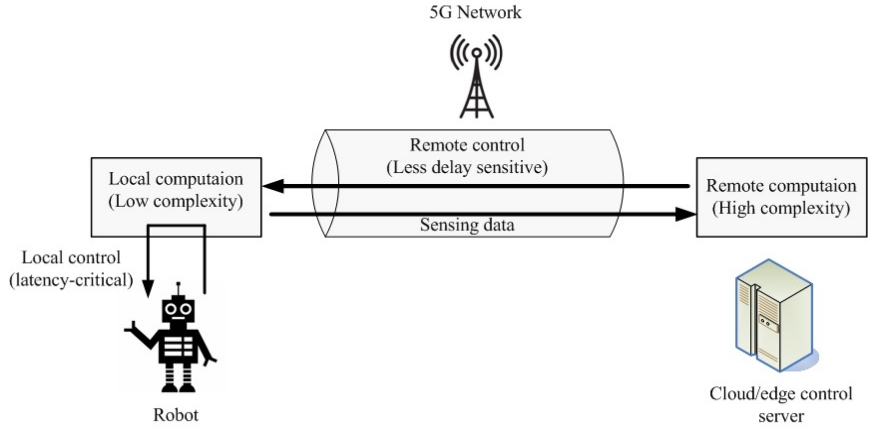
Enhanced media recognition: Deep Learning Based Vision Applications

Media quality enhancement: Video streaming upgrade

Example DNN-based Down/Up-scaler



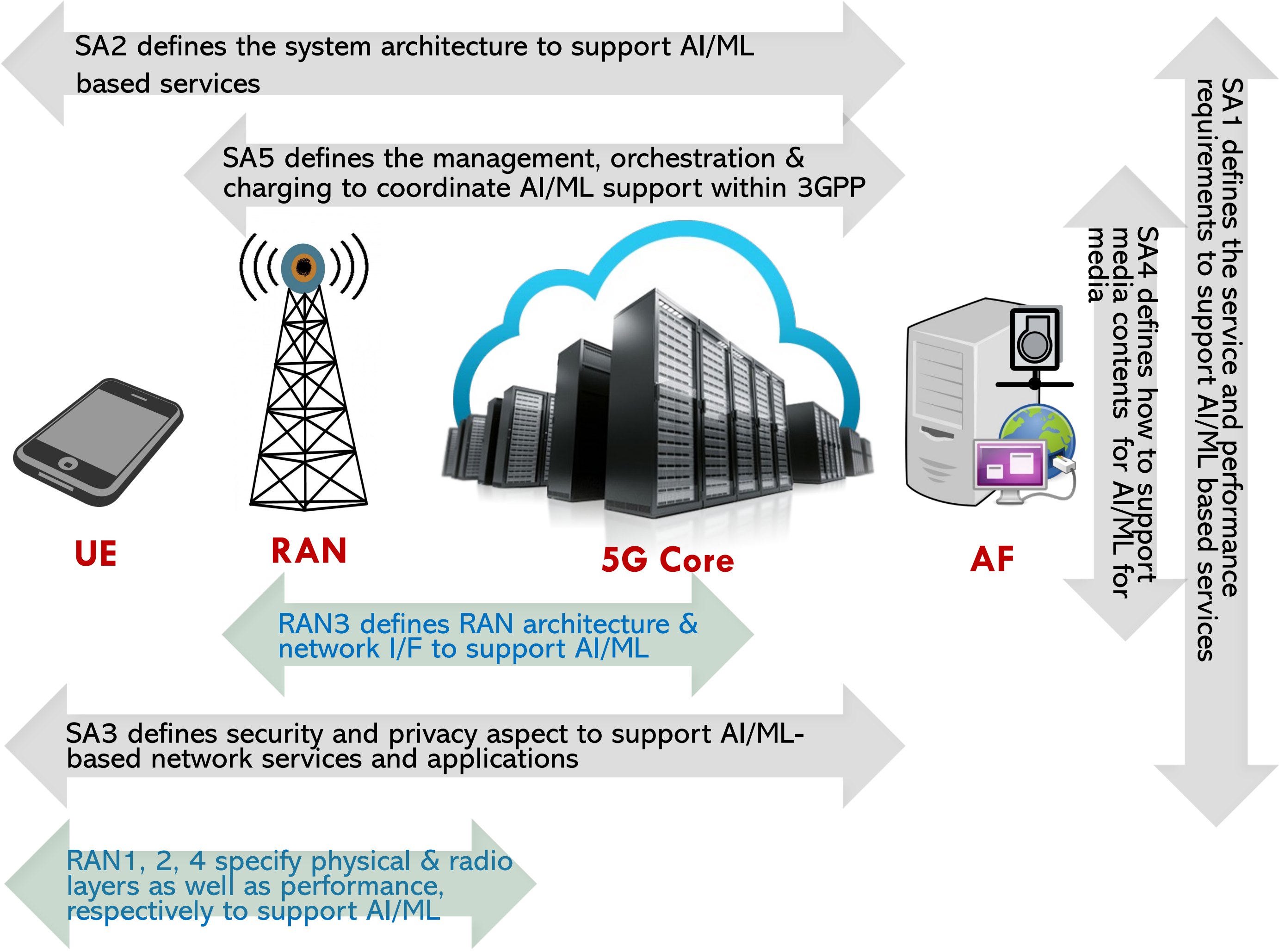
Split control of legged robot over mobile network



What 3GPP Working Groups Do on AI/ML

- ✓ **SA WG-1 (SA1):** Responsible for identifying service and performance requirements for 3GPP systems, in Rel-18, SA1 focused on defining the AI/ML model transfer in 5G.
- ✓ **SA WG-2 (SA2):** Responsible for developing system architecture, in Rel-18, SA2 worked on 5G system support for intelligent transport for the AI/ML-based services.
- ✓ **SA WG-3 (SA3):** Responsible for security and privacy aspects. For AI/ML, SA3 examined and determined the system security and privacy impacts towards 5G Core when supporting AI/ML-based network services and applications.
- ✓ **SA WG-4 (SA4):** Responsible for defining media codec for the system and delivery aspects of the media contents, inRel-18, SA4 defined the AI/ML for media.
- ✓ **SA WG-5 (SA5):** Responsible for management, orchestration, and charging for 3GPP systems, in Rel-18, SA5 defined AI/ML management to coordinate AI/ML functions across 5G system.
- ✓ **RAN WG-3 (RAN3):** Responsible for the overall RAN architecture and the specification of protocols for the related network interfaces, in Rel-17 and 18, RAN3 defined the initial support for AI/ML for next-generation RAN (NG-RAN).
- ✓ **RAN WG-1, 2, and 4 (RAN1, RAN2, and RNA4):** Responsible for physical layer, radio layer and performance of the radio Interfaces for UE, Evolved UTRAN, NG-RAN, and beyond, respectively, in Rel-18, these WGs define AI/ML for new radio (NR) air interface which is led by RAN1.

SA6 responsible for application layer services for "vertical markets": automotive, drones, smart factories



ARCHITECTURE

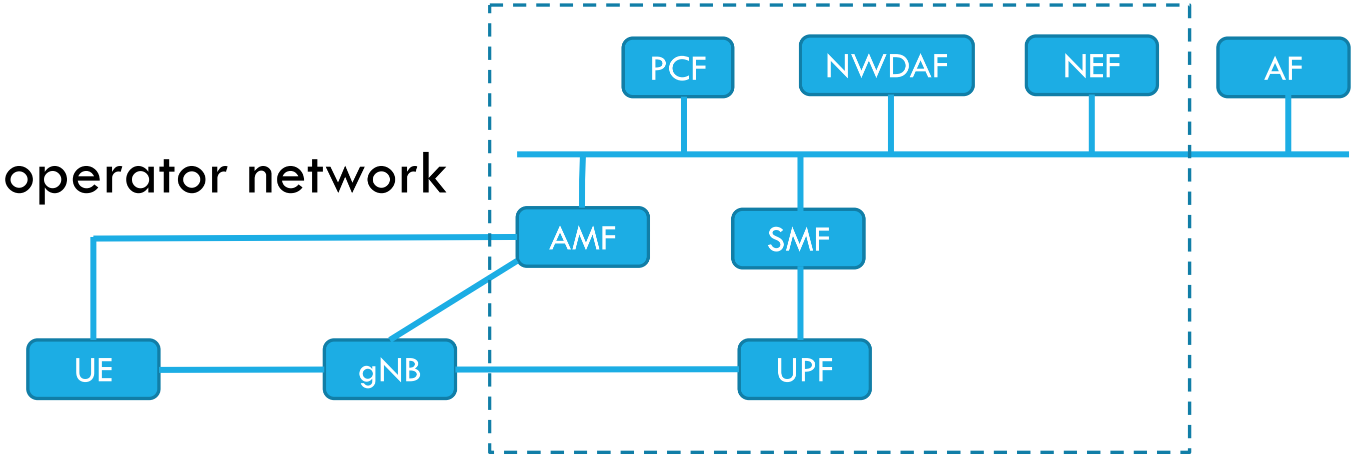
Network Data Analytics Function (NWDAF) provides analytics to Mobile Core Network Functions (NFs) and Operations and Management

The network exposure function (NEF) in Mobile Core to support monitoring and configuration capability for detection and/or reporting of monitoring events to authorized external party

Application Function AF

Applications outside of operator network

Or in the network



SA2 architecture enhancement for network AI/ML operation (eNA)

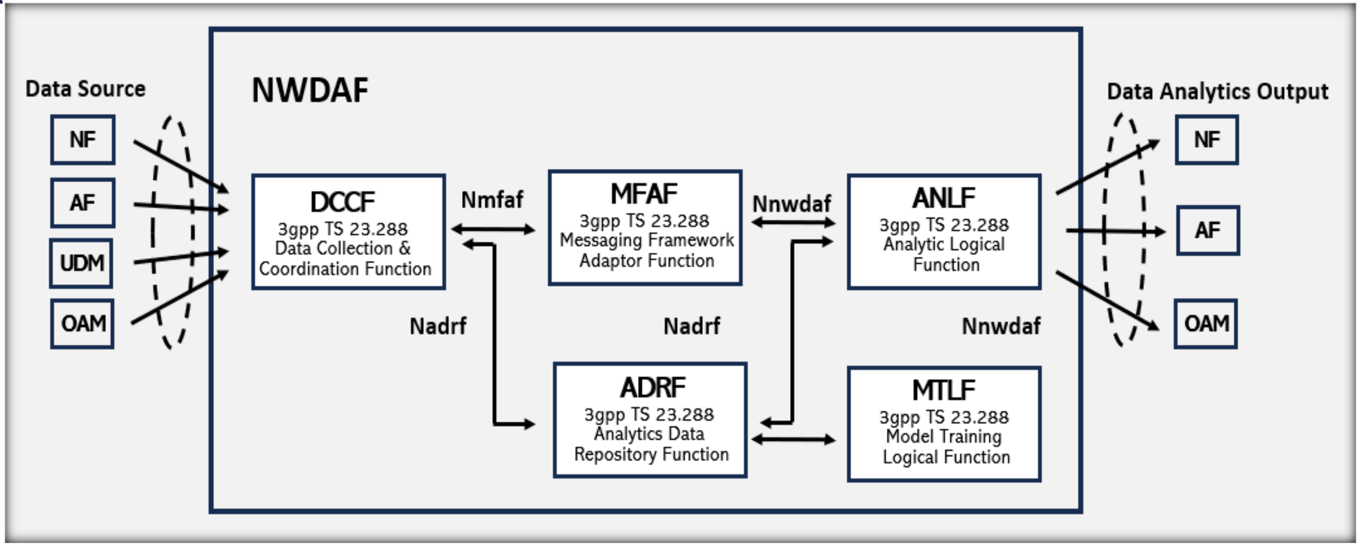
What Is Network Data Analytics Function (NWDAF)?

NWDAF as defined in 3GPP TSs 23.288 & 29.520 incorporates standard interfaces from the service-based architecture to collect data by subscription or request model from other network functions.

NWDAF defined in 3GPP TS 29.520 incorporates standard interfaces from the **service-based architecture** to collect data by subscription or request model from other NFs and similar procedures. This is to deliver analytics functions in the network for automation or reporting, solving major custom interface or format challenges.

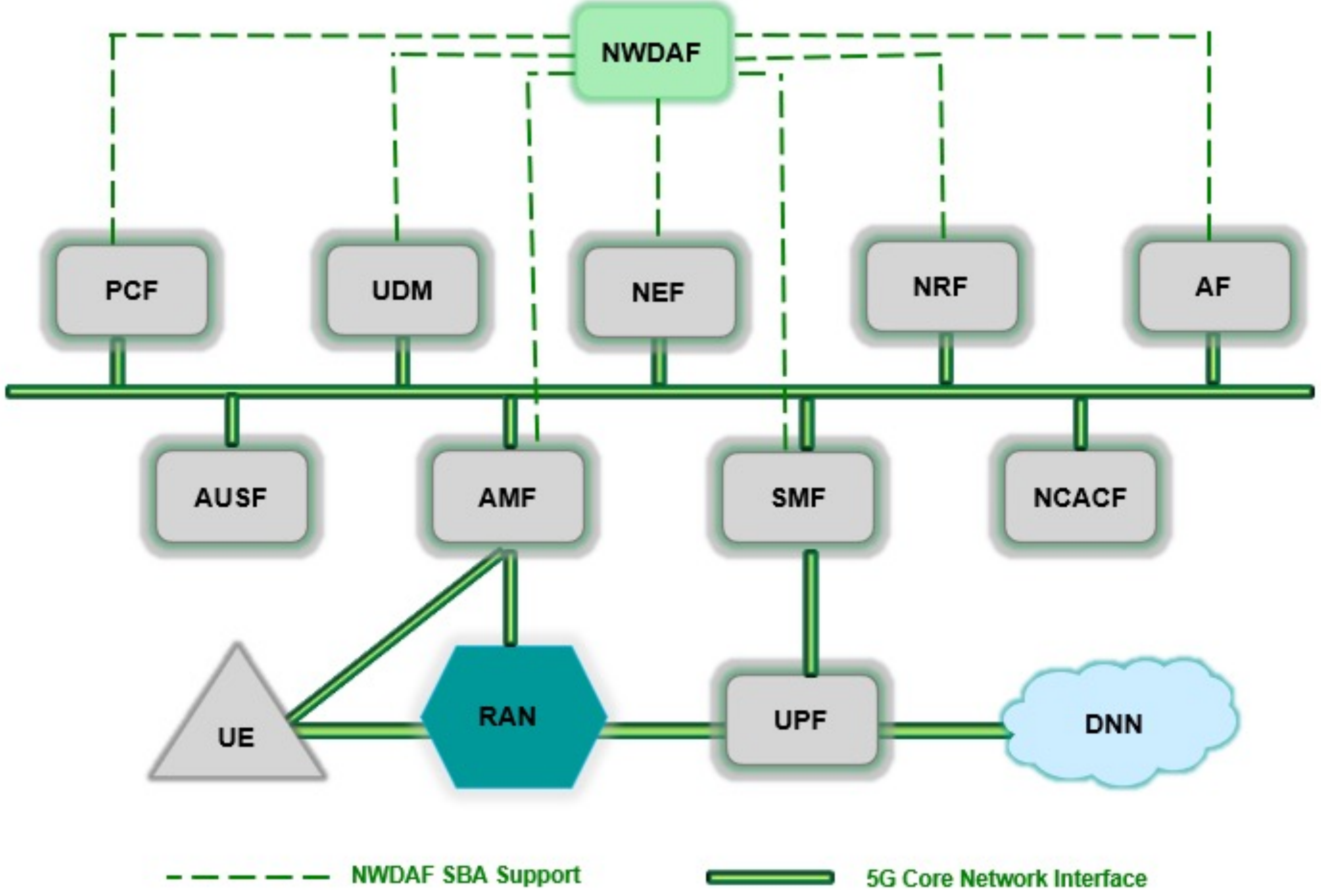
Group of standard functions that defined by 3GPP for supporting data analytics to support 5G Network Operation:

- ❑ NWDAF-ANLF – Analytical Logical Function
- ❑ NWDAF-MTLF – Model Training Logical Function
- ❑ DCCF – Data Collection Coordination (& Delivery) Function
- ❑ ADRF – Analytical Data Repository Function
- ❑ MFAF – Messaging Framework Adaptor Function



SA2 architecture enhancement for network AI/ML operation (eNA)

3GPP Mobile Core Service Based Architecture w.r.t NWDAF



What are the Key Functionalities of NWDAF?

- ✓ Support data collection from NFs and AFs.
- ✓ Support data collection from OAM.
- ✓ NWDAF service registration and metadata exposure to NFs and AFs.
- ✓ Support analytics information provisioning to NFs and AFs.
- ✓ Support Machine Learning (ML) model training and service provisioning to NWDAF-MTLF & NWDAF-AnLF

Analytical Logical Function
Model Training Logical Function

SA2 architecture enhancement for network AI/ML operation (eNA)

Referring to 3GPP TS 23.288, clause 5.3 ....

Federated learning among multiple NWDAFs is a machine learning technique in core network that trains an ML Model across multiple decentralized entities holding local data set, without exchanging/sharing local data set. This approach stands in contrast to traditional centralized machine learning techniques where all the local datasets are uploaded to one server, thus allowing to address critical issues such as data privacy, data security, data access rights.

NOTE 1: Horizontal Federated Learning is supported among multiple NWDAFs, which means the local data set in different FL client NWDAFs have the same feature space for different samples (e.g. UE IDs).

For Federated Learning supported by multiple NWDAFs containing MTLF, there is one NWDAF containing MTLF acting as FL server (called FL server NWDAF for short) and multiple NWDAFs containing MTLF acting as FL client (called FL client NWDAF for short), the main functionality includes:

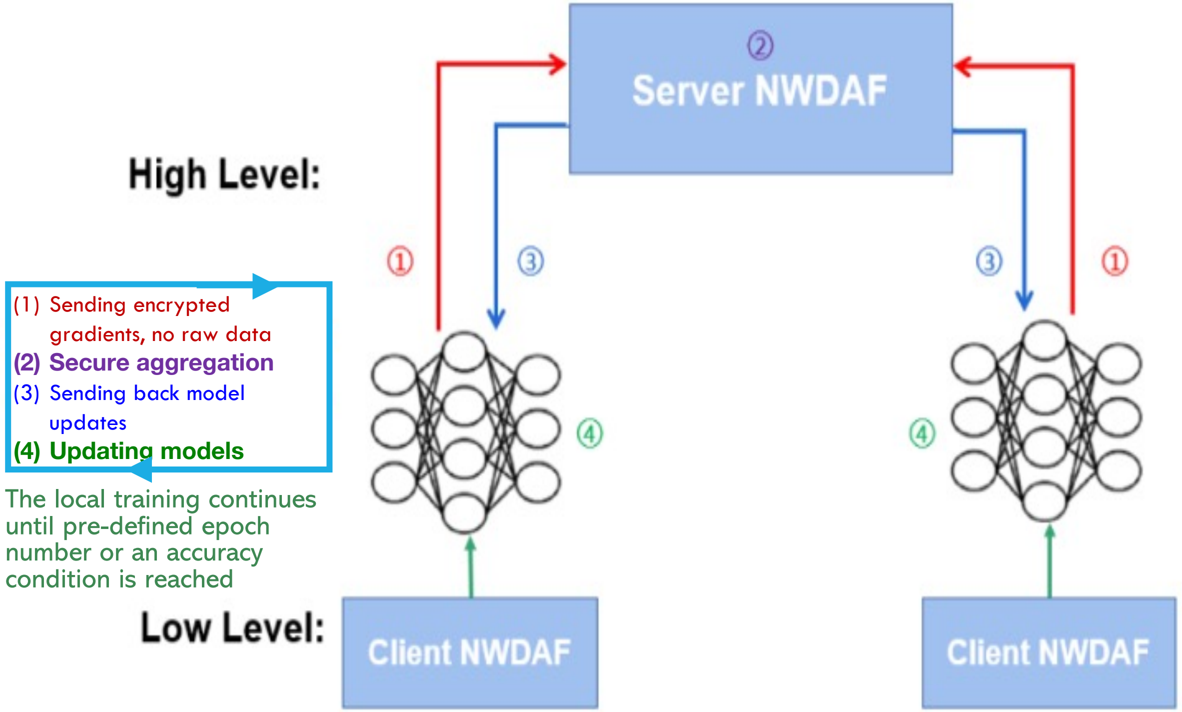
FL server NWDAF:

- discovers and selects FL client NWDAFs to participant in an FL procedure
- requests FL client NWDAFs to do local model training and to report local model information.
- generates global ML model by aggregating local model information from FL client NWDAFs.
- sends the global ML model back to FL client NWDAFs and repeats training iteration if needed.

FL client NWDAF:

- locally trains ML model that tasked by the FL server NWDAF with the available local data set, which includes the data that is not allowed to share with others due to e.g. data privacy, data security, data access rights.
- reports the trained local ML model information to the FL server NWDAF.
- receives the global ML model feedback from FL server NWDAF and repeats training iteration if needed.

FL server NWDAF or FL client NWDAF register to NRF with their FL capability information as described in clause 5.2.



Basic Architecture Framework for Federated Learning is supported in TODAY's Mobile Core

SECURITY AND PRIVACY ISSUES

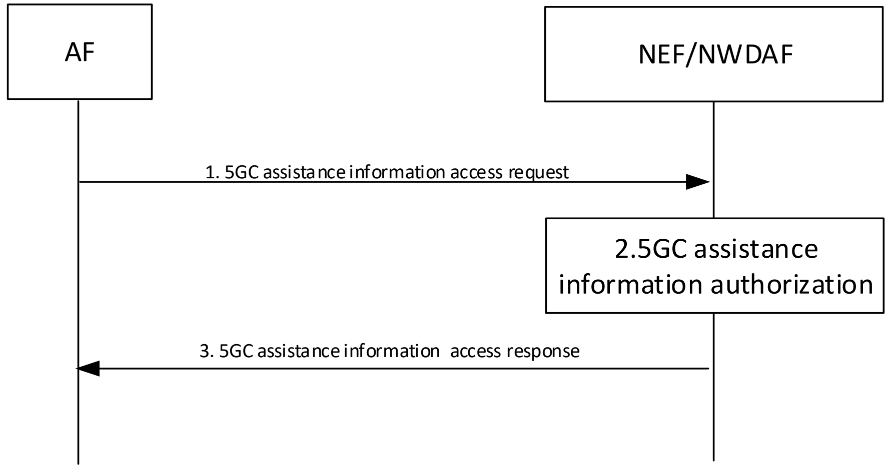
- Many key issues identified, only one has so far been worked out
- Key Issue: Privacy and authorization for 5GC assistance information exposure to AF
- the exposure of different types of assistance information such as traffic rate, packet delay, packet loss rate, network condition changes, candidate FL members, geographical distribution information, etc. to AF for AI / ML operations.
- Privacy issues Some of assistance information such as candidate FL members, geographical distribution information could be user privacy sensitive. In some cases a single piece of information alone would not be considered as privacy-sensitive, but the combination of that piece of information along with other seeming unrelated privacy data could potentially reveal user privacy. The mobile core network needs to determine which assistance information is required by AF to complete AI/ML operation and to avoid exposing information that is unnecessary for AI/ML operations.
- Security threats Without proper privacy protection mechanism, UE's privacy information may be leaked resulting in loss of user privacy. Unauthorized access of the mobile core network assistance information by AF can lead to misuse and user privacy leakage
- Security requirements The mobile core network shall support the protection of user privacy sensitive assistance information being exposed to AF.
  The mobile core network shall support authorization of AF for accessing assistance information.

SECURITY AND PRIVACY ISSUES

- UE profile based solution
- UE privacy profile/local policies may also contain protection policies that indicate how 5GC assistance information should be protected (e.g., encryption, integrity protection, etc.).
- The UE privacy profile is stored in the UDM/UDR. For each UE, the UE privacy profile determines whether the specific AF can request or modify specific information of a specific UE.
- UE profile includes UE identity (e.g., SUPI, SUCI, IMPI, Application layer ID of UE, GPSI), expected service identifier, data type of target 5GC assistance information (e.g., location information), granularity of target 5GC assistance information type (e.g., TAI for location information), expiration time (expiration), authorization policies (e.g., specific UE related 5GC assistance information can be handled by a specific service.), protection policies (e.g., a specific UE related 5GC assistance information needs to be encrypted before sharing to AFs).
- TLS is used to provide integrity protection, replay protection and confidentiality protection for the interface between the NEF and the AF. The support of TLS is mandatory

AUTHORIZATION EXAMPLE

- AF sends 5GC (mobile core) assistance information request to the NEF/NWDAF. The request includes the AF identity (e.g., AF_ID, Application layer ID, FQDN), expected service identifier, data type of target mobile core assistance information (e.g., location information), details of target 5GC assistance information (e.g., TAI), target UE identity (e.g., IMPI, Application layer ID of UE, GPSI).
- Upon receiving the request, NEF/NWDAF identifies the UE privacy profile according to the target UE identity. If NEF/NWDAF does not contain the UE privacy profile, NEF/NWDAF obtain the profile from UDM/UDR.
- NEF/NWDAF leverages the local policies/UE profile to check if the UE authorizes the AF to access the UE-related mobile core assistance information
- NEF/NWDAF sends the UE-related mobile core assistance information to AF when the local policies/UE privacy profile authorize the AF to access the information. According to the local policies/UE privacy profiles, NEF/NWDAF may need to protect the mobile core network assistance information with security mechanisms.
- For authorization OAuth (RFC 6749) is used
- For security TLS 1.3 (RFC 8446) is used



CONCLUSIONS

So far we presented 3GPP work as of early 2024, mostly in Release 19

3GPP 23.700-80-i100 has identified 7 key issues

3GPP 23.700-82-030 which is SA6 work has identified 7 key issues relevant for SA6

3GPP 33.898-i01 provides privacy and security solutions for one issue as discussed in earlier slides

No normative work is done probably will be done in Release 20 when 6G work starts (?)

We suggest it is good time for IETF to be involved and look into the security and privacy key issues and develop solutions

Q&A