# AI ML Mobile Device (selected) Specification Requirements

# with

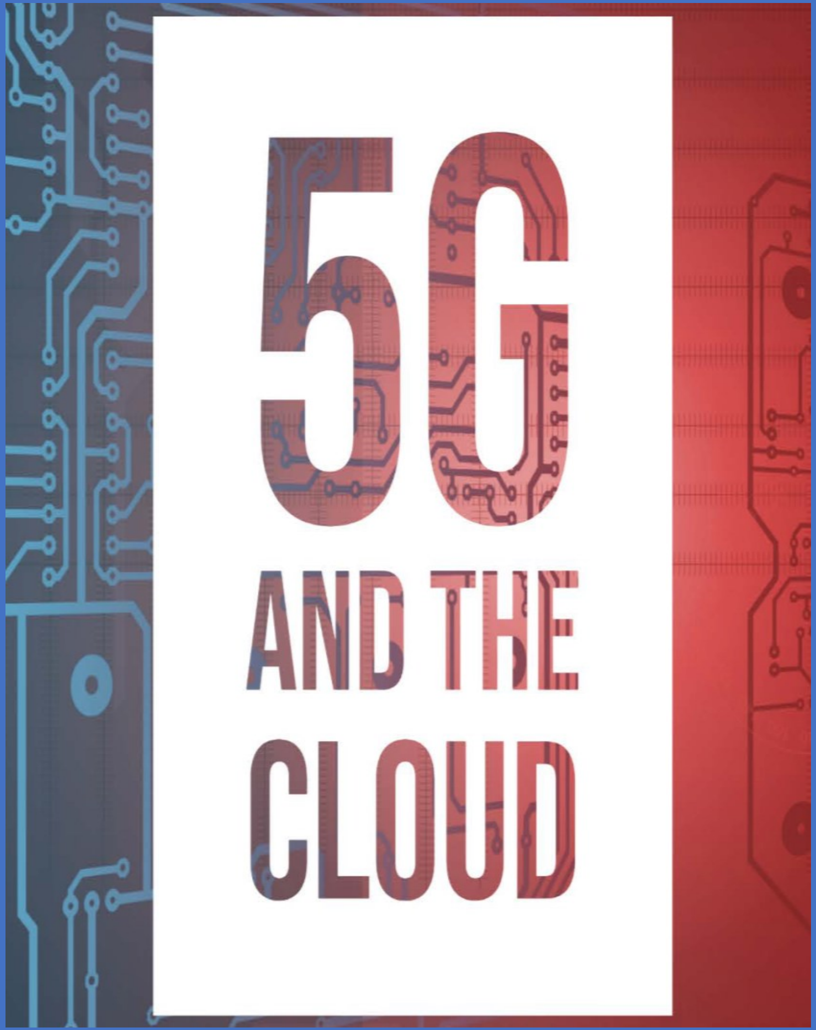
# (selected) use of AI ML in 5G Network

Ike Alisson

2023 - 11 -20  Rev PA03

# Table of Contents

## 1. CLOUD NATIVE PHILOSOPHY-RELATED ISSUES




The Cloud Native issues appear because the whole of the Cloud Native Development Philosophy has been applied:

- *without consideration* of the *Actual Deployment and Operational Environments.*

In brief, **the *Positive and Negative Aspects of Cloud Native from a 5G SA/SBA Network Function (NF) Perspective* are summarized as follows:**

| POSITIVE ( + ) | NEGATIVE ( - ) |
|---|---|
| Cloud Native has undeniably improved:<br><br>- Development,<br>- Delivery and Test,<br>- In-Service Upgrades<br>- Improved Version Management | The Context in which *Cloud Native* was designed is being *misrepresented or abused in two (2) senses:*<br><br>1. Cloud Native was *designed for People who write & operate the Applications*.<br><br>In today's Cellular Network, this clearly is not the case<br><br>2. ***Cloud Native* was designed for *Applications in which long interruptions*** *are tolerable*, therefore, *good Reliability is measured* ***in minutes of outage per month.***<br><br>This is also clearly not the case for **(2G, 3G, 4G, 5G) Cellular Communication Networks** where the expectation is that **outages last less than 5.26 minutes per year.** |

## 1. Cloud & Communications Systems' (current) Challenges & Issues

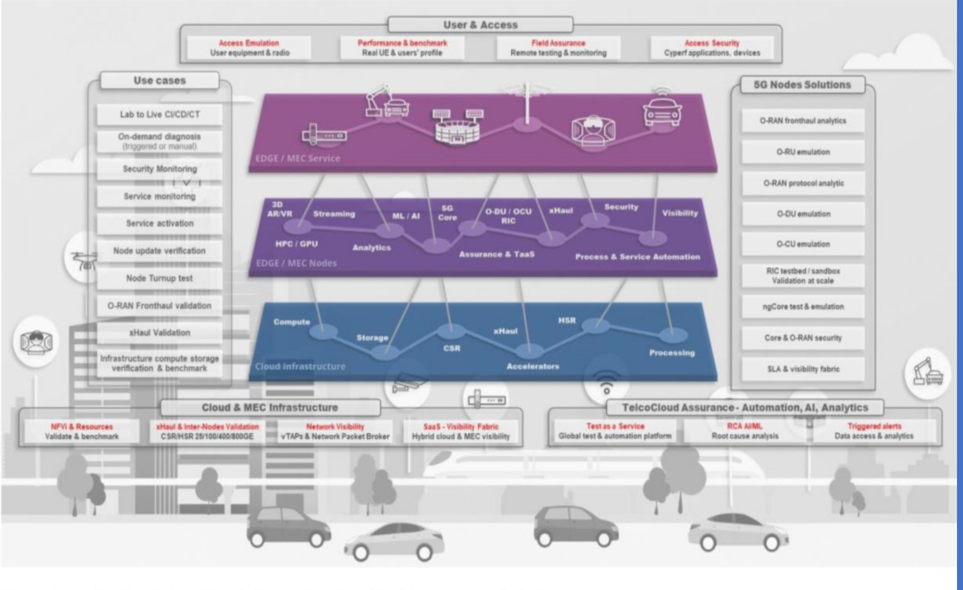



Figure : Telco Edge Cloud, Next-Gen Service Assurance at Scale

*Today's Cloud and Communications Systems are NOT CAPABLE of*

- *Capturing,*
- *Transmitting,*
- *Storing, and*
- *Analysing*

*the Petabytes of Data generated by the soon-to-be trillions of Sensors operating 24/7.*

***They are also NOT PREPARED to deliver the Compute needed for Real-Time AI/ML Inferencing required to drive such demands that we anticipate will come from:***

- *FoF (Factory of the Future)*
- *VR/XR/MR (Virtual, Extended, Mixed Reality and Extended Reality) with Haptic Interactions,*
- *NPNs/SNPNs Non Public Network/Stand-alone NPNs*
- *PINs and CPNs (Personal IoT Network/Customer Premises Networks)*
- *(V2X) Connected Vehicles,*
- *Assisted living, or*
- *Merging of Physical & Digital worlds with 5G & B5G*

# 1. The Cloud is "Changing"

**1st - Applications want to be deployed anywhere & change deployment anytime.**

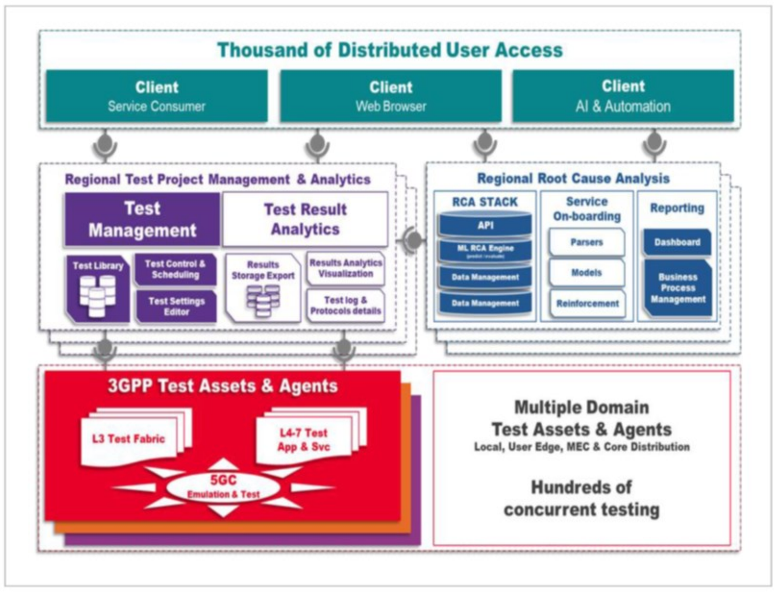



Fig – Nationwide Distributed Cloud based Network Test Bed Orchestration

***The focus moves from "Sharing Resources" to "Composing Dynamic Capabilities, in Real-time, even after Deployment.***

***Applications will be Delay- and Latency Sensitive*, on *varying Time-scales* with *different Hard- & Soft Boundaries.***

**Communication, Compute, and Storage** must be considered as an ***Integrated Set of Changeable Configurations*** that provide the required Service to an application.

**2nd - "Centre of Gravity is moving toward the "Devices" ("End-points"*) & Interactions in a Cyber-Physical World** best suited for these tasks and configure any required communication between all end points in important areas such as

- IoT,
- Industry 4.0,
- 5G NPNs/SNPNs/PINs, or
- Retail and Public Services.
- eHealth & Ageing and Living well

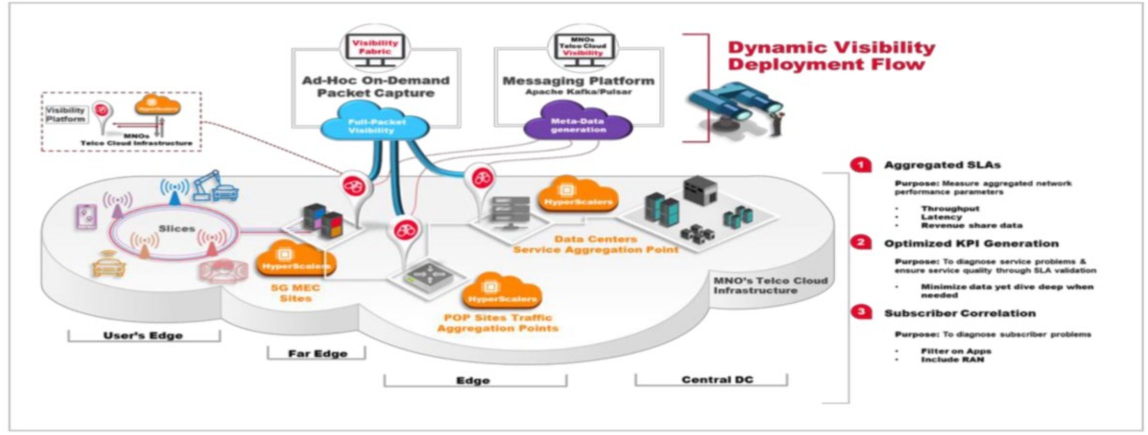



Figure : Hybrid Network Visibility Platform architecture

**You might be vigilant with the terms you use w.r.t. the terms "end-points" &/or "Edge" from Service E2E Solution Architecture fulfilling the 3GPP specified 5QI (QoS) Service Requirements & KPIs.*

## Management of Resources and Workloads:

**Most Orchestration Frameworks today use a Centralized Approach** (where) One (1) Entity has knowledge of all the Resources in the System and Plan how the Workloads will be mapped.

With the start of Docker & containers, the Kubernetes Project was started to provide a lightweight & scalable Orchestration solution.

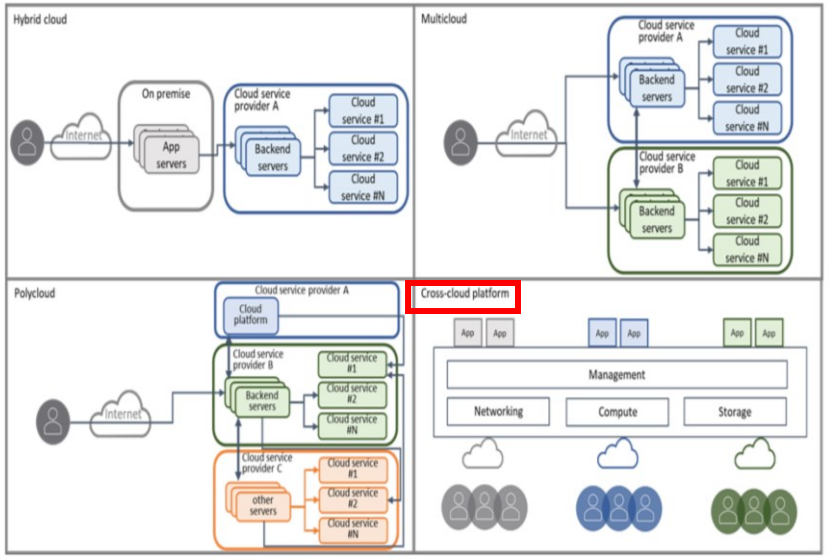
Most existing Compute Systems today, including Edge Computing Systems, rely **on "Static Provisioning"**.

Thus, the SW & the Services needed to perform the Compute are already residing at the Edge Server prior to an Edge node requests a Service & the pool of HW resources is also known a priori to Kubernetes.

*This Architecture works well for Cloud & the (ETSI) MEC where a Centralized Orchestration is used.*

Since the Resources of the Pervasive Edge are independently owned, the ***Orchestration Frameworks need to be extended to handle Dynamic and Multi-Tenant Resources*** *in a secure manner.*

Figure multi-cloud deployment models



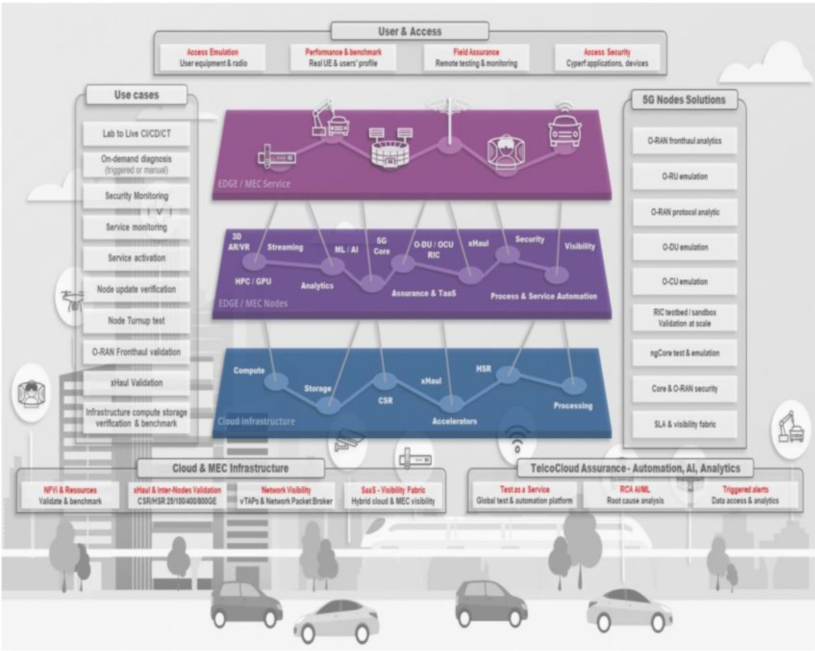




Figure :Telco Edge Cloud, Next-Gen Service Assurance at Scale

# 1. 1. Introducing new Authentication Layer in 3GPP 5G Advanced Rel-19 - 1

5G System Architecture specifications foresee the *introduction of an optional, User-centric Authentication Layer on top of the existing Subscription Authentication, supporting various Authentication Mechanisms and interactions with external Authentication Systems* as well as a Degree of Confidence (i.e. a Value that allows Differentiated Service Policies *depending on the Reliability of the User Identifier).*

The *New Authentication Layer* shall not replace existing Subscription Credentials.

The Security and Privacy of Subscriber or End User Data shall not be compromised.

Use Cases (UCs) are developed and potential requirements derived how to use the new User Identifier within the 3GPP System e.g. to provide Customized Services and enhanced Charging and how to provide this Identifier to external Entities to enable Authentication for Systems and Services outside 3GPP.

Use Cases (UCs) for use within 3GPP include:

- Providing different Users using the same UE with Customized Services

- identifying Users of Devices behind a GW with a 3GPP Subscription, but without the Devices having a dedicated 3GPP Subscription.

- Using a User Identifier being linked to a Subscription to access 3GPP Services via Non-3GPP access

- Using a User Identifier for Slice Authorization.

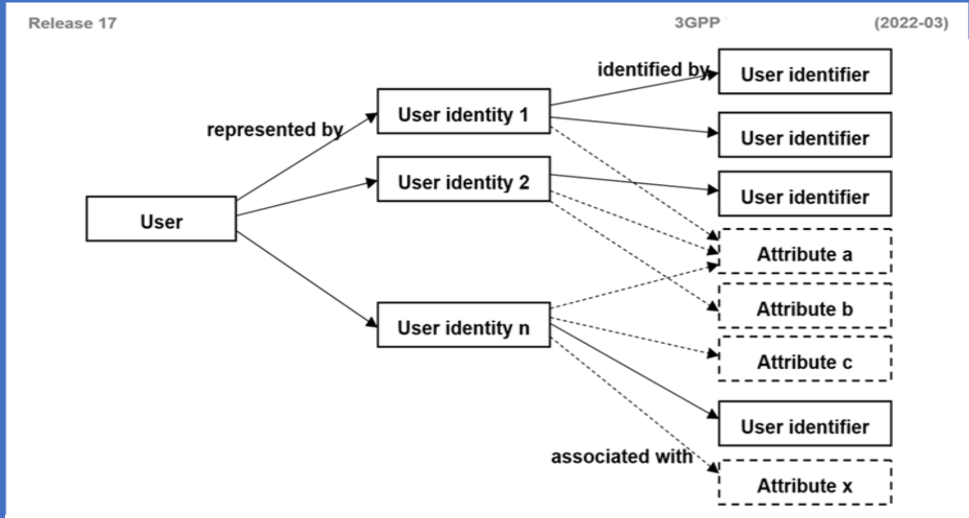



Figure: 5G System foreseen User-centric Authentication Layer on top of the existing Subscription Authentication showing the Relation between User, Identities, Identifiers and Attributes

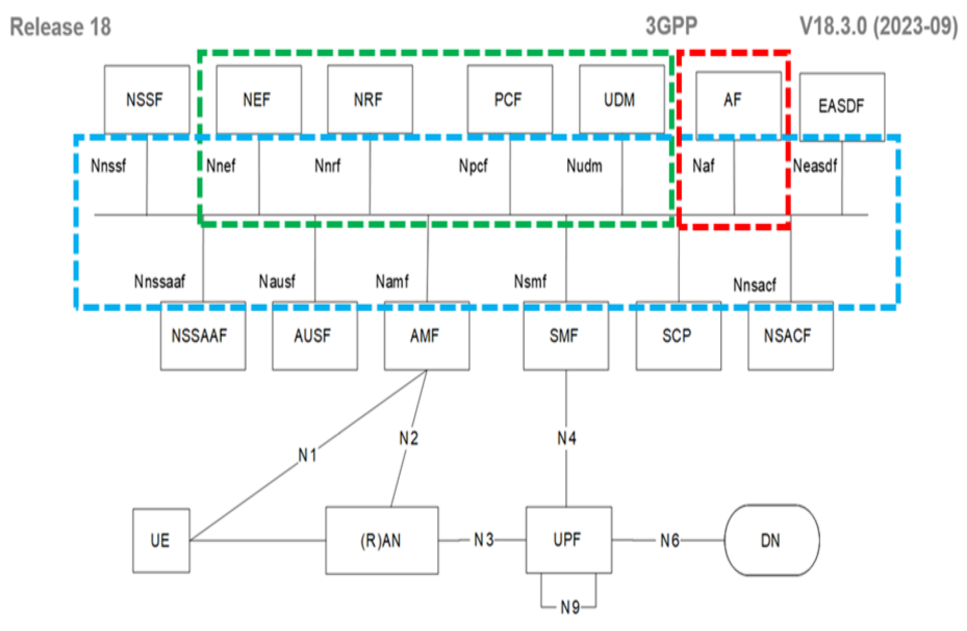



Figure: 5G System Non-Roaming Architecture

Current Mobile Networks are Subscription-centric, which allows Mobile Operators to protect the Access to the Network and respect legal obligations.

From a UC perspective, this was sufficient in times when a User typically only had one (1) Phone with one (1) Subscription, using only a few Services provided by the Operator such as:

A) Telephony and
B) SMS.

## ***However, Times have Changed:***

Today, a Person may have different kinds of Devices (e.g. Phones, Tablets, Laptops), some of which might belong to the User, others might be shared with someone else or belong to some other Party to access various Operator and Non-Operator Services.

Things are increasingly connected (Sensors, GWs, Actuators etc.) and there are many different flavours in the relation between:

A) the Owner of the Thing,
B) the Holder of the Subscription and
C) the Actual User of the Thing.

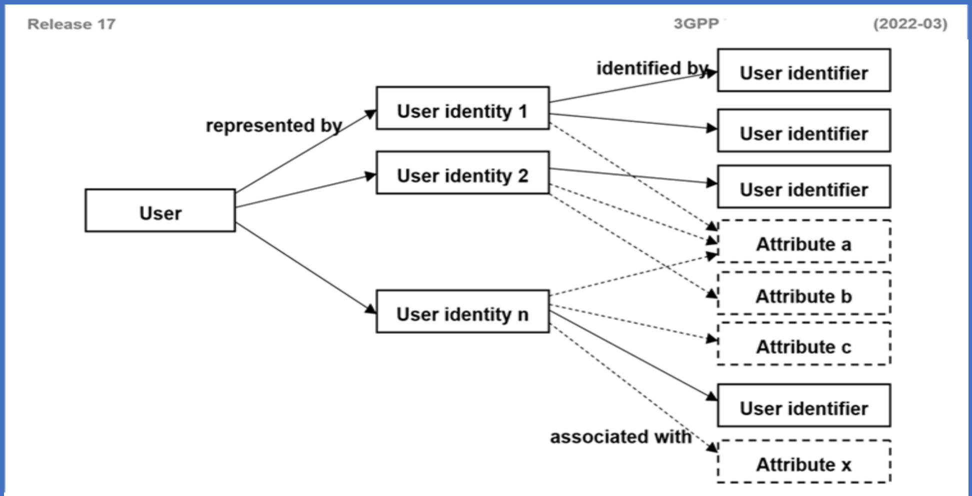



Figure: 5G System foreseen User-centric Authentication Layer on top of the existing Subscription Authentication showing the Relation between User, Identities, Identifiers and Attributes

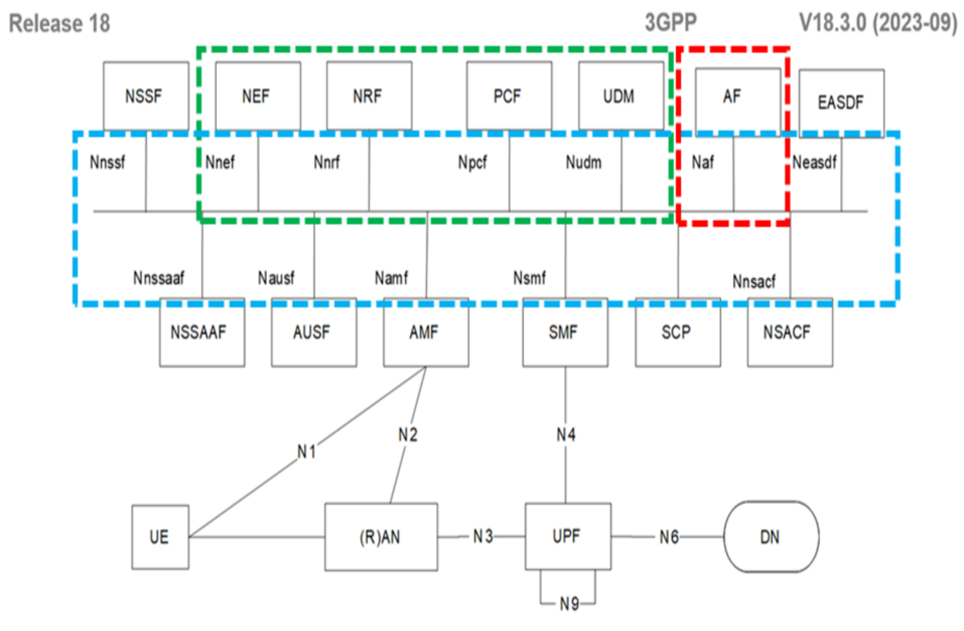



Figure: 5G System Non-Roaming Architecture

Presently, it is common for each Service to perform its own Authentication, often based on User-name and Password/Passkey.

For Users it becomes more and more cumbersome to manage the different Credentials of the growing number of services.

So-called identity Providers address the above problem by providing Identity Information to "Entities" and Authentication to Services for those Entities.

Such Mechanisms could be used over the top of any Data Connections, but integration or interworking with Operator Networks provides additional advantages.

Identifying the User in the Operator Network (by means of an "Identity" provided by some external party or the Operator) enables to provide an enhanced User Experience and optimized Performance as well as to offer Services to Devices that are not part of 3GPP Network.

*The "**User**" to be identified could be:*
*A) **Individual Human User,** using a UE with a certain Subscription, or*
*B) **Application** running on or connecting via a UE, or*
*C) **Device ("Thing")** behind **a GW UE** (e.g. **5G PIN PEGC**)*

***Network Settings can be adapted and Services offered to Users according to their "Needs", independent of the Subscription that is used to establish the Connection.***

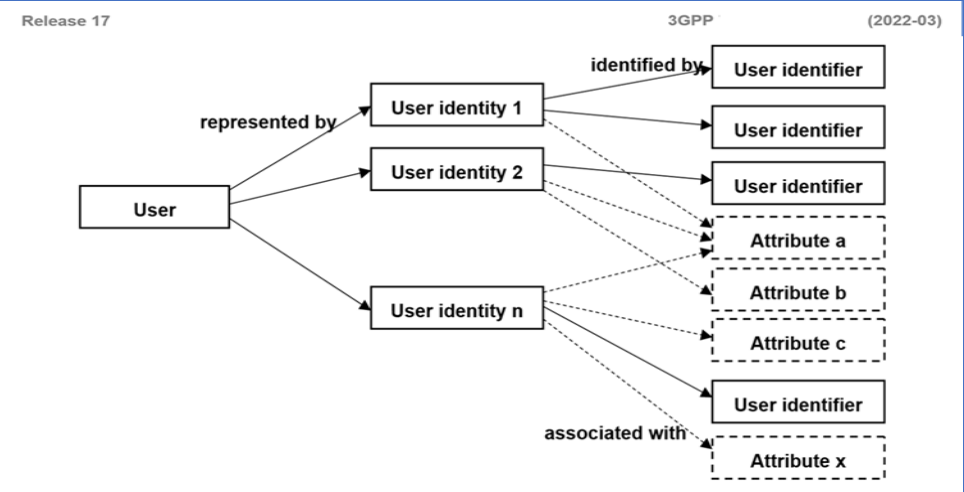



Figure: 5G System foreseen User-centric Authentication Layer on top of the existing Subscription Authentication showing the Relation between User, Identities, Identifiers and Attributes

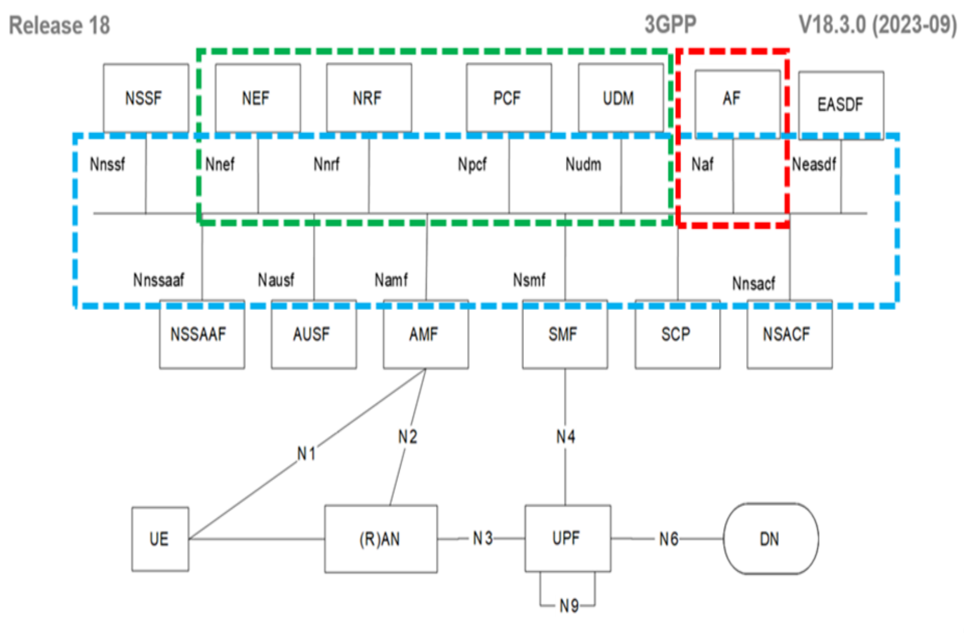



Figure: 5G System Non-Roaming Architecture

In the context of identity management something outside a system that needs to be identified in the system is referred to as "entity".

In 3GP, such an "Entity*" is called a User.

A "User "is not necessarily a Person, it could also be an Application or a Device ("Thing").

***The "Entity*" is uniquely represented by an Identity in the System.***

The Identity can dependent on the Role of the Entity in the System (e.g. which Kind of Service is used for which purpose).

*As such, a "User" can have several "User Identities": e.g. :*
*- one (1) User Identity representing the Professional Role of the (Human) User and*
*- another one representing some aspects of her Private Life.*

***There is a 1:n Relation between User and User Identity.***

**The state related to "Entity" as being re-defined within the updated definition of "Context" used in 3GPP 5G System Architecture, as part of 5G NDL (Network Data Layer) and ETSI*

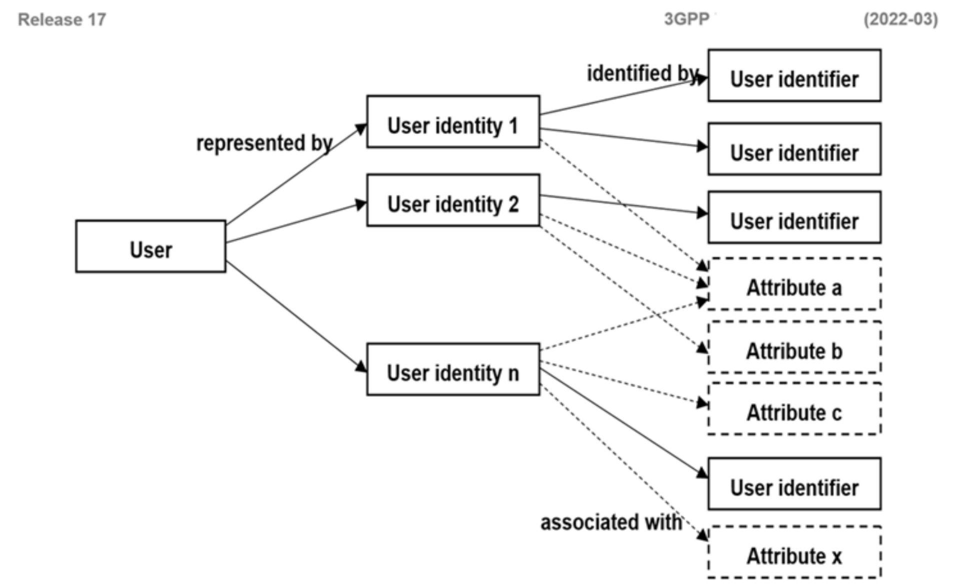



Figure: 5G System foreseen User-centric Authentication Layer on top of the existing Subscription Authentication showing the Relation between User, Identities, Identifiers and Attributes

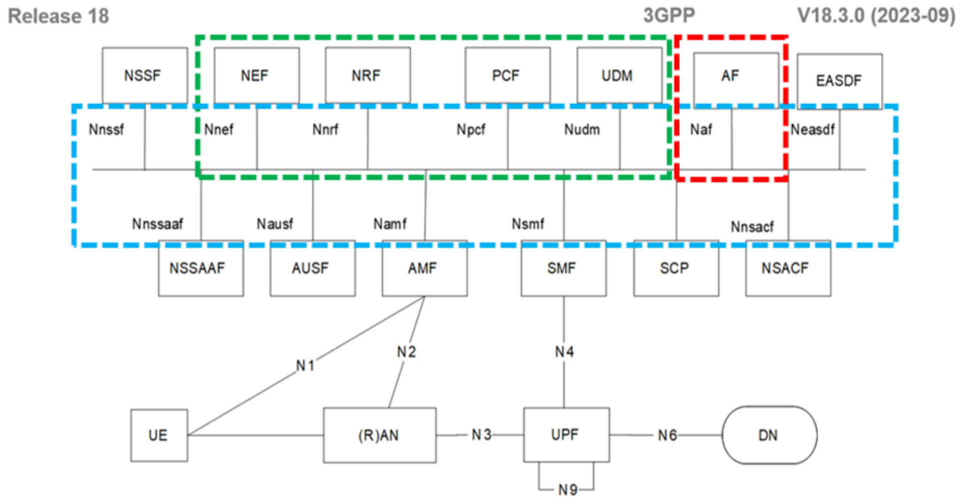



Figure: 5G System Non-Roaming Architecture

***There is a 1:n Relation between "User" and "User Identity".***

A "U*ser Identity*" is associated with some Pieces of Information, which are generally called "Attributes".

*One "Special Form of Attributes" are "Identifiers".*

The relation between "Identity" and "identifier" is 1:n.

Each User Identity is identified in the System by one (1) or more User Identifiers.

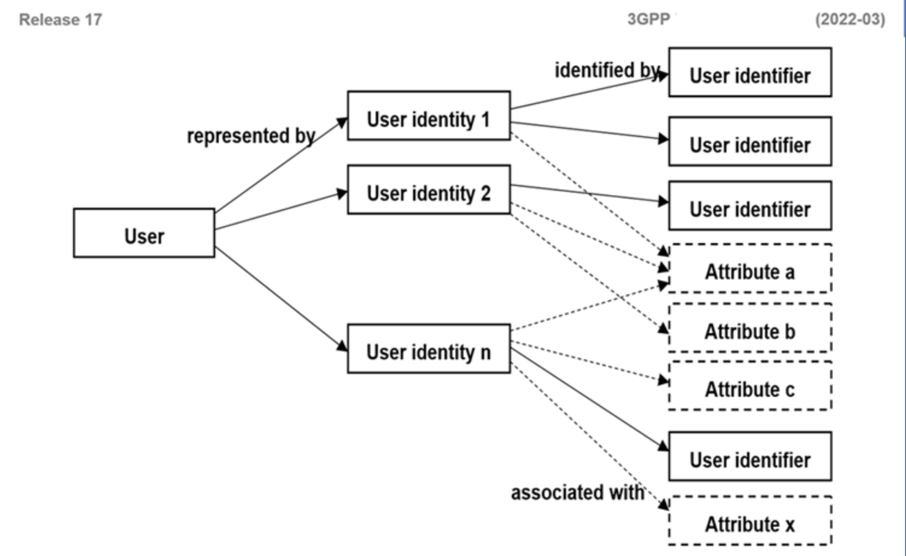



Figure: 5G System foreseen User-centric Authentication Layer on top of the existing Subscription Authentication showing the Relation between User, Identities, Identifiers and Attributes

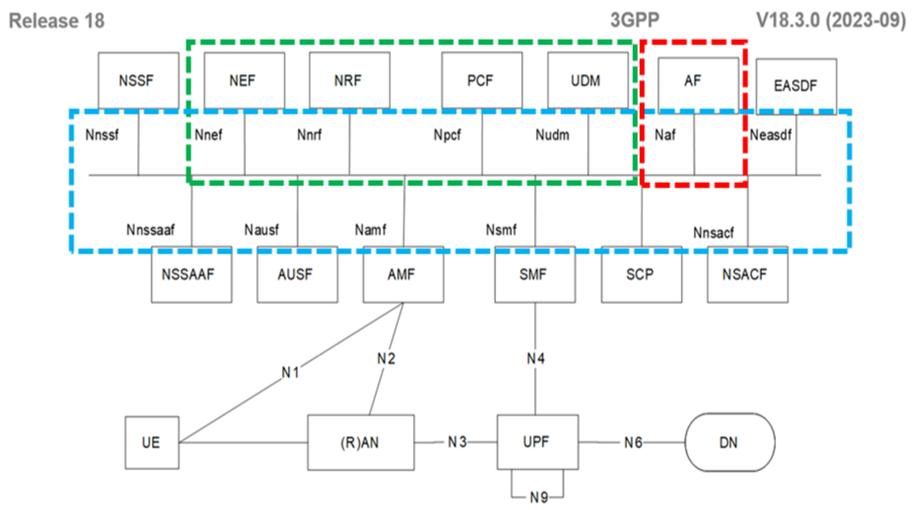



Figure: 5G System Non-Roaming Architecture

## Each User Identity is identified in the System by one (1) or more User Identifiers.

An "Identifier" could take the form of an
- NAI (Network Access Identifier)
- Email Address or
- some Number,
- could be "Permanent" (comparable to the *IMSI e.g. SUPI/SUCI*), or
- "Temporary" (comparable to the "TMSI").

E.g., in the internet-world a user might choose to use her company email address when registering and using services (access to web portals) that she needs for her work.

For access to other sites, e.g. online shopping or login to information servers concerning some hobby, she might use other email addresses. In this example the email addresses are the user identifiers that identify the different identities of the user for certain Web Services.

Other attributes could contain information about
- the Date of Birth of a User,
- the Private Address,
- the Company Name & Address,
- Job title etc.

Attributes that are no identifiers may be associated with more than one identity, e.g. Date of Birth might be relevant in the Professional as well as in the Private context.

*One Identity typically is associated with several Attributes.* With having multiple user accounts the above information is distributed over multiple servers. An identity provider creates, manages and stores this information in one place, authenticates a selected user identity (i.e. verifies a claimed user identity) for a service and provides the result and necessary attributes to the service.

Release 17 3GPP (2022-03)

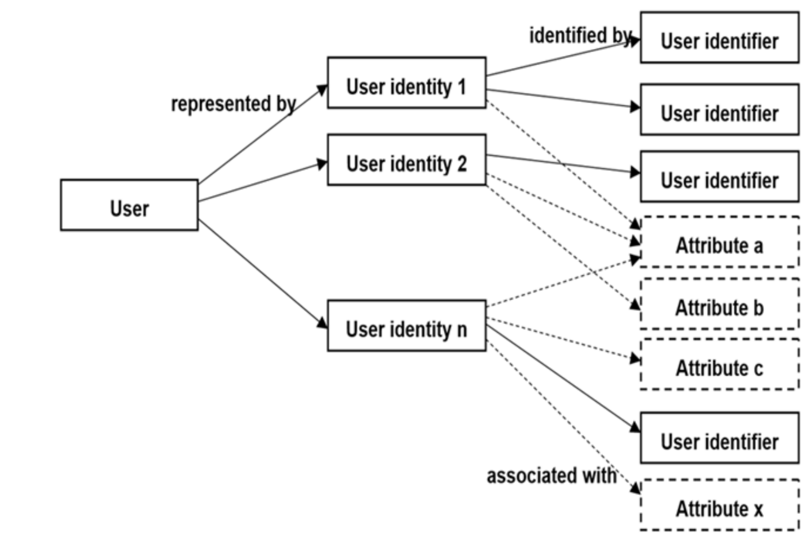



Figure: 5G System foreseen User-centric Authentication Layer on top of the existing Subscription Authentication showing the Relation between User, Identities, Identifiers and Attributes

Release 18 3GPP V18.3.0 (2023-09)

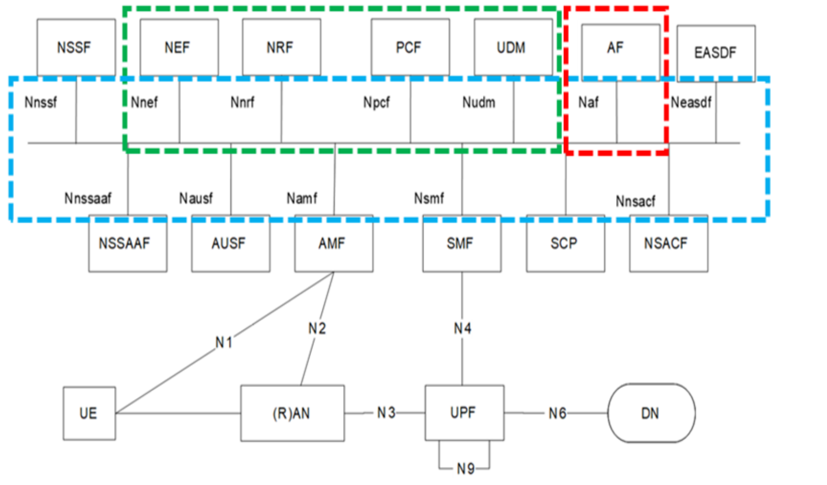



Figure: 5G System Non-Roaming Architecture

# 1. Introducing new Authentication Layer in 3GPP 5G Advanced Rel-19 - 7

*Impact on the 3GPP 5G System*

The Objective of this evolvement does not aim to define an Identity Provisioning Service.

The actual process of identity creation, provisioning, managing, authentication etc. does not need to be defined within 3GPP.

The focus is to outline the interaction of such a Service with the 3GPP System w. r. t.:

- How to take a User Identity into account for adapting Network and Operator-deployed Service Settings (e.g. Policies) and for Network Slice (SST) Selection;

- Support of Providing the User Identity to External Services via the 3GPP Network;

- Extending 3GPP Services to Non-3GPP Devices that are identified by User Identifiers, e.g. to enable Network and Service Access by these Devices and to make them addressable and reachable from the Network ;

- additionally, if the operator acts as Identity Provider, How to improve the Level of Security or Confidence in the Identity by taking into account Information from the Network

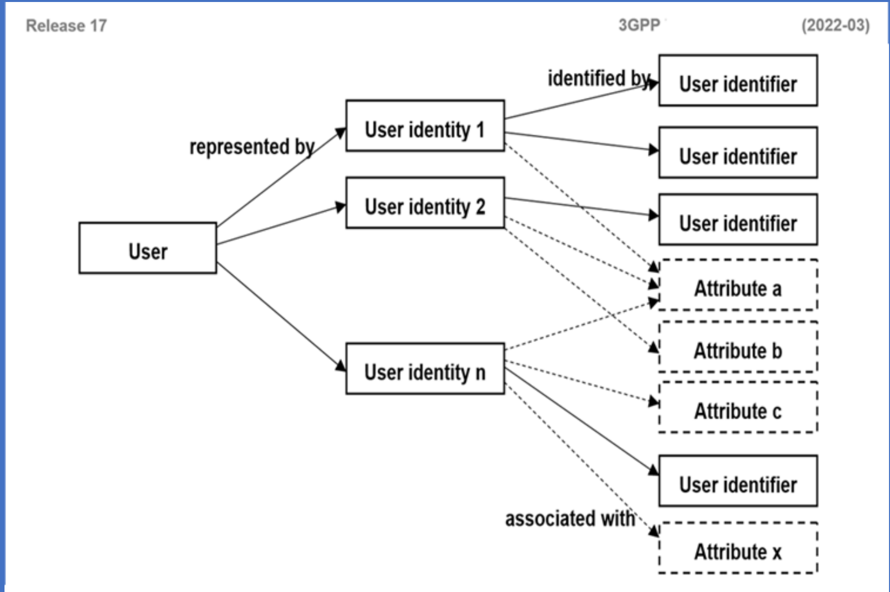



Figure: 5G System foreseen User-centric Authentication Layer on top of the existing Subscription Authentication showing the Relation between User, Identities, Identifiers and Attributes

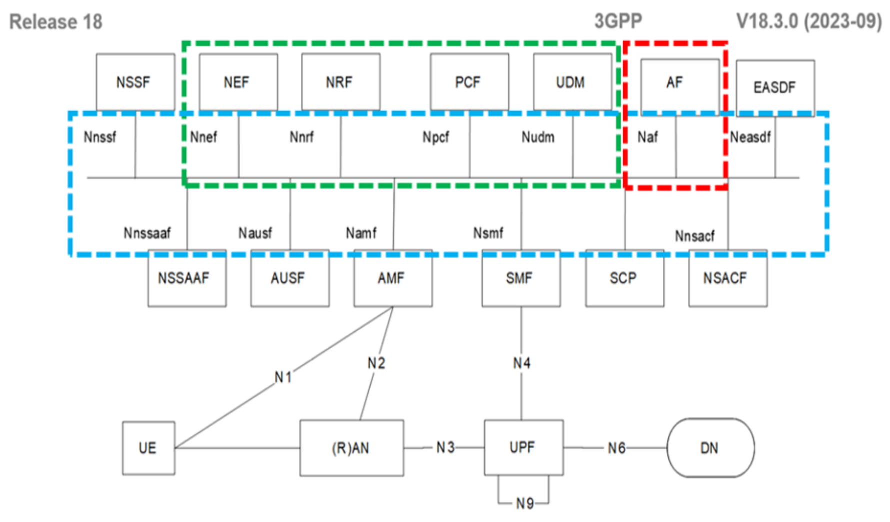



Figure: 5G System Non-Roaming Architecture

## 2. AI ML Mobile Device (selected) Requirements Specification - 1

### *AI Mobile Device** Definition

An *AI Mobile Device* refers to **a Mobile Device that has all of the following Characteristics**:

1. *On-Device Computational Resources* to enable AI Deep Learning (DL) and other AI Algorithms based on either *dedicated AI Hardware (HW) or general HW* to support Deep Learning (DL) AI Applications.

2. *On-Device Software (SW) Framework* to *support the updating of AI Deep Learning Neural Networks* (DLNN).

3. *On-Device AI Software (SW)* to perform *Inferencing* using Deep Neural Network (DNN) Models.

**Added Information on selected AI ML Mobile Devices information from selected Device (& chip) suppliers as e.g. Qualcomm, Google, Apple, Huawei, etc. and their native AI ML Mobile Device API specifications is included in slides 6-8*

Release 19 3GPP V19.1.0 (2023-09)

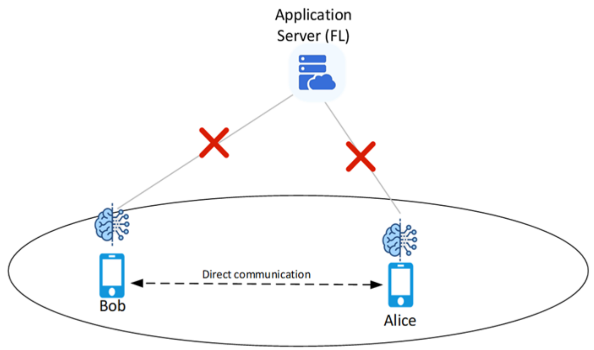



**Figure: 5G AI ML Mobile Devices, two UEs, performing Decentralized Federated Learning (FL) using Direct Device Connection**

Release 19 3GPP V19.4.0 (2023-09)

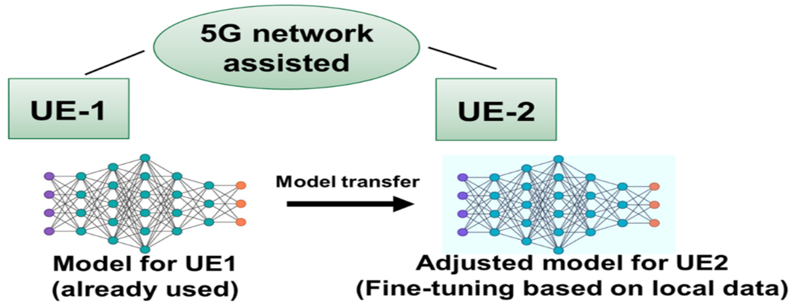



**Figure: 5G Network-assisted AI/ML Model Transfer Learning from Source to Target UE**

Release 19 3GPP V19.1.0 (2023-09)

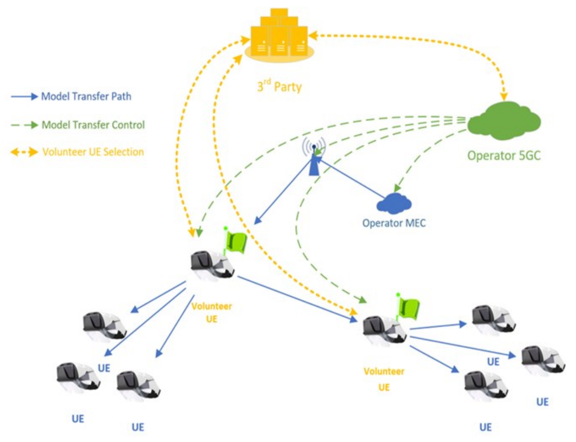



**Figure: 5G System Architecture AI/ML Model Management through Direct Device Connection**

## 2. AI ML Mobile Device (selected) Requirements Specification - 2

### The Requirements of AI ML Mobile Device - *AI ML Device Hardware (HW) Requirements*

AI Mobile Device HW is required to support AI SW Applications efficiently. HW Performance Measurements can be found in the Table below using the modified VGG 16. Alternatively, a better Network might be used.

| Requirement for the modified VGG 16 network | |
|---|---|
| TS47_3.1_REQ_001 | An AI Mobile Device SHOULD have a minimum of (1) int8 TOPS. |
| TS47_3.1_REQ_002 | An AI Mobile Device SHOULD have a minimum of (0.5) float16 TOPS. |
| TS47_3.1_REQ_003 | An AI Mobile Device SHOULD have a minimum of (0.5) int8 TOPS/Watt. |
| TS47_3.1_REQ_004 | An AI Mobile Device SHOULD have a minimum of (0.3) float16 TOPS/Watt. |

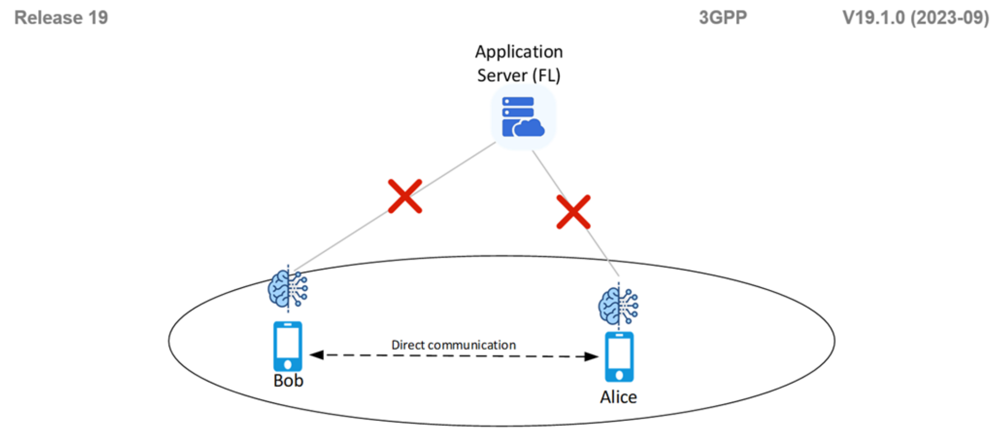



Figure: 5G AI ML Mobile Devices, two UEs, performing Decentralized Federated Learning (FL) using Direct Device Connection

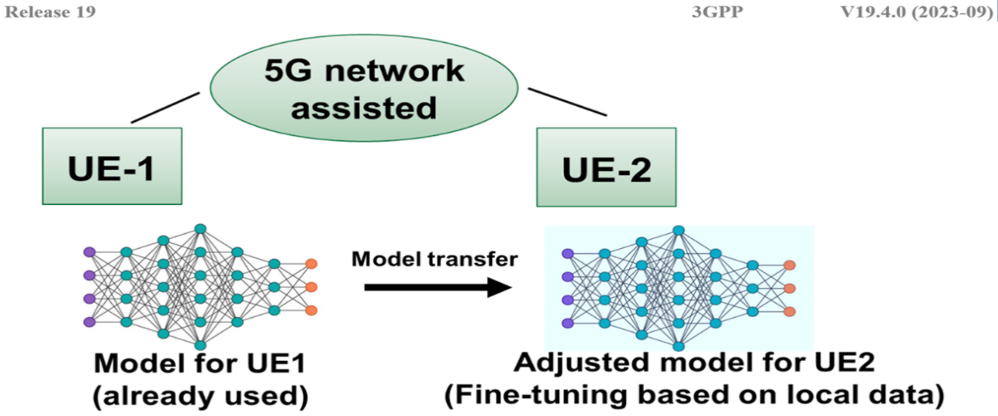



Figure: 5G Network-assisted AI/ML Model Transfer Learning from Source to Target UE

# 2. AI ML Mobile Device (selected) Requirements Specification - 3

## The Requirements of AI ML Mobile Device - *AI ML Device Software (SW) Requirements*

| | |
|---|---|
| REQ_001 | An AI Mobile Device SHALL support on-device model updates of an existing deep learning network. |
| REQ_002 | An AI Mobile Device SHALL support native APIs to expose the AI hardware functions. |
| REQ_003 | An AI Mobile Device SHALL support application APIs (See Appendix A) for native and third-party applications to access Computer Vision (CV), Automatic Speech Recognition (ASR), Natural Language Understanding (NLU) models. |
| REQ_004 | An AI Mobile Device SHOULD provide an SDK to convert DNN models from an existing format to the native format of the AI mobile device. Non-exhaustive examples of DNN model file format are: *.ckpt or *.pb, *.tflite, *.prototxt, *.pb or *.pth or *.pt, *.json and *.onnx. |
| REQ_005 | An AI Mobile Device SHOULD provide an SDK to support definition of new customized Deep Learning operators. |

Release 19 3GPP V19.1.0 (2023-09)

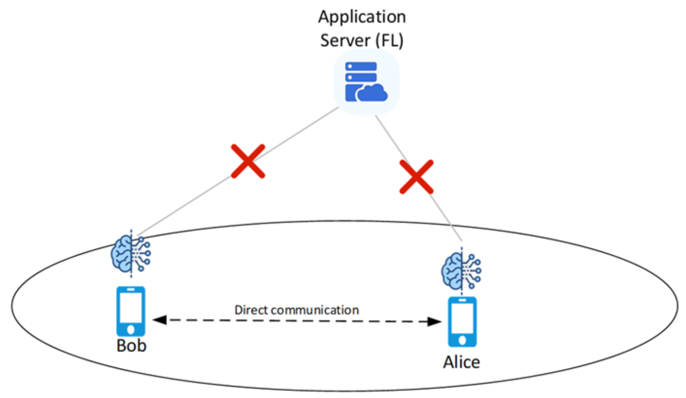



**Figure: 5G AI ML Mobile Devices, two UEs, performing Decentralized Federated Learning (FL) using Direct Device Connection**

Release 19 3GPP V19.4.0 (2023-09)

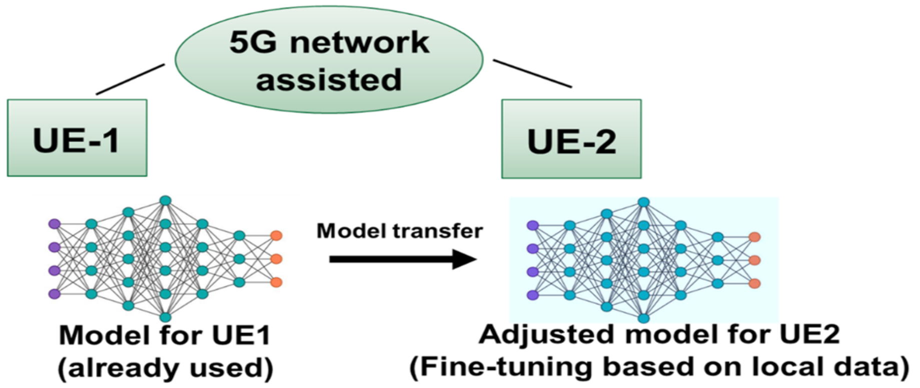



**Figure: 5G Network-assisted AI/ML Model Transfer Learning from Source to Target UE**

# 2. AI ML Mobile Device (selected) Requirements Specification - 4

The Requirements of AI ML Mobile Device - *AI ML Device Application Requirements*

*AI Applications may include, but are not limited to:*
*- Biometric Functions,*
*- Image Processing,*
*- Speech,*
*- Augmented Reality (AR) and*
*- System Optimization Categories.*

*If any such Functions are supported on the Device then the following requirements apply.*

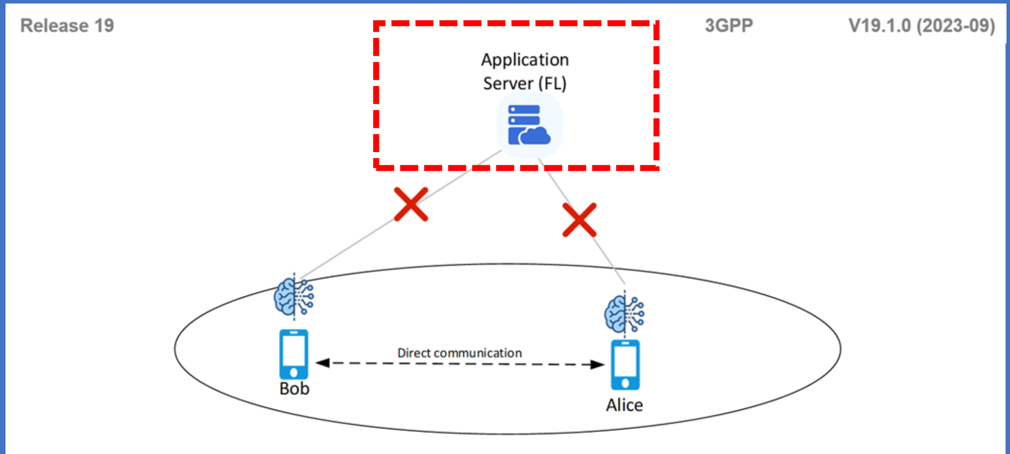



Figure: 5G AI ML Mobile Devices, two UEs, performing Decentralized Federated Learning (FL) using Direct Device Connection

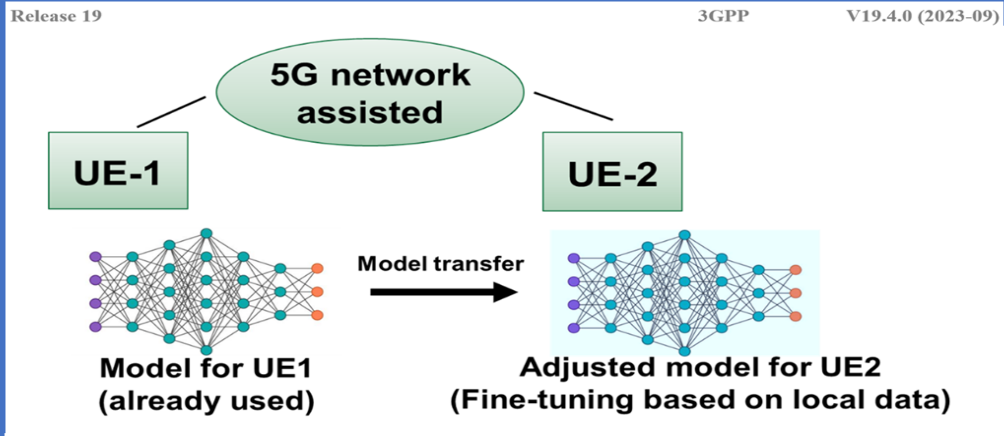



Figure: 5G Network-assisted AI/ML Model Transfer Learning from Source to Target UE

# 2. AI ML Mobile Device (selected) Requirements Specification - 5

## The Requirements of AI ML Mobile Device - *AI ML Device Application Requirements*

*AI Applications may include, but are not limited to:*
*- Biometric Functions,*

*If any such Functions are supported on the Device,*
*then the following requirements apply.*

**Biometric Performance Requirements**

| | |
|---|---|
| TS REQ_001 | An AI Mobile Device SHOULD support a 2D facial biometric system. |
| TS REQ_002 | An AI Mobile Device SHOULD support a 3D facial biometric system. |
| TS REQ_003 | An AI Mobile Device SHOULD support a fingerprint biometric system. |
| TS REQ_004 | An AI Mobile Device supporting 2D facial biometric system SHALL support the biometric KPI requirement TS47_3.4.1_REQ_004.1 for each of the use cases: Device Unlock, Application Login and Payment Authorization. |
| TS REQ_004.1 | 2D Facial FAR <= (0.002)% and FRR <= (3)% simultaneously |
| TS REQ_005 | An AI Mobile Device supporting 3D facial biometric system SHALL support the biometric KPI requirement TS47_3.4.1_REQ_005.1 for each of the use cases: Device Unlock, Application Login and Payment Authorization. |
| TS REQ_005.1 | 3D Facial FAR <= (0.001)% and FRR <= (3)% simultaneously. |
| TS REQ_006 | An AI Mobile Device supporting fingerprint biometric system SHALL support the biometric KPI requirement TS47_3.4.1_REQ_006.1 for each of the use cases: Device Unlock, Application Login and Payment Authorization. |
| TS REQ_006.1 | Fingerprint FAR <= (0.002)% and FRR <= (3)% simultaneously. |
| TS REQ_007 | The biometric key performance indicators (KPIs) for the supported biometric system SHOULD be certified by one or more of the following programs:<br>Fast IDentity Online (FIDO) Alliance Biometric Component Certification Program.<br>Internet Finance Authentication Alliance (IFAA) biometric Certification Program. |

Release 19 3GPP V19.1.0 (2023-09)

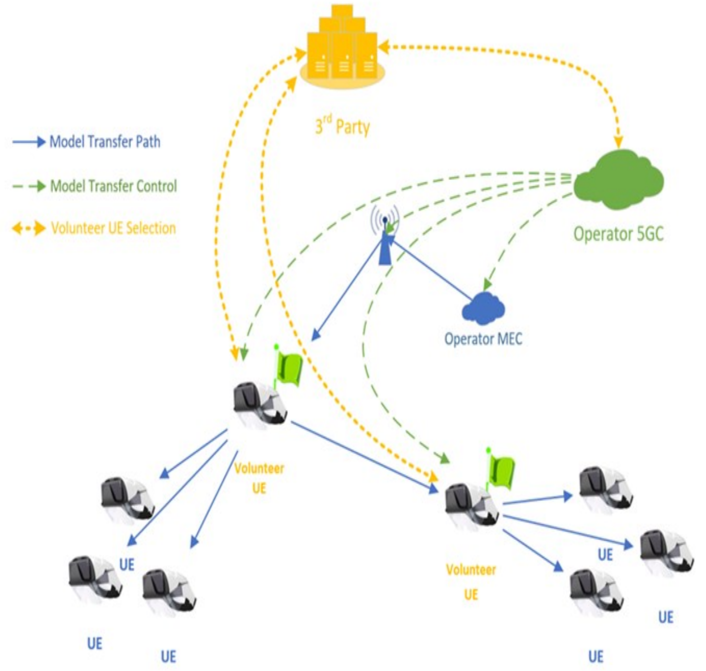



**Figure: 5G System Architecture AI/ML Model Management through Direct Device Connection**

## The Requirements of AI ML Mobile Device - *AI ML Device Application Requirements*

*AI Applications may include, but are not limited to:*

- *Image Processing,*

### *On-Device Image Processing Requirements*

| | | |
|---|---|---|
| TS | REQ_001 | An AI Mobile Device SHOULD have optical character recognition (OCR) capability on the device. |
| TS | REQ_002 | An AI Mobile Device SHOULD have image detection, image classification and image segmentation capabilities on the device. |
| TS | REQ_003 | An AI Mobile Device SHOULD have face detection and face clustering capabilities within a group of photos on the device. |
| TS | REQ_004 | An AI Mobile Device SHOULD have video super-resolution capabilities on the device. |
| TS | REQ_005 | An AI Mobile Device SHOULD have video classification capabilities on the device. |

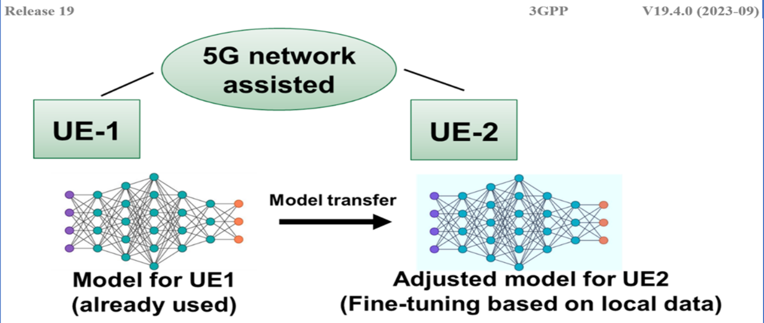


Figure: 5G Network-assisted AI/ML Model Transfer Learning from Source to Target UE

### On-Device Image Processing Applications Requirements

**On-Device Image Processing Applications**

| | | |
|---|---|---|
| TS | REQ_001 | The AI Mobile Device SHOULD support photo scene detection and recognition where the User has the ability to consent to their use. |
| TS | REQ_001.1 | If REQ_001 is supported then the AI Mobile Device SHALL support Identification of one or more objects in different scenes such as portraits, landscapes, foods, night scenes and texts, etc. |
| TS | REQ_001.2 | If REQ_001 is supported then the AI Mobile Device SHALL support Scene detection capabilities to optimize camera settings for image capture based on scene content. |
| TS | REQ_002 | The AI Mobile Device SHOULD support text detection and recognition of installed language packages, where the User has the ability to consent to the text detection and recognition use. |
| TS | REQ_003 | The AI Mobile Device SHOULD support automatic language detection. |
| TS | REQ_004 | The AI Mobile Device SHOULD provide personalized FPE for Users based on gender, age, and skin tone. |
| TS | REQ_005 | The AI Mobile Device SHOULD support FPE of multiple people in a single photo. |
| TS | REQ_006 | The FPE functionality SHOULD be switched off by default and the AI Mobile Device SHOULD support User adjustment of the FPE level from no enhancement to the max FPE. |
| TS | REQ_007 | The AI Mobile Device SHOULD support automatic classification of photos in an album by different categories. |

**Note**: *FPE Functionality is recommended to be automatically off by default in order to give the User the choice of whether to turn this Feature on. This is in recognition of Mental Health & Ethical Concerns*.

# 2. AI ML Mobile Device (selected) Requirements Specification - 7

The Requirements of AI ML Mobile Device - *AI ML Device Application Requirements*

*AI Applications may include, but are not limited to: - AI ML Device Speech Requirements,*

AI ML Device Speech Requirements for Speech ability include such functions as:

*- Voice Recognition, Text to Speech, Voice Activation etc.*

| | | |
|---|---|---|
| TS | REQ_001 | The AI Mobile Device SHOULD have speech ability. |
| TS | REQ_002 | The AI Mobile Device SHOULD support Automatic speech recognition (ASR) capabilities where the User has the ability to consent to ASR. |
| TS | REQ_003 | The AI Mobile Device SHOULD support Natural Language Understanding (NLU) capabilities where the User has the ability to consent to NLU. |
| TS | REQ_004 | The AI Mobile Device SHOULD support Synthesized Voice (Text-To-Speech (TTS) capabilities where the User has the ability to consent to TTS. |
| TS | REQ_005 | If the AI Mobile Device supports Voice Assistant then the requirements SHALL apply. |

**Voice assistant**

| | | |
|---|---|---|
| TS | REQ_001 | AI Mobile Device SHALL support the following functions.<br>Automatic speech recognition (ASR) capabilities.<br>Natural Language Understanding (NLU) capabilities.<br>Synthesized Voice (Text-To-Speech (TTS)) capabilities. |
| TS | REQ_002 | The AI Mobile Device SHALL support voice trigger, and its specific requirements are listed in the following sub requirements: |
| TS | REQ_002.1 | The AI Mobile Device SHOULD support voiceprint recognition for preventing people other than the device's owner from triggering voice assistant. |
| TS | REQ_002.2 | In a quiet environment, the following SHALL be required:<br>The true acceptance rate (TAR) >= (90)%, and the false acceptance rate (FAR) of voiceprint recognition <= (20)%. |
| TS | REQ_002.3 | In a noisy environment, the following SHALL be required:<br>TAR >=(80)%, and FAR of voiceprint recognition <= (20)%. |
| TS | REQ_003 | The AI Mobile Device SHALL have on-device speech recognition library (i.e. with no access to the Internet) for changing the system |
| | | setting (e.g. Turn Bluetooth on/off via voice assistant) and invoking the native applications (e.g. send SMS via voice assistant). |
| TS | REQ_004 | The AI Mobile Device SHOULD have access to different categories of applications and invoke these applications' services and functions via voice assistant. |
| TS | REQ_005 | The AI Mobile Device SHALL support information search by on-device voice assistant. |
| TS | REQ_006 | The AI Mobile Device SHOULD support interaction with smart devices (e.g. home appliances) via voice assistant. |

The Requirements of AI ML Mobile Device - *AI ML Device Application Requirements*

*AI Applications may include, but are not limited to:*

*AI ML Device Augmented Reality (AR) Requirements*

**Augmented Reality (AR)**

| | | |
|---|---|---|
| TS | REQ_001 | The AI Mobile Device SHOULD provide the following AI capabilities for AR native and third-party applications:<br>1. Hand gesture recognition.<br>2. Hand skeleton tracking.<br>3. Human body pose recognition.<br>4. Human body skeleton tracking. |
| TS | REQ_002 | The AI Mobile Device SHOULD support the following applications:<br>1. AR Emoji<br>a. Creating customized AR-based Emoji.<br>b. Tracking User's facial movement and expression and render these on the AR-based Emoji.<br>2. AR video<br>a. Compositing real objects with virtual objects and/or virtual background.<br>b. Minimum (30) fps frame rate.<br>c. AR shadow effect and occlusion handling.<br>d. AR enhanced information text labels should not deviate or disappear from the actual target scene when the AI Mobile Device moves. |

Release 19 3GPP V19.4.0 (2023-09)

**Figure: 5G Network Service Requirements KPI Table for additional High Data Rate and Low Latency AR/VR and Cloud/Edge/Split Data Rendering Service**

| Use Cases | Characteristic parameter (KPI) | | | Influence quantity | | |
|---|---|---|---|---|---|---|
| | Max allowed end-to-end latency | Service bit rate: user-experienced data rate | Reliability | # of UEs | UE Speed | Service Area (note 2) |
| Cloud/Edge/Split Rendering (note 1) | 5 ms (i.e. UL+DL between UE and the interface to data network) (note 4) | 0,1 to [1] Gbit/s supporting visual content (e.g. VR based or high definition video) with 4K, 8K resolution and up to120 frames per second content. | 99,99 % in uplink and 99,9 % in downlink (note 4) | - | Stationary or Pedestrian (note 7) | Countrywide |
| Gaming or Interactive Data Exchanging (note 3) | 10ms (note 4) | 0,1 to [1] Gbit/s supporting visual content (e.g. VR based or high definition video) with 4K, 8K resolution and up to120 frames per second content. | 99,99 % (note 4) | ≤ [10] | Stationary or Pedestrian (note 7) | 20 m x 10 m; in one vehicle (up to 120 km/h) and in one train (up to 500 km/h) |
| Consumption of AR/VR content via tethered AR/VR headset (note 6) | [5 to 10] ms (note 5) | 0,1 to [10] Gbit/s (note 5) | [99,99 %] | - | Stationary or Pedestrian | – |

NOTE 1: Unless otherwise specified, all communication via wireless link is between UEs and network node (UE to network node and/or network node to UE) rather than direct wireless links (UE to UE).
NOTE 2: Length x width (x height).
NOTE 3: Communication includes direct wireless links (UE to UE).
NOTE 4: Latency and reliability KPIs can vary based on specific use case/architecture, e.g. for cloud/edge/split rendering, and can be represented by a range of values.
NOTE 5: The decoding capability in the VR headset and the encoding/decoding complexity/time of the stream will set the required bit rate and latency over the direct wireless link between the tethered VR headset and its connected UE, bit rate from 100 Mbit/s to [10] Gbit/s and latency from 5 ms to 10 ms.
NOTE 6: The performance requirement is valid for the direct wireless link between the tethered VR headset and its connected UE.
NOTE 7: Similar user-experienced data rates may be achievable also at higher UE speeds. [50]

# 2. AI ML Mobile Device (selected) Requirements Specification - 9

The Requirements of AI ML Mobile Device - *AI ML Device Application Requirements*

*AI Applications may include, but are not limited to:*

*- AI ML Device System Optimization Categories Requirements*

**System Optimization**

| | | |
|---|---|---|
| TS | REQ_001 | Only with the explicit permission of the User in order to respect the User's right to privacy around their habits: the AI Mobile Device SHOULD support dynamic system resource allocation and optimization based on feedback provided by on-device sensors measuring environmental conditions combined with continuous learning of User habits and behaviours or device or network usage or performance indicators:<br>1. Dynamic application management (e.g. pre-loading, closing, put to sleep, control network access) based on User's habits (e.g. usage duration, frequency).<br>2. Dynamic application management based on abnormal behaviour detection (e.g. increased memory usage, abnormal power consumption, self-starting in the background). |
| | | 3. Dynamic system resource management based on continuous learning of system performance (e.g. memory and storage defragmentation, off-line storage during off-peak periods).<br>4. Dynamic system resource allocation for high performance applications (e.g., gaming and video). |

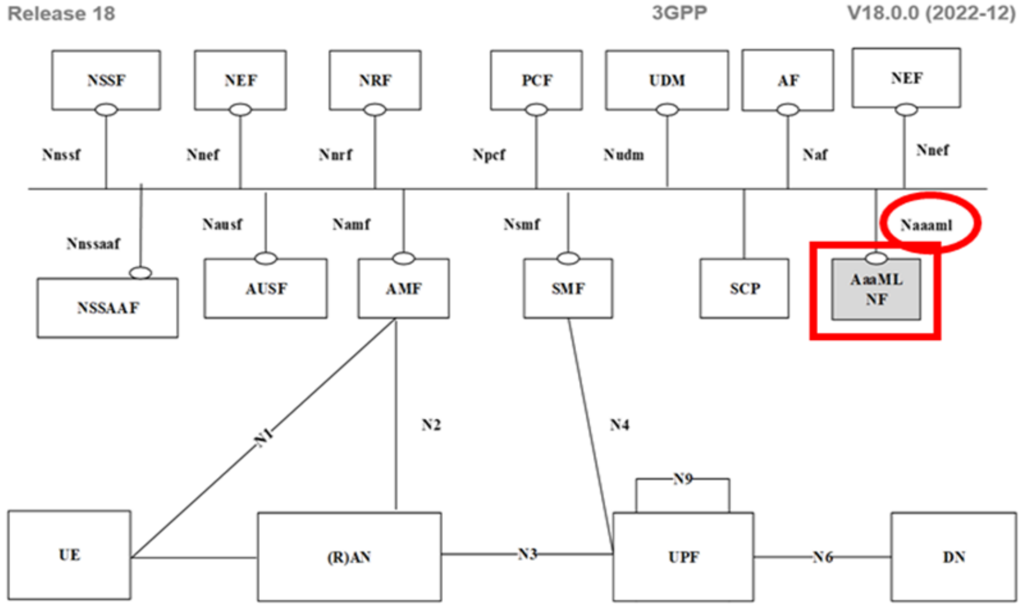



**Figure: 5G System Service Architecture with AaaML NF**

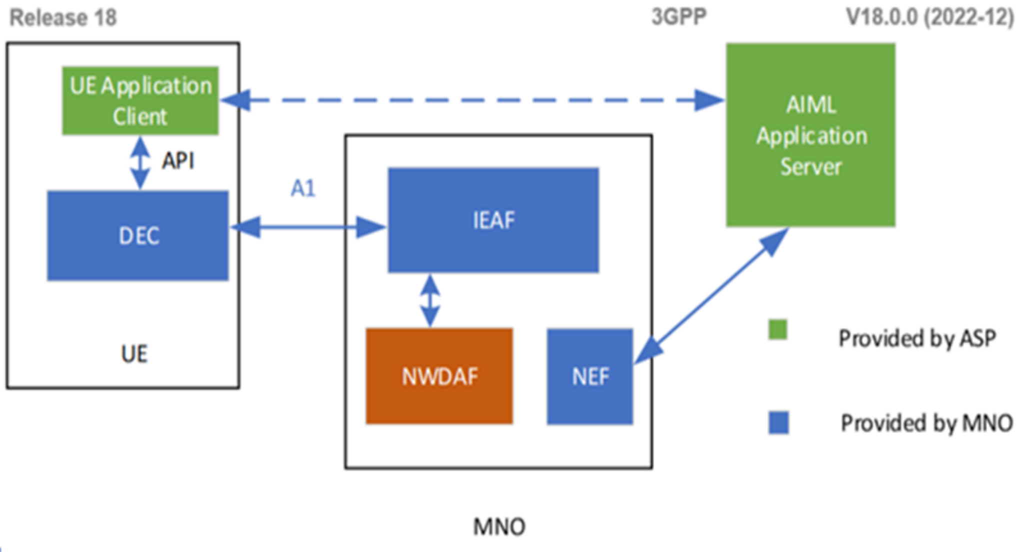



**Figure: 5G IEAF (Data Information AF)**

# 2. AI ML Mobile Device (selected) Requirements Specification - 10

*AI ML Device Privacy and Security Requirements*

Applicable Law(s) and Regulations as related to Privacy and Data Protection must be complied with in connection with AI on Mobile Device.

For avoidance of doubt, where Laws are not in place in certain jurisdictions, *Manufacturers should respect the User and not leave AI Functionality 'on' by default.*

## *It should be 'Private by Design and by Default'.*

Any choice to turn off Functionality by the User must be fully respected and Techniques, such as 'Dark Patterns', that seek to manipulate a User's Free Choice should be avoided.

**Privacy Requirements**

| | | |
|---|---|---|
| TS | REQ_001 | AI on mobile device SHOULD comply with the privacy laws in the country where the device is commercially retailed. |
| TS | REQ_002 | Appropriate technical and organisational safeguards SHOULD be implemented to ensure that, by default, only the personal data reasonably necessary for a specific purpose are processed. |
| TS | REQ_003 | AI Applications that process Personal Data SHALL be off by default unless processing exclusively takes place locally on the device. |
| TS | REQ_003.1 | The User SHOULD be allowed to control whether individual AI applications are switched on. |
| TS | REQ_003.2 | The User SHOULD be allowed to control whether individual AI applications are switched off. |
| TS | REQ_004 | The AI Application on the AI Mobile Device SHALL be designed in such a way that a Data Processor will have the responsibility to:<br>1) Be transparent with the User on the nature of the input data used in the AI processing (e.g. personal files, biometrics, ...).<br>2) Forbid transferring personal data processing off the device except if the User has explicitly agreed or other legal basis has been satisfied in accordance with the law. |
| | | 3) Forbid transferring results of on-device AI processing containing personal data off the device except if the User has explicitly agreed or other legal basis has been satisfied in accordance with the law. |

*1. See, e.g., GDPR Article 35(1) requiring a Data Protection Impact Assessment where the processing, "is likely to result in a high risk to the rights and freedoms of natural persons", and GDPR Article 34(2) requiring a notice of Personal Data breach where it, "is likely to result in a high risk to the rights and freedoms of natural persons."*

Release 18 3GPP V18.0.0 (2023-09)

**Table: 5G System Reference Architecture for Data Collection and Reporting Methods invoked by the UE Application on the Direct Data Collection Client**

| Method name | Type | Description |
|---|---|---|
| *registerUeApplication* | State change | UE Application registers with the Direct Data Collection Client, including a callback listener for receiving event notifications. |
| *deregisterUeApplication* | State change | UE Application deregisters with the Direct Data Collection Client. |
| *setUserConsent* | | UE Application grants permission for the Direct Data Reporting Client to include the GPSI when creating Data Reporting Sessions. |
| *getDataCollectionAndReportingConfiguration* | Configuration request | UE Application obtains its UE data collection and reporting configuration from the Direct Data Collection Client. |
| *reportUeData* | Data report | UE Application reports collected UE data to the Direct Data Collection Client according to its configuration.<br>The UE Application may indicate (by setting a Boolean method parameter to *true*) that the data report includes UE data requiring expedited processing by the Direct Data Collection Client and, consequently, by the Data Collection AF. |
| *resetClientReportingIdentifier* | | UE Application requests that the Direct Data Collection Client generates a new opaque client reporting identifier for use in data reporting until further notice.<br>This requires any existing Data Reporting Session to be destroyed and a new one (including the replacement client reporting identifier) to be created. |
| *uEApplicationBusy* | Notification | UE Application notifies the Direct Data Collection Client that it is temporarily unable to perform UE data collection and reporting due to a busy or stalled condition. |
| *impendingUeApplicationFailure* | Notification | UE Application notifies the Direct Data Collection Client of an impending fatal error condition that will cause abrupt shutdown of the UE Application. |

Release 18 3GPP V18.1.0 (2023–09)

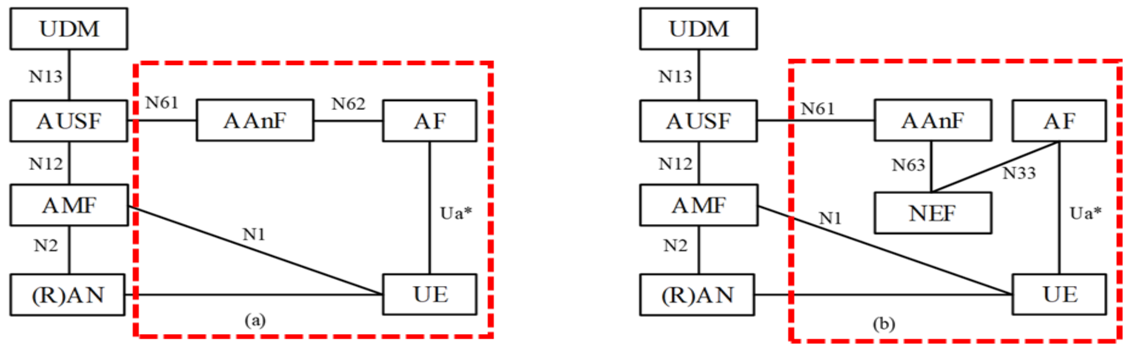



**Figure: 5G System Architecture Fundamental Network Model for Authentication and Key Management for Applications (AKMA) Architecture in Reference Point representation for (a) Internal HPLMN AFs and (b) External AFs**

# 2. AI ML Mobile Device (selected) Requirements Specification - 11

*AI ML Device Security Requirements*

Applicable Law(s) and Regulations as related to Security and Data Protection must be complied with in connection with AI on Mobile Device.

For avoidance of doubt, where Laws are not in place in certain Jurisdictions, manufacturers should respect the User and not leave AI Functionality ‘on’ by default. From a Security perspective this also follows the ‘Principle of least privilege’, ensuring that Systems have no more access than is necessary, as a default starting point.

The AI Mobile Device needs to operate as ‘Secure by Default’. Any choice to turn off functionality by the User must be fully respected and techniques, such as ‘Dark Patterns’ that seek to manipulate a User’s free choice should be avoided. This assists in retaining User Trust and helps prevent subversion by malicious actors.

## Security Requirements

| | | |
|---|---|---|
| TS | REQ_001 | The AI Mobile Device SHALL use reasonable safeguards appropriate to the sensitivity, confidentiality and integrity of the information. |
| TS | REQ_002 | Except as required or permitted by applicable law, the User SHALL always remain in control of the collection of their personal data and its usage, in order to minimise the risk of malicious usage or data leakage. |
| TS | REQ_003 | Off ‘toggle’ switches SHALL turn off the functionality, except as permitted or required by applicable law. |
| TS | REQ_004 | Techniques, such as ‘Dark Patterns’, that manipulate the User’s choice SHALL NOT be used. |

## Security for AI Applications

| | | |
|---|---|---|
| TS | REQ_001 | The AI models used by an AI Mobile Device SHOULD be secure and robust, and be protected with appropriate safeguards to prevent and to mitigate attacks. |
| TS | REQ_002 | Defence techniques SHOULD be employed to protect the training data for protecting models. For example, in evasion attacks, data can be manipulated to mislead AI models. |
| TS | REQ_003 | Autonomous AI Mobile Device operations SHALL be controlled, and/or authorized by the authenticated User. |
| TS | REQ_004 | AI Mobile Device operations SHOULD be performed in the Secured Environment [4], e.g. a secure boot and upgrade is enforced, and the system integrity is protected. |
| TS | REQ_005 | Data and metadata for AI Mobile Device SHALL be stored with encryption with keys that are stored securely in a Secured Environment, e.g. Trusted Execution Environment (TEE) [4]. |
| TS | REQ_006 | Biometric Data, which are processed by an AI Application (e.g. templates) used for authentication within the AI Mobile Device, SHALL NOT be transferred off the device. |

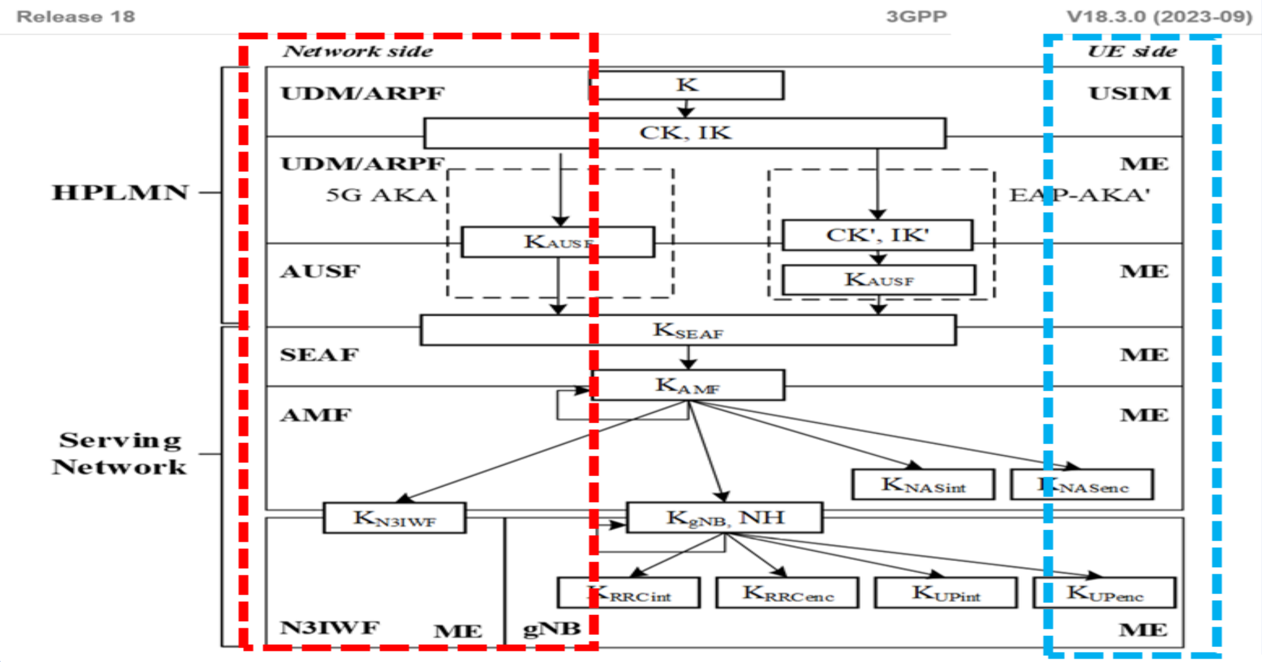



**Figure: 5G System Security Architecture Key Hierarchy Generation**

2. AI ML Mobile Device (selected) Requirements Specification - 12

AI ML Mobile Device SDK & API Requirements

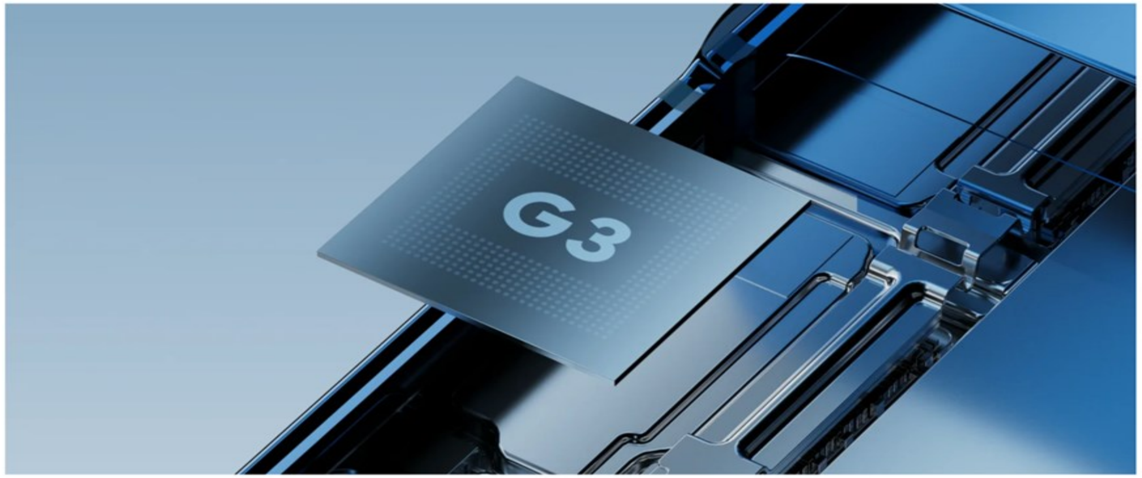
**Currently, each Chipset Vendor has its own Set of APIs, which leads to a Fragmented Ecosystem.**

***Standardizing and Unifying Application APIs is very Necessary and Highly Recommended.***

***Google - Android Neural Networks API (NNAPI)***

The *Android Neural Networks API (NNAPI)* is *an Android C API designed* for running computationally intensive operations for Machine Learning (ML) on Mobile Devices.

NNAPI is designed to provide a base Layer of Functionality for Higher-Level Machine Learning (ML) Frameworks (such as TensorFlow Lite, Caffe2, or others) that Build and train neural networks.




Our third-generation Google Tensor G3 chip brings the latest in AI to the Pixel 8 and Pixel 8 Pro.

Release 18 3GPP V18.1.0 (2023-09)

## OpenAPI definition of the AI/ML NRM

OpenAPI document AiMlNrm.yaml

```yaml
openapi: 3.0.1
info:
  title: AI/ML NRM
  version: 18.1.0
  description: >-
    OAS 3.0.1 specification of the AI/ML NRM
    © 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
    All rights reserved.
externalDocs:
  description: 3GPP          AI/ML Management
```

Release 18 3GPP V18.1.0 (2023-09)

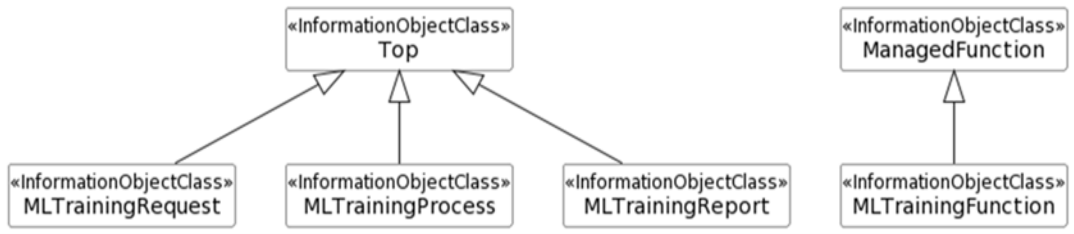



**Figure: 5G System AI/ML Management Inheritance Hierarchy for ML Training related NRM**

Release 18 3GPP V18.1.0 (2023-09)

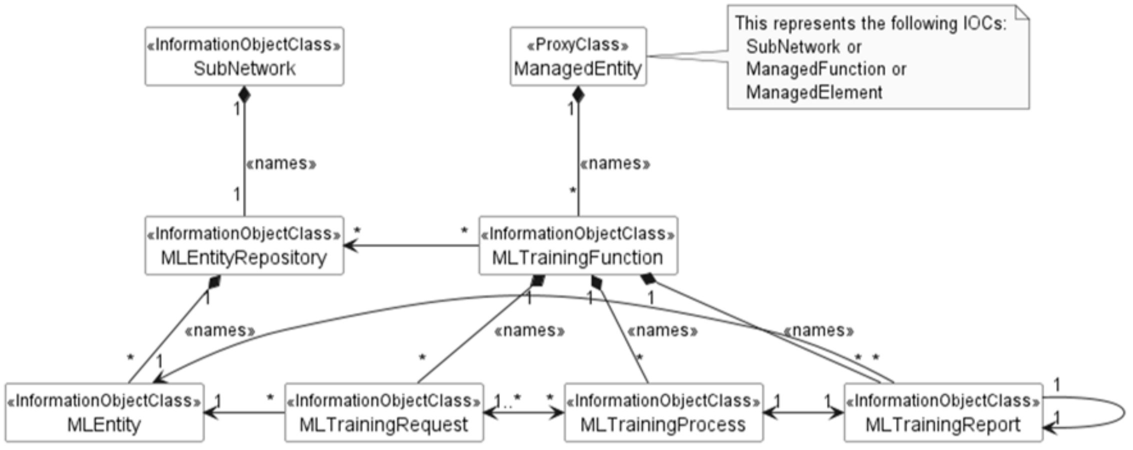



**Figure: 5G System AI/ML Management NRM fragment for ML Training**

## AI ML Mobile Device SDK & API Requirements

### ***Qualcomm - Snapdragon Neural Processing Engine (SNPE)***

The *Snapdragon Neural Processing Engine (**SNPE**)* is a Qualcomm Snapdragon SW accelerated runtime for the execution of Deep Neural Networks (DNN). The Qualcomm Neural Processing SDK for AI ML is designed to help developers run one (1) or more Neural Network Models trained in Caffe/Caffe2, ONNX, or TensorFlow on Snapdragon Mobile Platforms, whether that is the CPU, GPU or DSP.



Release 18 3GPP V18.1.0 (2023-09)

OpenAPI definition of the AI/ML NRM

OpenAPI document AiMlNrm.yaml


```
openapi: 3.0.1
info:
  title: AI/ML NRM
  version: 18.1.0
  description: >-
    OAS 3.0.1 specification of the AI/ML NRM
    © 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
    All rights reserved.
externalDocs:
  description: 3GPP AI/ML Management
```


Release 18 3GPP V18.1.0 (2023-09)

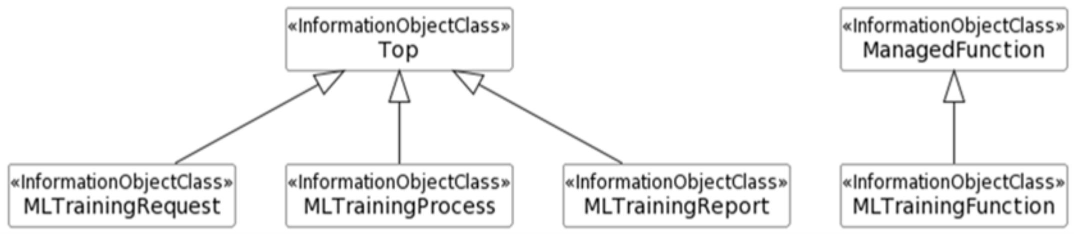



**Figure: 5G System AI/ML Management Inheritance Hierarchy for ML Training related NRM**

Release 18 3GPP V18.1.0 (2023-09)

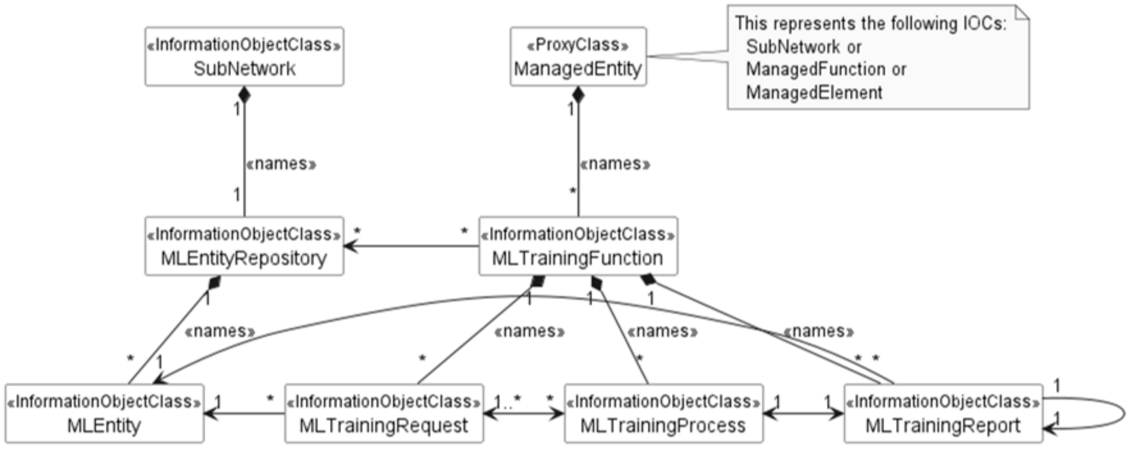



**Figure: 5G System AI/ML Management NRM fragment for ML Training**

## AI ML Mobile Device SDK & API Requirements

### ***MediaTek NeuroPilot***

It embraces the advantages of 'Edge AI', which means the AI Processing is done On-Device rather than relying on a fast Internet Connection and Cloud Service. However, NeuroPilot doesn't have to use a dedicated AI Processor. Its SW can intelligently detect what Compute Resources are available, between CPU, GPU and APU, and automatically choose the best one. MediaTek's Next-Generation APU Architecture incorporates a HW Generative AI Engine, enabling faster and safer edge AI Computing.

The 6th generation MediaTek APU has Maximum Effective Performance initiative with a new 'Dual-Mode' design that can enact an eXtreme Power Savings mode specifically for AI-Noise Reduction (AI-NR) in photography and videography, and AI-Super Resolution (AI-SR) used in AI-Camera, AI-GPU or AI-Video activities.


Source: Mediatek

Release 18 3GPP V18.1.0 (2023-09)

### OpenAPI definition of the AI/ML NRM

OpenAPI document AiMlNrm.yaml

```
openapi: 3.0.1
info:
  title: AI/ML NRM
  version: 18.1.0
  description: >-
    OAS 3.0.1 specification of the AI/ML NRM
    © 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
    All rights reserved.
externalDocs:
  description: 3GPP           AI/ML Management
```

Release 18 3GPP V18.1.0 (2023-09)

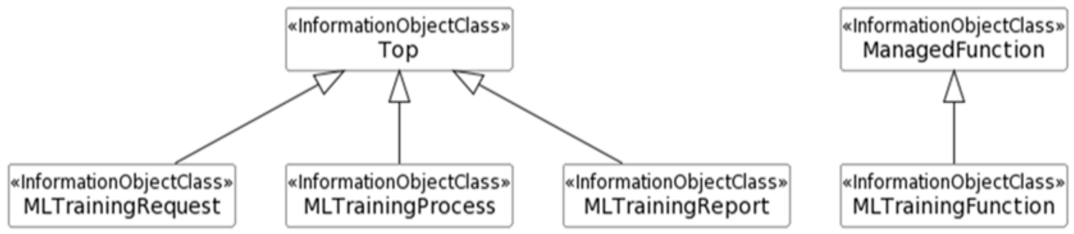



**Figure: 5G System AI/ML Management Inheritance Hierarchy for ML Training related NRM**

Release 18 3GPP V18.1.0 (2023-09)

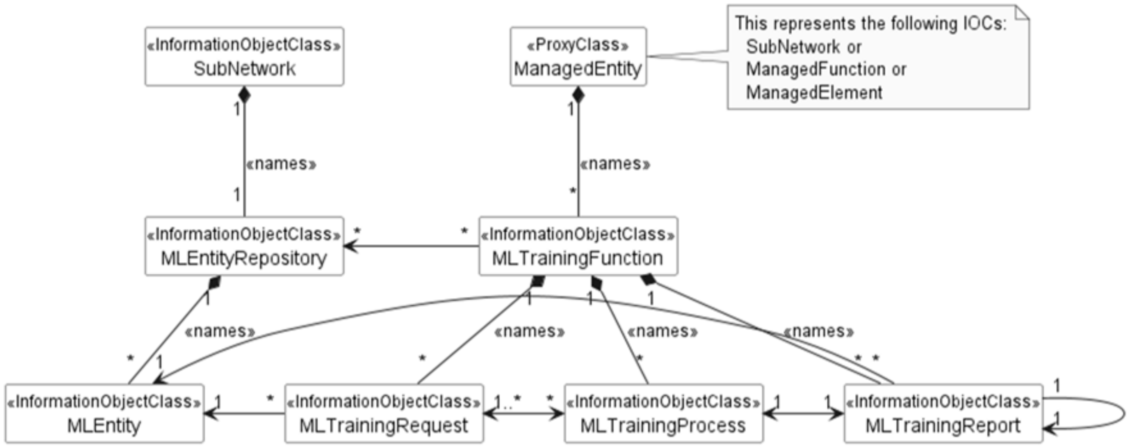



**Figure: 5G System AI/ML Management NRM fragment for ML Training**

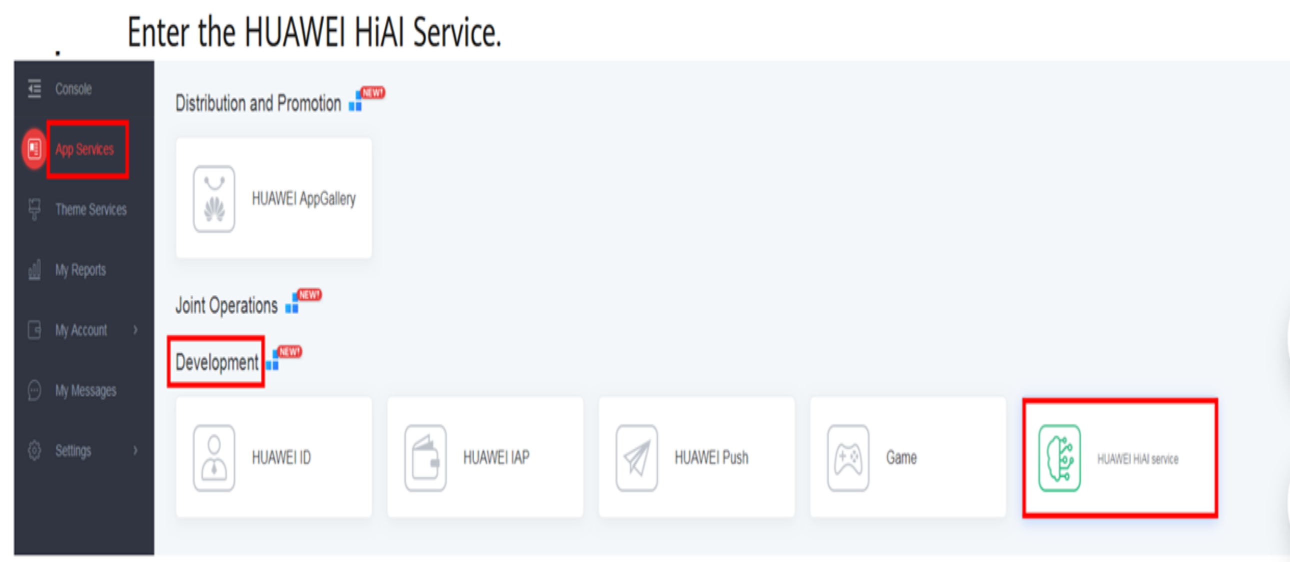
## AI ML Mobile Device SDK & API Requirements

***Huawei* - HiAI**

HiAI is a Mobile Terminal–oriented artificial AI Computing Platform that constructs three (3) Layers of Ecology:

1. Service Capability Openness,
2. Application Capability Openness, and
3. Chip Capability Openness

The three (3) - Layer Open Platform that integrates Terminals, Chips, and the Cloud brings more extraordinary experiences for Users and Developers.




Release 18 3GPP V18.1.0 (2023-09)

### OpenAPI definition of the AI/ML NRM

OpenAPI document AiMlNrm.yaml


```
openapi: 3.0.1
info:
  title: AI/ML NRM
  version: 18.1.0
  description: >-
    OAS 3.0.1 specification of the AI/ML NRM
    © 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
    All rights reserved.
externalDocs:
  description: 3GPP AI/ML Management
```


Release 18 3GPP V18.1.0 (2023-09)

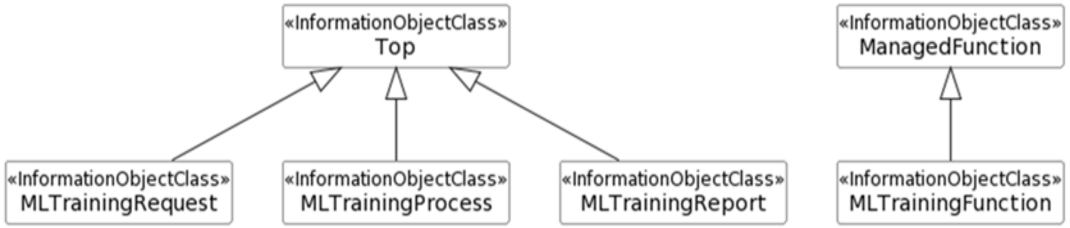



**Figure: 5G System AI/ML Management Inheritance Hierarchy for ML Training related NRM**

Release 18 3GPP V18.1.0 (2023-09)

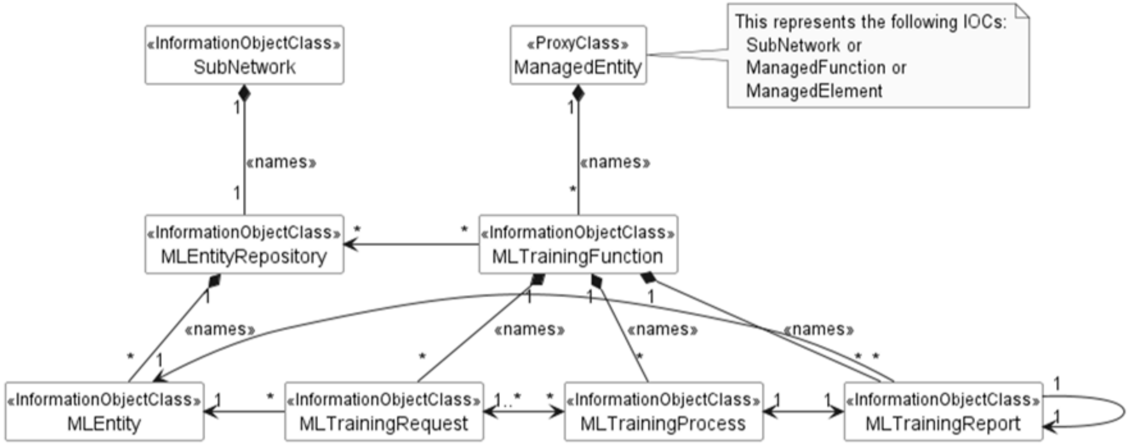



**Figure: 5G System AI/ML Management NRM fragment for ML Training**

## AI ML Mobile Device SDK & API Requirements

### ***Apple - Core ML***

Core ML is an Apple framework that allows developers to easily integrate machine learning (ML) models into apps. Core ML is available on iOS, watchOS, macOS, and tvOS. Core ML introduces a public file format (.mlmodel) for a broad set of ML methods including deep neural networks (convolutional and recurrent), tree ensembles (boosted trees, random forest, decision trees), and generalized linear models.

# Core ML

Use Core ML to integrate machine learning models into your app. Core ML provides a unified representation for all models. Your app uses Core ML APIs and user data to make predictions, and to train or fine-tune models, all on a person's device.


Release 18 3GPP V18.1.0 (2023-09)

## OpenAPI definition of the AI/ML NRM

OpenAPI document AiMlNrm.yaml


```
openapi: 3.0.1
info:
  title: AI/ML NRM
  version: 18.1.0
  description: >-
    OAS 3.0.1 specification of the AI/ML NRM
    © 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
    All rights reserved.
externalDocs:
  description: 3GPP           AI/ML Management
```


Release 18 3GPP V18.1.0 (2023-09)

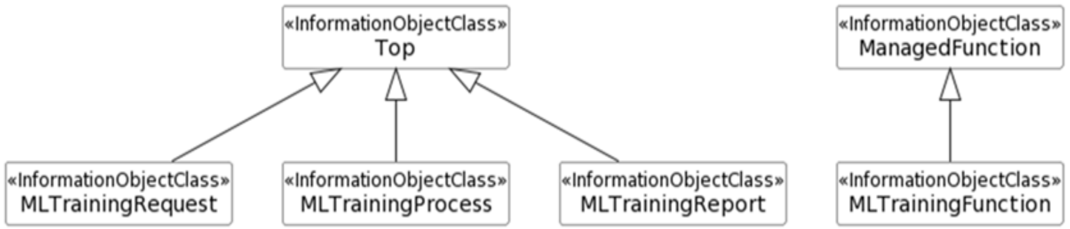



**Figure: 5G System AI/ML Management Inheritance Hierarchy for ML Training related NRM**

Release 18 3GPP V18.1.0 (2023-09)

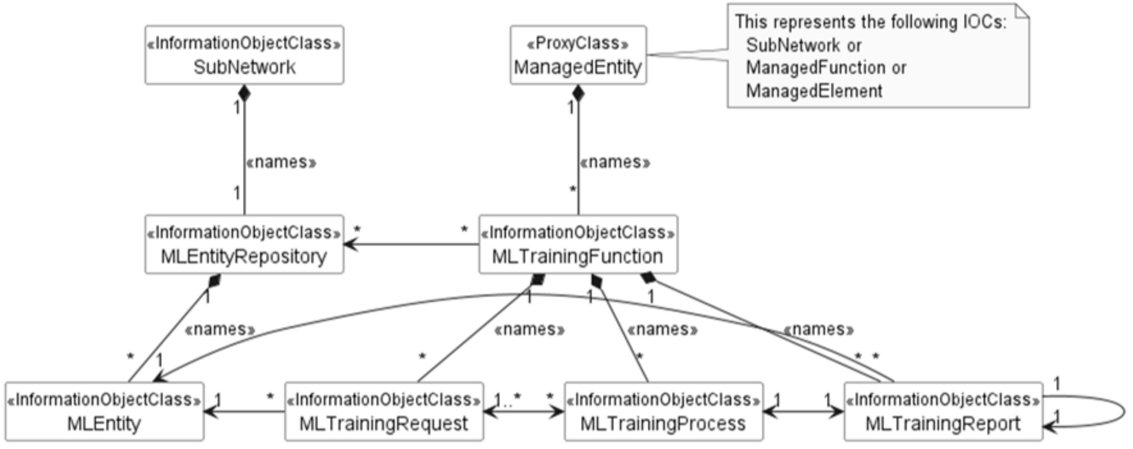



**Figure: 5G System AI/ML Management NRM fragment for MLTraining**

## AI ML Mobile Device SDK & API Requirements

***Mobile AI Compute Engine (MACE)*** is a Deep Learning (DL) Inference Framework optimized for mobile heterogeneous computing on Android, iOS, Linux and Windows Devices. The Design focuses on the following targets:

*1. Performance:* Runtime is optimized with NEON, OpenCL and Hexagon, and Winograd algorithm is introduced to speed up convolution operations. The initialization is also optimized to be faster.

*2. Power consumption:* Chip dependent power options like big.LITTLE scheduling, Adreno GPU hints are included as advanced APIs.

*3. Responsiveness:* UI Responsiveness Guarantee is sometimes obligatory when running a Model. Mechanism like automatically breaking OpenCL Kernel into small units is introduced to allow better pre-emption for the UI Rendering Task.

*4. Memory Usage and Library Footprint:* Graph Level Memory Allocation Optimization and Buffer re-use are supported. The Core Library tries to keep minimum external dependencies to keep the Library Footprint small.

*5. Model Protection*: Model protection has been *the Highest Priority* since the beginning of the Design. Various Techniques are introduced like Converting Models to C++ Code and literal obfuscations.

*6. Platform Coverage:* Good coverage of recent *Qualcomm, MediaTek, Pinecone and other ARM based Chips*. *CPU Runtime* supports *Android, iOS and Linux*.

*7. Rich Model Formats Support:* TensorFlow, Caffe and ONNX Model Formats are supported.

Release 18 3GPP V18.1.0 (2023-09)

### OpenAPI definition of the AI/ML NRM

OpenAPI document AiMlNrm.yaml


```yaml
openapi: 3.0.1
info:
  title: AI/ML NRM
  version: 18.1.0
  description: >-
    OAS 3.0.1 specification of the AI/ML NRM
    © 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
    All rights reserved.
externalDocs:
  description: 3GPP AI/ML Management
```


Release 18 3GPP V18.1.0 (2023-09)

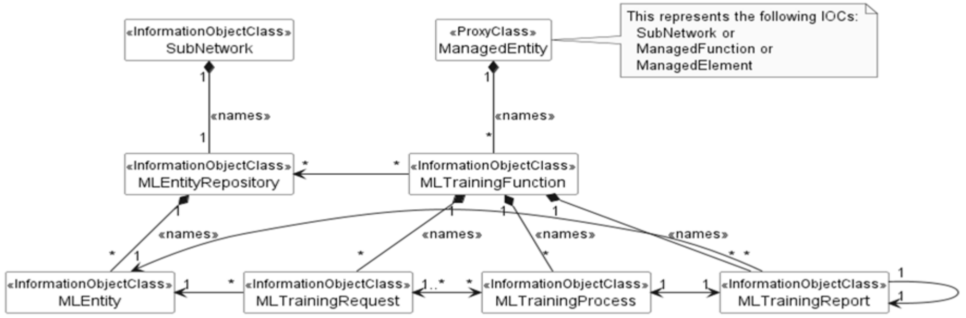
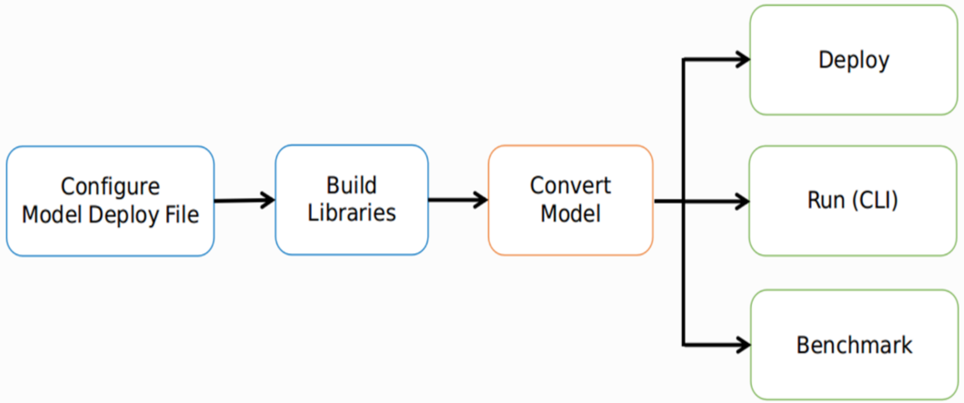



**Figure: 5G System AI/ML Management NRM fragment for MLTraining**

The following figure shows the basic work flow of MACE.



# 3. 5G System use of AI/ML

The AI/ML Techniques and relevant Applications are being increasingly adopted by the wider Industries and proved to be successful. These are now being applied to Telecommunication Industry including Mobile Networks.

Although AI/ML Techniques, in general, are quite mature nowadays, some of the relevant aspects of the Technology are still evolving while New Complementary Techniques are frequently emerging.

The AI/ML Techniques can be generally characterized from different perspectives including the followings:

- ***Learning Methods :*** The Learning Methods include Supervised Learning, Semi-Supervised Learning, Unsupervised Learning and Reinforcement Learning. Each Learning Method fits one (1) or more specific Category of Inference (e.g. Prediction), and requires Specific Type of Training Data. A brief comparison of these learning methods is provided in the Table:

- *Learning complexity:* As per the Learning Complexity, there are Machine Learning (i.e. basic learning) and Deep Learning (DL).

- *Learning Architecture:* Based on the Topology and Location where the Learning Tasks take place, the AI/ML can be categorized to Centralized Learning, Distributed Learning and Federated Learning.

- *Learning Continuity:* From Learning Continuity Perspective, the AI/ML can be off-line Learning or Continual Learning.

Release 18 3GPP V18.1.0 (2023-09)

**Table : Comparison of AI/ML Learning Methods**

| | Supervised learning | Semi-supervised learning | Unsupervised learning | Reinforcement learning |
|---|---|---|---|---|
| Category of inference | Regression (numeric), classification | Regression (numeric), classification | Association, Clustering | Reward-based behaviour |
| Type of training data | Labelled data (Note) | Labelled data (Note), and unlabelled data | Unlabelled data | Not pre-defined |
| NOTE: The labelled data means the input and output parameters are explicitly labelled for each training data example. | | | | |

Artificial Intelligence/Machine Learning (AI/ML) Capabilities are used in various Domains in 5G System, including:

- Management and Orchestration for Data Analytics (MDA)
- 5G Networks Data Analytics (NWDAF)
- NG-RAN, e.g. RAN Intelligence.

The AI/ML-Inference Function in the 5GS uses the ML Model and/or AI Decision Entity for Inference. Each AI/ML Technique, depending on the adopted specific Characteristics, suitable for supporting certain Type/Category of Use Case(s) in 5G System.

To enable and facilitate the AI/ML Capabilities with the suitable AI/ML Techniques in 5GS, the ML Model and AI/ML Inference Function need to be managed.

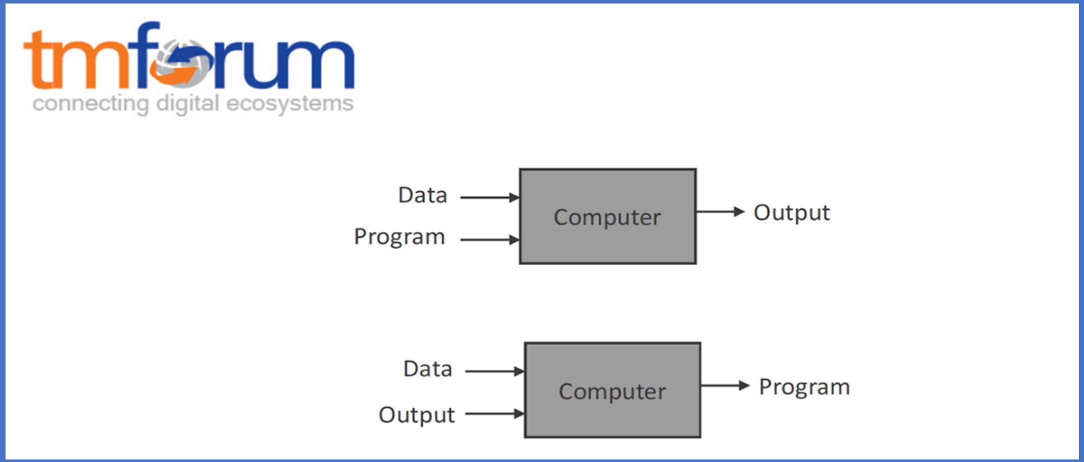
# 3. 5G System use of AI/ML

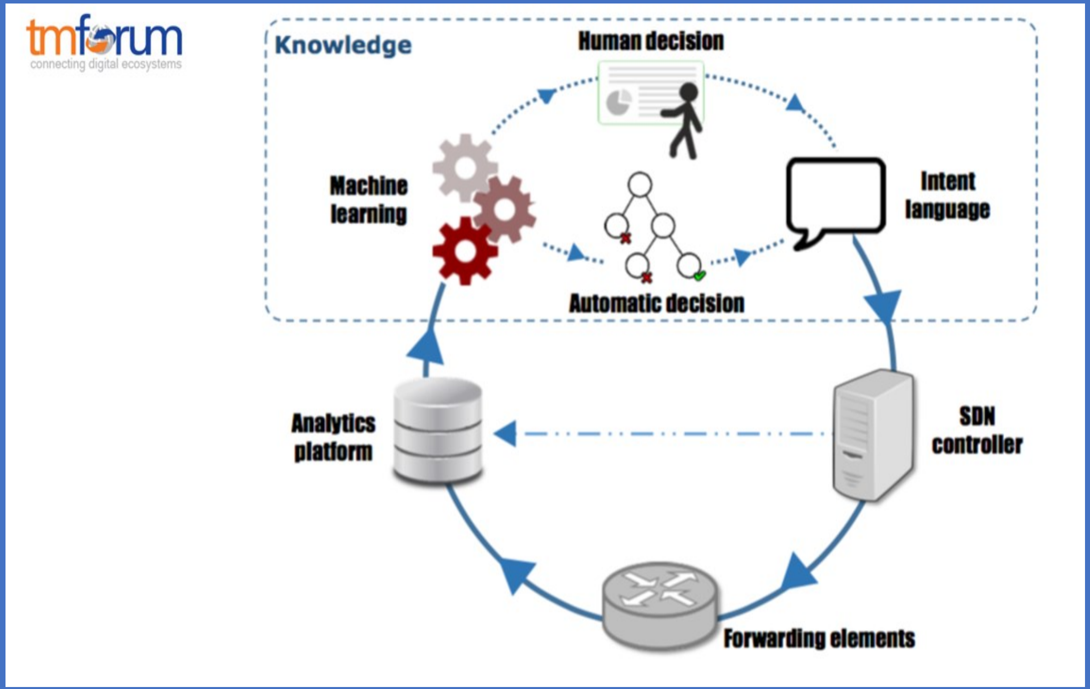
In **Programming**, a Human writes a Computer Program and provides the Data, which the Computer processes to create the Output.

***In Machine Learning (ML),*** Humans provide the Data along with the Desired Output, Rules and Constraints, and the Computer (Algorithms with trained Models) writes the Program to deliver this.

A *Knowledge-defined Network (**KDN**)* operates by means of a Control Loop to provide:
- *Automation,*
- *Recommendation,*
- *Optimization,*
- *Validation and*
- *Estimation.*

CSPs are beginning to *use AI and Machine Learning (ML) in three (3) Key Areas*:
1. Customer Experience Management
2. Service Management and Optimization
3. Network Management and Optimization




***The Knowledge Plane (KP)*** is a distributed & decentralized construct within the Network that
- Gathers,
- Aggregates, and
- Manages

***Information about Network behavior and Operation***, and provides an integrated view to all parties (Operators, Users, and the Network itself). The Goal is to enlarge our view of what constitutes ***the Network to match the intuition of a User,*** and to enhance our ability to manage the network intelligently, without disturbing the open and unknowing forwarding plane (Ref. D.C., KP for I., v4.6 05/03).

# 3. 5G System use of AI/ML

5G System AI/ML Model Transfer

The ***5G System*** can at least support *three (3) types of AI/ML Operations:*

**1.AI/ML Operation splitting between AI/ML (Network) End-points:** The AI/ML Operation/Model is split into Multiple Parts according to the current Task and Environment. The intention is to *off-load the Computation-Intensive, Energy-Intensive Parts to Network End-points*, whereas *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 End-point executes the remaining Parts/Layers and feeds the Inference Results back to the Device.

***2. AI/ML Model/Data Distribution and Sharing over 5G System***: Multi-functional Mobile Terminals might need to switch the AI/ML Model in response to task and environment variations. The condition of adaptive model selection is that the models to be selected are available for the Mobile Device. However, given the fact that the AI/ML Models are becoming increasingly diverse, and with the *limited storage resource in a UE*, it can be determined to *not pre-load all candidate AI/ML Models on-board*. *Online model distribution (i.e. New Model Downloading) is needed*, in which an AI/ML Model can be distributed from a NW end-point 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 needs to be monitored constantly.*

***3. Distributed/Federated Learning (FL) over 5G System:*** The Cloud Server trains a Global Model by aggregating Local Models partially-trained by each End devices. 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 to the Cloud server 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.

In Mobile Communications Systems, Mobile Devices (e.g. Smartphones, Automotive, Robots) are increasingly replacing conventional Algorithms (e.g. Speech Recognition, Image Recognition, Video Processing) with AI/ML Models to enable Applications.

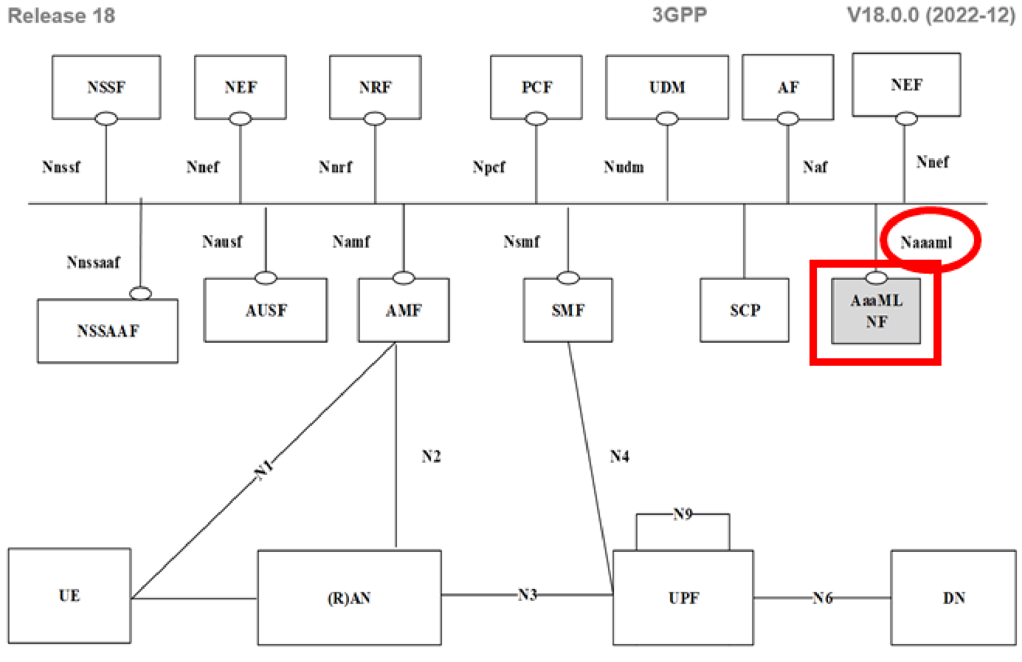



Figure: 5G System Service Architecture with AaaML NF

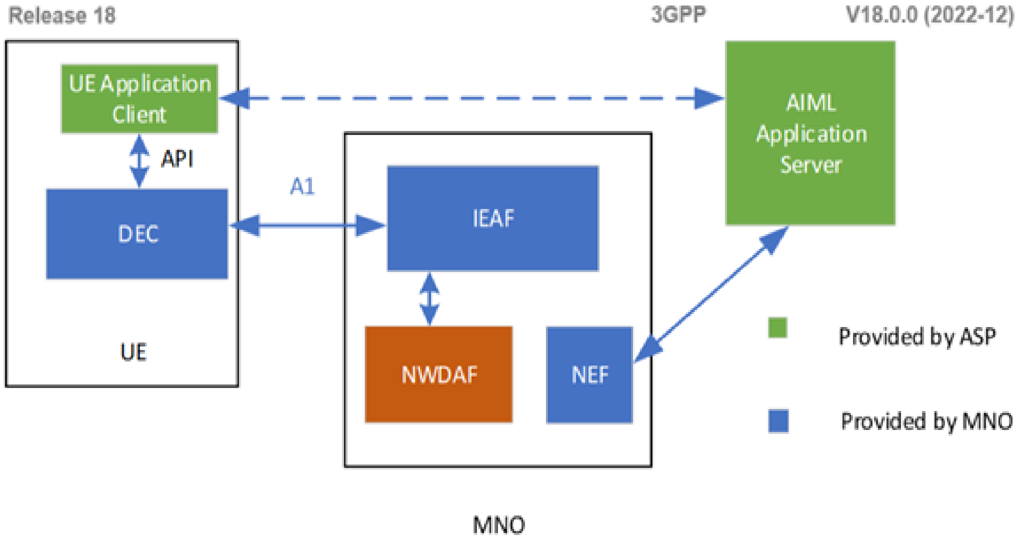



Figure: 5G IEAF (Data Information AF)

# **Split AI/ML Operation between AI/ML End-Points** for *AI Inference by leveraging Direct Device Connection*

Proximity based Work Task Off-loading for AI/ML Inference

*The Model Splitting is the most significant Feature for AI Inference.*

As some ***3GPP 5G R18 UCs*** show, the *Number of Terminal Computing Layers* and the *Amount of Data Transmission* are corresponding to *Different Model Splitting Points.*

For example, as the Figure shows, the General Trend is that the more Layers the UE calculated, the less Intermediate Data needs to be transmitted to Application Server (AS).

In another word, *when UE has Low Computation Capacity* (e.g. due to Low Battery), the *Application can change the Splitting Point* to *let UE calculate fewer Layers while increasing the Data Rate in Uu for transmitting a Higher Load of Intermediate Data to Network.*

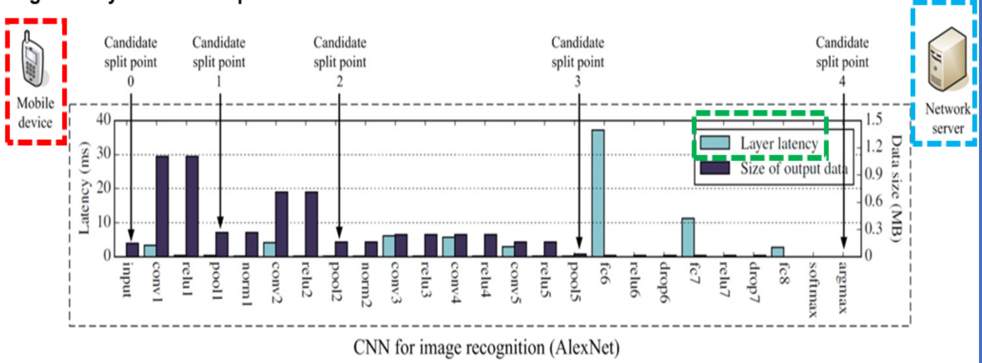
However, sometimes the *Data Rate cannot be increased due to Radio Resource Limitation,* in such circumstances, *UE with Low Computation Capacity* needs to *off-load the Computation Task to a Proximity UE* (likely a ***Relay UE***), but still keeping the Computation Service and let the ***Proximity UE*** to send the Calculated Data to Network. Thus, by *off-loading the Work Task using "Direct" Device Connection,* the *original UE's computation load* will be released while the *Data Rate in Uu interface* will not necessarily be increased either, which leads to a more ideal performance.

A UE uses the AI Model (AlexNet) for image Recognition. As predetermined by Application, *there are five (5) Alternative Splitting Points which are corresponding to intermediate Data Size and Data Rate, while fewer the layers being calculated implies fewer the workload being performed by UE.*

The Specific Values are shown in the Table about Split AI/ML Image Recognition.

Release 19 3GPP V19.1.0 (2023-09)

**Figure: Layer-level Computation/Communication Resource Evaluation for an AlexNet Model**



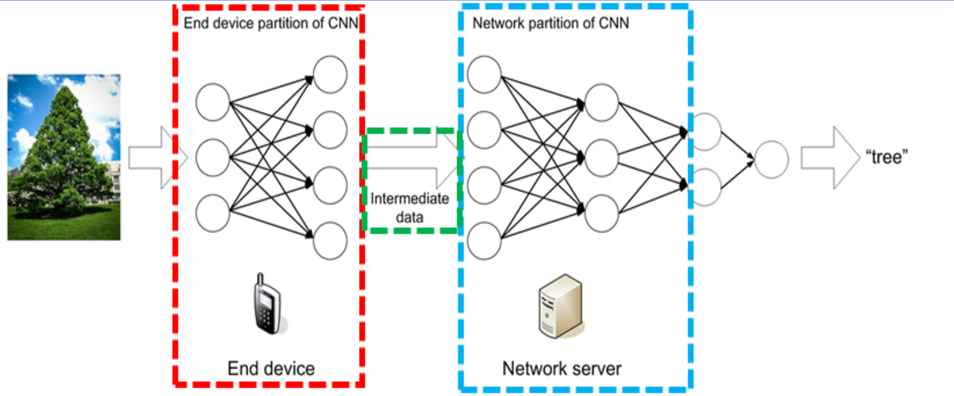




**Figure: Example of split AI/ML Image Recognition**

Release 19 3GPP V19.1.0 (2023-09)

**Table: Required UL Data Rate for different Split Points of AlexNet Model for Video Recognition at 30 Frame Per Second (FPS)**

| Split point | Approximate output data size (MByte) | Required UL data rate (Mbit/s) |
|---|---|---|
| Candidate split point 0 (Cloud-based inference) | 0.15 | 36 |
| Candidate split point 1 (after pool1 layer) | 0.27 | 65 |
| Candidate split point 2 (after pool2 layer) | 0.17 | 41 |
| Candidate split point 3 (after pool5 layer) | 0.02 | 4.8 |
| Candidate split point 4 (Device-based inference) | N/A | N/A |

# 3. 5G System use of AI/ML

As shown on the left side (a) "No Task Off-loading" in the of Figure, UE-A is doing *Image Recognition using AlexNet Model.* The *involved AI/ML End-Points* (e.g. UE, AI/ML Cloud/Edge Server) run Applications providing the capability of AI/ML Model Inference for control task, and support the split control operation. *The 5G System has the ability to provide 5G Network related information to the AI/ML Server.*

It selects "*Splitting Point-3" for the AI Inference.*

*The E2E Service Latency (including Image Recognition Latency and Intermediate Data Transmission Latency) is 1 second.*

When the UE-A's battery becomes low, it cannot afford the heavy work task for the *AlexNet Model (i.e. calculating Layer 1-15 for AlexNet Model in Local side.*

Being managed by 5G Network, the *UE-A discovers UE-B (a Customer Premise Equipment, CPE*) which has installed the same Model and is willing to *take the off.loading task from UE-A*.

**NOTE 1**: *The 5G Network does not store UE-A and UE-B's Location Data.*

Then *UE-A established the side-link (direct device connection) to UE-B*. During the sidelink establishment, the UE-B also gets the information of the total service latency (including the image recognition latency and intermediate data transmission latency) and the processing time consumed by UE-A for computing layer 1-4.

*Since the UE-B has acquired the E2E Service Latency and the processing time consumed by UE-A, and also it knows its own processing time for computing layer 5-15, the UE-B can determine the QoS parameters applied to both Uu and Sidelink while keeping the E2E service latency same as the E2E service latency described in step-1.*

**NOTE 2**: *It is assumed that the UE-A and UE-B have the same Computation Capacity, i.e. the time used for computing the certain AlexNet Model Layers are the same for UE-A and UE-B. Otherwise, the Data Rate on Uu and Sidelink may be changed accordingly.*



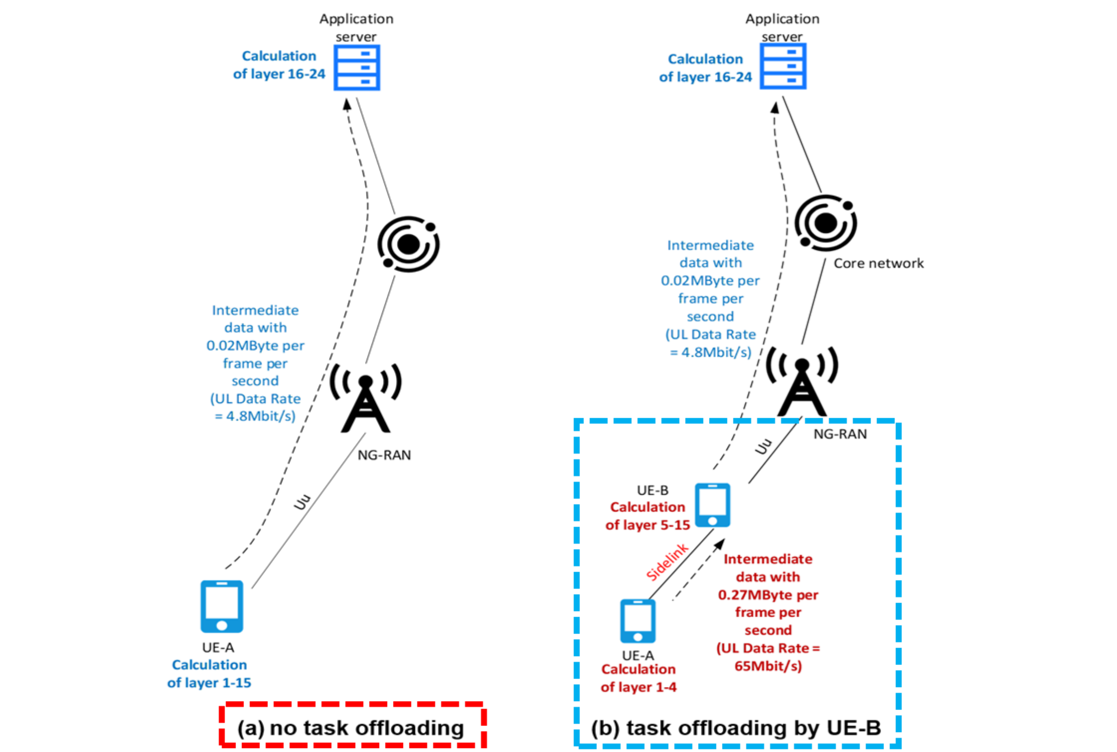



**Figure: Using Direct Device Connection (SL) to realize the Proximity-based Work Task off-loading with the Data Rate on Uu need not be increased while the original UE's Computation Load is off-loaded**



**Table: KPI Requirements for Proximity-based Work Task offloading**

| | UL data size (for sidelink) | UL data rate (for sidelink) | Intermediate data uploading latency (including sidelink+Uu) | Image recognition latency |
|---|---|---|---|---|
| AlexNet model with 30FPS (NOTE 1) | 0.15 - 0.02 Mbyte for each frame | 4.8 – 65 Mbit/s | - 2ms for Remote driving, AR displaying/gaming, and remote-controlled robotics;<br>- 10ms for video recognition;<br>- 100ms for One-shot object recognition, Person identification, or photo enhancement in smart phone | 1s |
| VGG-16 model with 30FPS | 0.1 - 1.5 Mbyte for each frame | 24 - 720 Mbit/s | | 1s |
| NOTE 1: FPS stands for Frame Per Second | | | | |

# 3. 5G System use of AI/ML

***Latency*** is a critical requirement. The Figure summarizes the Latency Cost in three (3) scenarios:

(A) The *Inference of Model M is done locally*.
***Latency*** is denoted $L_{LI}$.

(B) The *Inference Process* is fully *off-loaded on a second (2nd) device*.
***Latency*** is denoted $L_{FO}$.

(C) The *Inference Process* is partially *off-loaded on a second (2nd) device*.
***Latency*** is denoted $L_{PO}$.

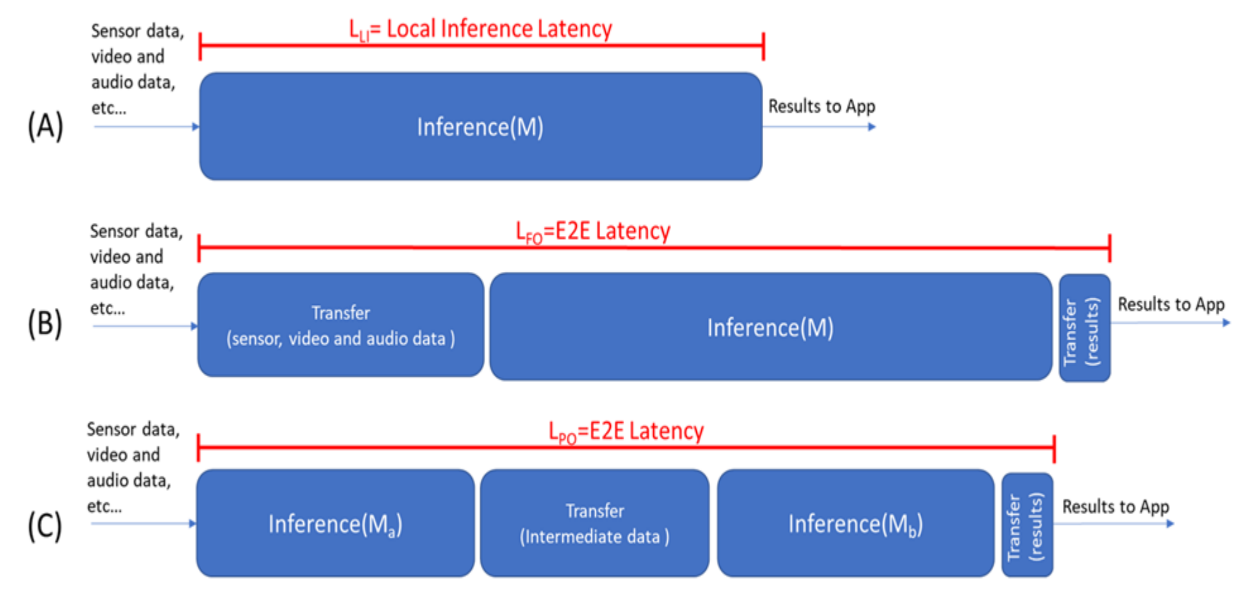
There are three types of AIML Operations such as:

- AI/ML Operation Splitting between AI/ML End-Points;
- AI/ML Model/Data Distribution and Sharing over 5G System (5GS);
- Distributed/Federated Learning (FL) over 5G System (5GS).

*Use Cases (UCs)* corresponding to ***the three (3) Types of AIML Operations incorporates the assistance of Direct Device Connection***.

Release 19 3GPP V19.4.0 (2023-09)

**Figure: AI/ML Model Inference Latency summary**




Release 19 3GPP V19.4.0 (2023-09)

**Table: 5G System Architecture Service Requirements KPI Table of AI/ML Model/Data Distribution and sharing by leveraging direct device connection**

| Max allowed end-to-end latency (NOTE 1) | Experienced data rate (NOTE 1) | Payload size (NOTE 1) | Communication service availability (NOTE 1) | Remark |
|---|---|---|---|---|
| 1s | ≤ 1.92 Gbit/s | ≤ 240 MByte | 99.9 % | AI Model Transfer Management through Direct Device Connection |
| 3s | ≤ 81.33 Mbyte/s | ≤ 244 MByte | - | transfer learning for trajectory prediction |
| NOTE 1: The KPIs in the table apply to data transmission using direct device connection.<br>NOTE 2: The AI/ML model data distribution is for a specific application service | | | | |

*5G System-assisted Transfer Learning for Trajectory Prediction*

AIML Model Transfer Learning is beneficial for lowering cost and raising effectiveness when training a Model using a Target UE based on a pre-training model. The Principle of transfer learning is to use the knowledge from the Source Domain to train a Model in the Target Domain to achieve more expedient and higher accuracy efficiency.

Since ***the AI Model is a kind of Knowledge***, when the Centralized Application Server acquires *enough Number of AIML Model used by UEs*, it may perform a backward inference/inversion attacks to derive the feature of *UE's Local Data Set*, which means a *Privacy Risk exists.*

In order to resolve the Privacy concern for Transfer Learning, the Model Transfer via direct device connection is a better to be used so that the Network Node (e.g. Application Server) cannot acquire the AI/ML Model used by UE and no way to do backward inference.

Release 19 3GPP V19.4.0 (2023-09)

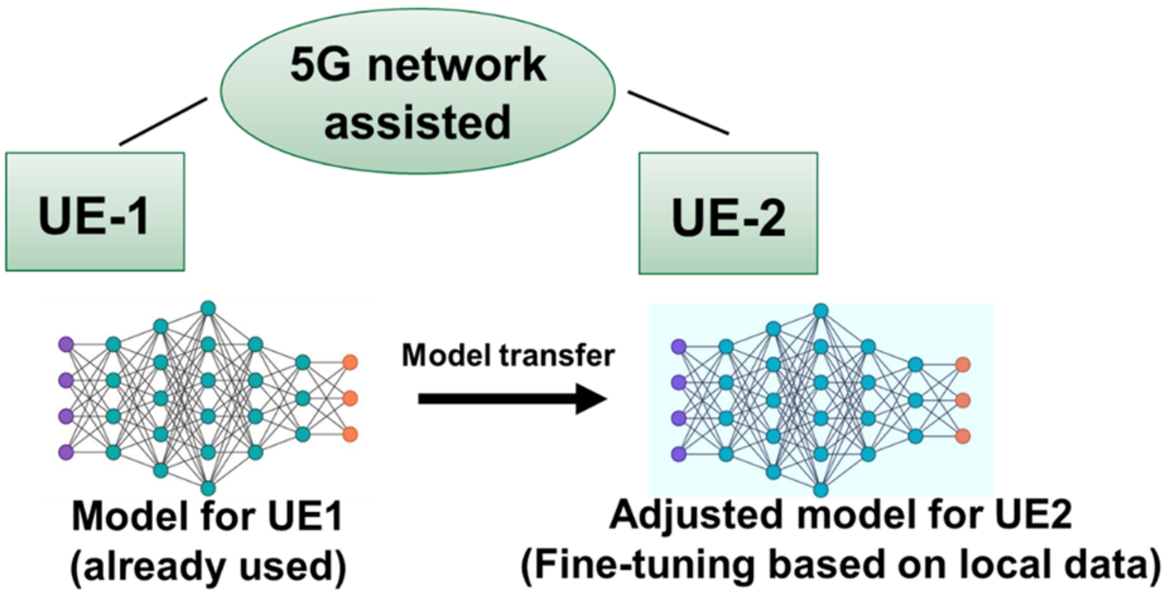



Figure: 5G Network-assisted AI/ML Model Transfer Learning from Source to Target UE

## 3. 5G System use of AI/ML

***5G System Architecture AI/ML Model/Data Distribution and sharing by leveraging Direct Device Connection***

Operators can provide Services to help manage and distribute the AI/ML Models especially in the "Edge" Server so that the UE can acquire a proper Model immediately.

However, when a lot of UEs requesting for the same Model at the same Time or the UE is blocked by barriers with poor connection with the Base Station, the Model Transfer Process will become longer than expected.

To overcome this difficulty, as shown in the Figure, a "volunteer" UE ,which is well connected to the Base Station, can help "Relaying" AI/ML Models or Receive & Store AI/ML Models first.

Then, the other UEs can download AI/ML Models from the "volunteer" UE through Direct Device Connection.

In this way, all UE can have a "stable" & "reliable" Model Transfer Process while the Radio Resource of the Base Station can be saved.

Besides, the "volunteer" UE can transfer the stored Models to other "volunteer" UEs under Operator's control.

*The Selection of "volunteer" UE can be realized by Local Network Policies and Strategies (utilizing the 5GC Functionality enhancement of support of LADN to DNN and S-NSSAI).* And it also can be exposed as a Capability to the 3rd Party Company when the Company wants to choose one (1) or a few "certain" UEs to be "volunteer" UEs in an activity.

Release 19 3GPP V19.1.0 (2023-09)

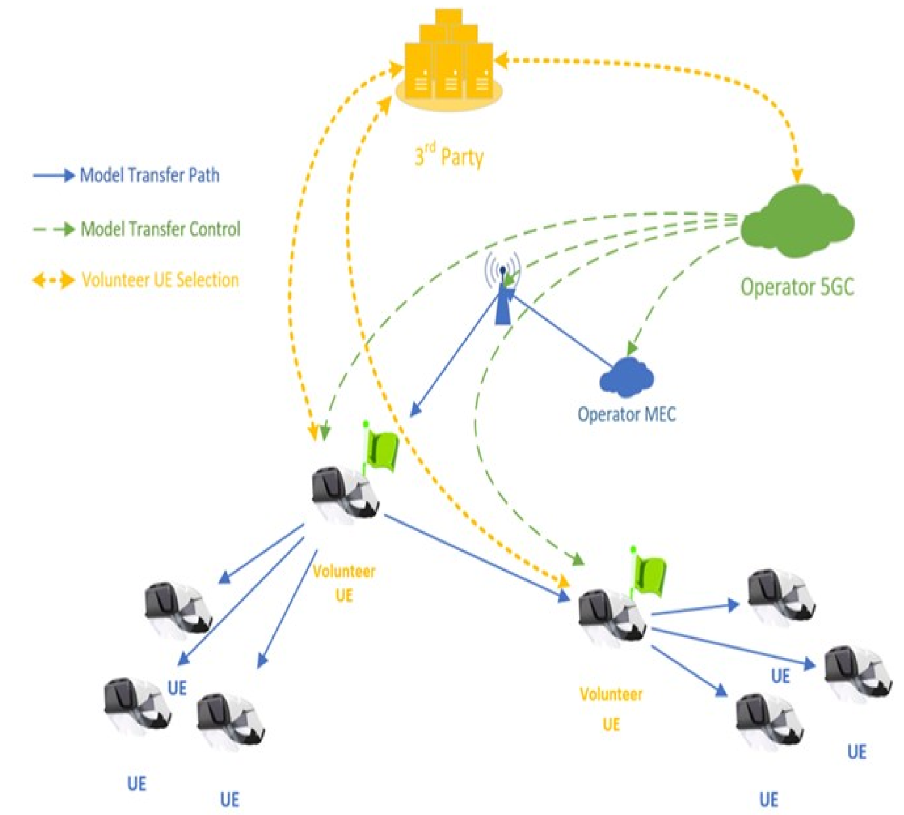



Figure: 5G System Architecture AI/ML Model Management through Direct Device Connection

E.g., a Travel Company may assign the tour (4) Guides' Augmented Reality (AR) Headsets as "volunteer" UEs in a Carnival through the Operator's Network exposure. The Travel Company may sign a Higher Quality Plan for Tourist Guides' Devices to provide better User Experience for following Tourists. Meanwhile, operator can benefit from the *alternative Open Service based on AI/ML Model Management Capabilities* and may avoid "low" Quality of Service (QoS) due to crowding "direct connections" to Base Stations during the Carnival.

# 3. 5G System use of AI/ML

*5G System (5GS) AI/ML Model Transfer KPIs*

The 5GS shall support *split AI/ML Inference* between *UE* and *Network Server/ Application Function* with Performance Requirements as given in the Table.

Release 19 3GPP V19.4.0 (2023-09)

**Table: 5G System AI/ML Model Transfer KPI of split AI/ML Inference between UE and Network Server/ Application Function (AF)**

| Uplink KPI | | | | | Downlink KPI | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Max allowed UL end-to-end latency | Experienced data rate | Payload size | Communication service availability | Reliability | Max allowed DL end-to-end latency | Experienced data rate | Payload size | Reliability | Remarks |
| 2 ms | 1.08 Gbit/s | 0.27 MByte | 99.999 % | 99.9 % | | | | 99.999 % | Split AI/ML image recognition |
| 100 ms | 1.5 Mbit/s | | | | 100 ms | 150 Mbit/s | 1.5 MByte/ frame | | Enhanced media recognition |
| | 4.7 Mbit/s | | | | 12 ms | 320 Mbit/s | 40 kByte | | Split control for robotics |
| NOTE 1: Communication service availability relates to the service interfaces, and reliability relates to a given system entity. One or more retransmissions of network layer packets can take place in order to satisfy the reliability requirement. | | | | | | | | | |

The 5GS shall support *AI/ML Model downloading* with *Performance Requirements* as given in the Table:

Release 19 3GPP V19.4.0 (2023-09)

**Table: 5G System AI/ML Model Transfer KPI Table of AI/ML Model Downloading**

| Max allowed DL end-to-end latency | Experienced data rate (DL) | Model size | Communication service availability | Reliability | User density | # of downloaded AI/ML models | Remarks |
|---|---|---|---|---|---|---|---|
| 1s | 1.1Gbit/s | 138MByte | 99.999 % | 99.9% for data transmission of model weight factors; 99.999% for data transmission of model topology | | | AI/ML model distribution for image recognition |
| 1s | 640Mbit/s | 80MByte | 99.999 % | | | | AI/ML model distribution for speech recognition |
| 1s | 512Mbit/s(see note 1) | 64MByte | | | | Parallel download of up to 50 AI/ML models | Real time media editing with on-board AI inference |
| 1s | | 536MByte | | | up to 5000~ 10000/km2 in an urban area | | AI model management as a Service |
| 1s | 22Mbit/s | 2.4MByte | 99.999 % | | | | AI/ML based Automotive Networked Systems |
| 1s | | 500MByte | | | | | Shared AI/ML model monitoring |
| 3s | 450Mbit/s | 170MByte | | | | | Media quality enhancement |

NOTE 1: 512Mbit/s concerns AI/ML models having a payload size below 64 MB. TBD for larger payload sizes.
NOTE 2: Communication service availability relates to the service interfaces, and reliability relates to a given system entity. One or more retransmissions of network layer packets can take place in order to satisfy the reliability requirement.

# 3. 5G System use of AI/ML

*5G System (5GS) AI/ML Model Transfer KPIs*

The 5G System shall support ***Federated Learning (FL) between UE and Network Server/Application Function (AF)*** with Performance Requirements as given in the Table:

Release 19 3GPP V19.4.0 (2023-09)

**Table: 5G System AI/ML Model Transfer KPI Table of Federated Learning between UE and Network Server/Application Function**

| Max allowed DL or UL end-to-end latency | DL experienced data rate | UL experienced data rate | DL packet size | UL packet size | Communication service availability | Remarks |
|---|---|---|---|---|---|---|
| 1s | 1.0Gbit/s | 1.0Gbit/s | 132MByte | 132MByte | | Uncompressed Federated Learning for image recognition |
| 1s | 80.88Mbit/s | 80.88Mbit/s | 10Mbyte | 10Mbyte | TBD | Compressed Federated Learning for image/video processing |
| 1s | TBD | TBD | 10MByte | 10MByte | | Data Transfer Disturbance in Multi-agent multi-device ML Operations |

The 5G System shall support Split AI/ML Inference between AI/ML End-points by leveraging ***Direct Device Connection*** with *Performance Requirements* as given in Table

Release 19 3GPP V19.4.0 (2023-09)

**Table: 5G System AI/ML Model Transfer KPI Table of Split AI/ML Operation between AI/ML End-points f for AI Inference by leveraging Direct Device Connection**

| Max allowed end-to-end latency (NOTE 1) | Payload size (Intermediate data size) (NOTE 1) | Experienced data rate (NOTE 1) | Service area dimension | Communication service availability (NOTE 1) | Reliability (NOTE 1) | Remarks |
|---|---|---|---|---|---|---|
| 10–100 ms | ≤ 1.5 Mbyte for each frame | ≤ 720 Mbps | | | | Proximity-based work task offloading for Remote driving, AR displaying/gaming, remote-controlled robotics, video recognition and One-shot object recognition |
| 10 ms | ≤ 1.6 MByte (8 bits data format) | ≤ 1.28 Gbps | 900 $m^2$ (30 m x 30 m) | 99.999 % | 99.99 % | Local AI/ML model split on factory robots |
| 10 ms | ≤ 6.4 Mbyte (32 bits data format) | ≤ 1.5 Gbps | | | | Local AI/ML model split on factory robots |
| NOTE 1: The KPIs in the table apply to UL data transmission in case of indirect network connection. | | | | | | |

*5G System (5GS) AI/ML Model Transfer KPIs*

The 5G System shall support ***AI/ML Model/Data Distribution and Sharing by leveraging Direct Device Connection*** with Performance Requirements as given in the Table:

Release 19 3GPP V19.4.0 (2023-09)

**Table: 5G System Architecture Service Requirements KPI Table of AI/ML Model/Data Distribution and sharing by leveraging direct device connection**

| Max allowed end-to-end latency (NOTE 1) | Experienced data rate (NOTE 1) | Payload size (NOTE 1) | Communication service availability (NOTE 1) | Remark |
|---|---|---|---|---|
| 1s | ≤ 1.92 Gbit/s | ≤ 240 MByte | 99.9 % | AI Model Transfer Management through Direct Device Connection |
| 3s | ≤ 81.33 Mbyte/s | ≤ 244 MByte | - | transfer learning for trajectory prediction |
| NOTE 1: The KPIs in the table apply to data transmission using direct device connection.<br>NOTE 2: The AI/ML model data distribution is for a specific application service | | | | |

The 5G System shall support ***AI/ML Model/Data Distribution and Sharing Federated Learning (FL) by leveraging Direct Device Connection*** with *Performance Requirements* as given in the Table:

Release 19 3GPP V19.4.0 (2023-09)

**Table: 5G System AI/ML Model Transfer KPI Table of Distributed/Federated Learning by leveraging Direct Device Connection**

| Payload size (NOTE 1) | Maximum latency (NOTE 1) | Experienced data rate (NOTE 1) | Reliability (NOTE 1) | Remark |
|---|---|---|---|---|
| 132 MByte | 2-3 s | ≤ 528 Mbit/s | | Direct device connection assisted Federated Learning (Uncompressed model) Asynchronous Federated Learning via direct device connection |
| ≤ 50 MByte | 1 s | ≤ 220 Mbit/s | 99.99% | |
| NOTE 1: The KPIs in the table apply to both UL and DL data transmission in case of indirect network connection. | | | | |

## *ML Knowledge Transfer Learning*

It is known that existing ML Capability can be *leveraged in producing or improving New or other ML Capability.*

Specifically, using Transfer Learning Knowledge contained in one (1) or more ML Entities may be transferred to another ML Entity.

Transfer Learning relies on Task and Domain similarity to deduce whether some parts of a deployed ML Entity can be re-used in another Domain / Task with some modifications.

As such, aspects of Transfer Learning that are appropriate in Multi-Vendor Environments need to be supported in Network Management Systems.

However, ML Entities are likely to not be Multi-Vendor Objects, i.e. it will, in most cases, not be possible to transfer an ML Entity from Function to another.

Instead, the Knowledge contained in the Model should be transferred instead of transferring the ML Entity itself as e.g. the Knowledge contained in an ML Entity deployed to perform Mobility Optimization by Day can be leveraged to produce a new ML Entity to perform Mobility Optimization by Night.

As such and as illustrated in the Figure, the Network or its Management System needs to have the required Management Services for ML Transfer Learning (MLKLT), where ML Transfer Learning refers to means to allow and support the usage and fulfilment of transfer learning between any two ML Entities.

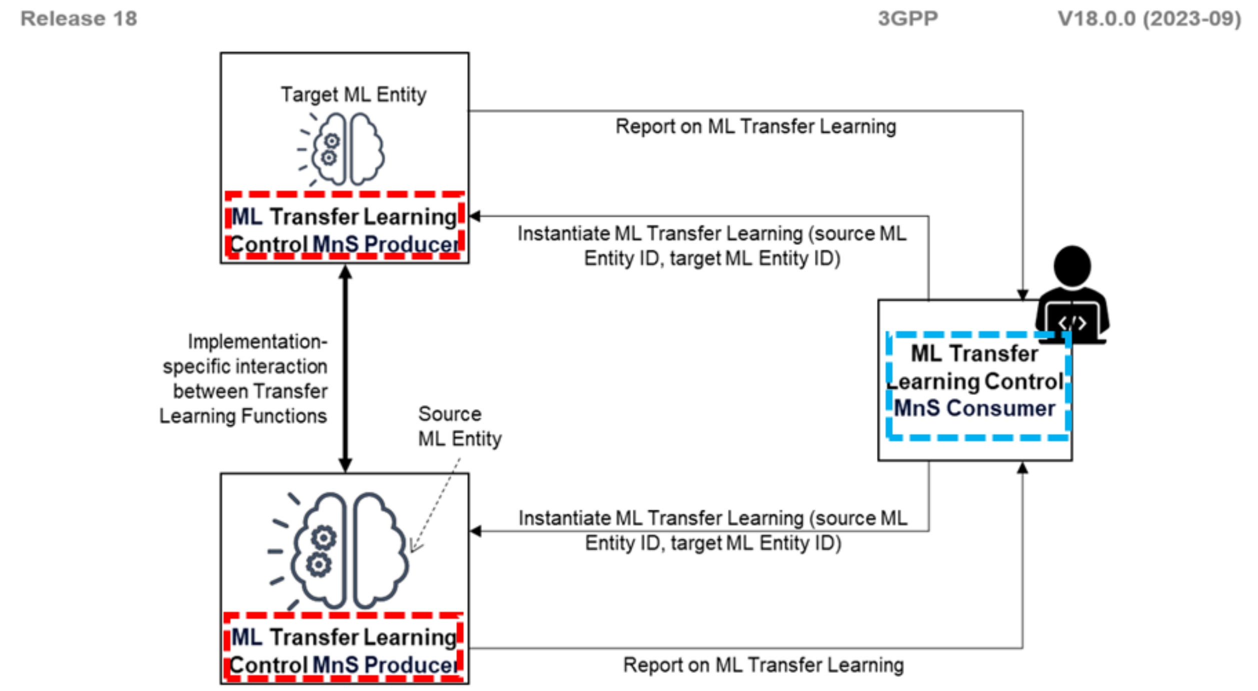



**Figure: ML Knowledge Transfer Learning (MLKLT) Flow between the Source MLKLT (which is the Entity with the Pre-trained ML Entity), the Peer MLKLT (which is the Entity that shall train a New ML Entity) and the MLKLT MnS Consumer (which may be the Operator or another Management Function that wishes to trigger or control MLKLT)**

### *Use Cases (UCs)*

*1. Discovering Sharable Knowledge*

*2. Knowledge Sharing and Transfer Learning*

## Use Cases (UCs)

### *Discovering Sharable Knowledge*

For the Transfer Learning, it is expected that the ***Source ML Knowledge Transfer Learning MnS Producer*** shares its Knowledge with the Target ML Training Function, either simply as Single Knowledge Transfer Instance or through an Interactive Transfer Learning Process.

The Concept of Knowledge here represents any Experiences or Information gathered by the ML Entity in the *ML Knowledge Transfer Learning MnS Producer* through

- *Training,*
- *Inference,*
- *Updates, or*
- *Testing.*

This *Information or Experiences* can be in the form of - but not limited to *Data Statistics or other Features of the underlying ML Model.*

It may also be the output of an ML Entity.

The 3GPP Management Systems should provide means for an *MnS Consumer* to discover this potentially shareable knowledge as well as means for the provider of MLKLT to share the *Knowledge with the MnS Consumer.*

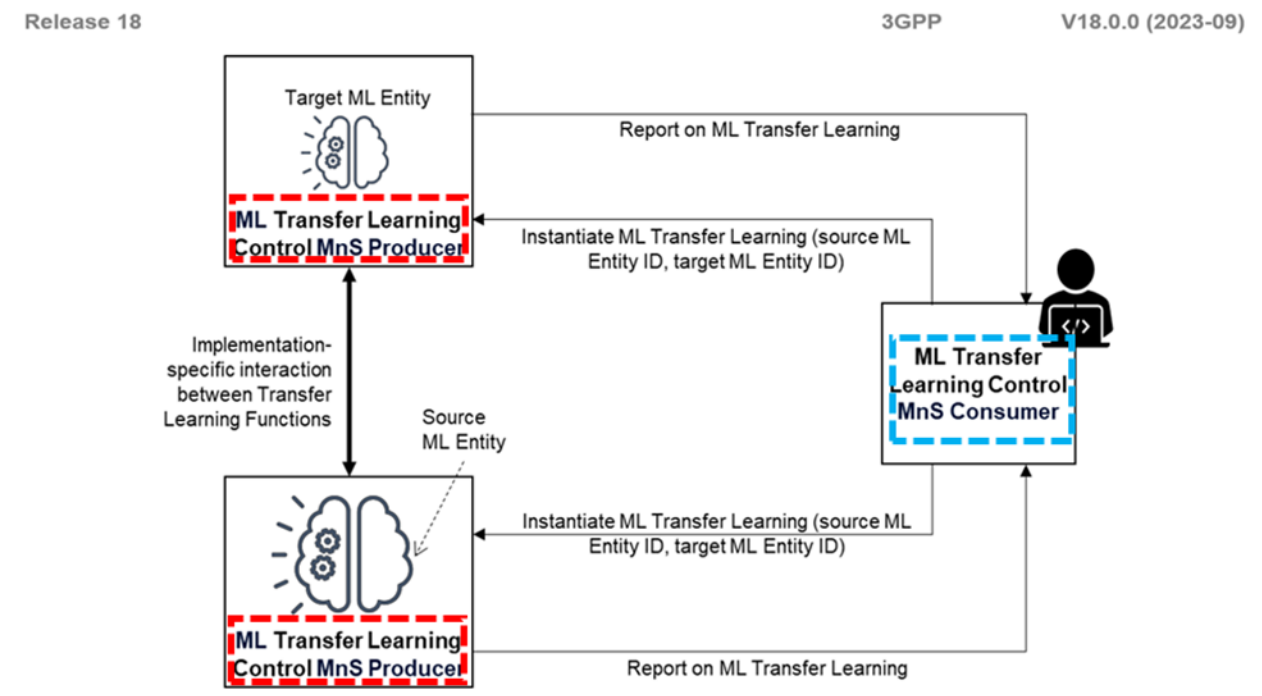



**Figure: ML Knowledge Transfer Learning (MLKLT) Flow between the Source MLKLT (which is the Entity with the Pre-trained ML Entity), the Peer MLKLT (which is the Entity that shall train a New ML Entity) and the MLKLT MnS Consumer (which may be the Operator or another Management Function that wishes to trigger or control MLKLT)**

## Knowledge Sharing and Transfer Learning

The Transfer Learning may be triggered by a *MnS Consumer* either to fulfil the learning for itself or for it to be accomplished through another ML Training Function.

The Entity containing the Knowledge may be an Independent Managed Entity (the ML Entity).

Alternatively, the ML Model may also be an Entity that is not independently managed but is an attribute of a managed ML Entity or ML Function in which case MLKLT does not involve sharing the ML Model or parts thereof but may imply implementing the means and services to enable the sharing of knowledge contained within the ML Entity or ML-enabled Function.

The 3GPP Management System should provide means and the related Services needed to realize the ML Transfer Learning Process.

Specifically, the 3GPP Management System should provide means for an MnS Consumer to request and receive Sharable Knowledge as well as means for the Provider of MLKLT to share the Knowledge with the MnS Consumer or any stated Target ML Training Function. Similarly, the 3GPP Management System should provide means for an MnS Consumer to manage and control the MLKLT Process and the related requests associated with Transfer Learning between two (2) ML Entities or between the two (2) ML Entities and a Shared Knowledge Repository.

The two (2) Use Cases (UCs) should address the four (4) Scenarios described in the Figures.

***Note** that, the UC and Requirements focus on the Required Management Capabilities*.

Release 18 3GPP V18.0.0 (2023-09)

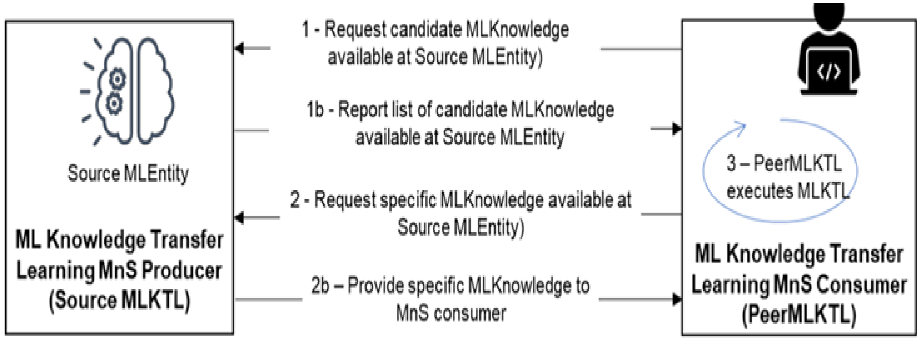



Figure: Scenario 1 - Interactions for ML-Knowledge Transfer Learning (MLKLT) to support Training at the ML Knowledge Transfer MnS Consumer - the ML Knowledge Transfer MnS Consumer obtains the ML Knowledge which it then uses for Training the New ML Entity based on Knowledge received from the MLKLT Source MnS Producer

Release 18 3GPP V18.0.0 (2023-09)

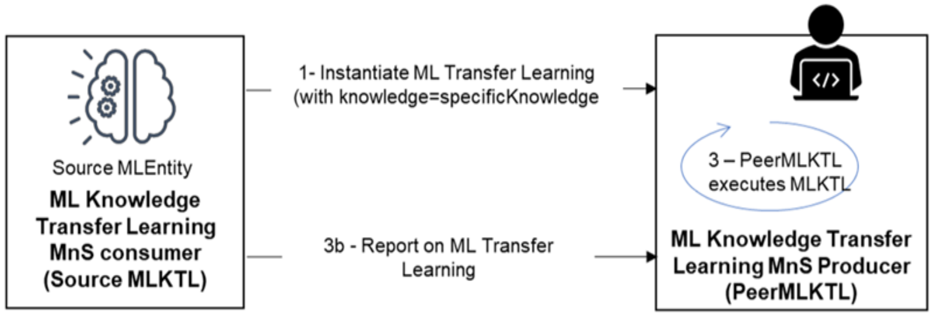



Figure: Scenario 2 - Interactions for ML-Knowledge Transfer Learning (MLKTL) to support Training at the ML Knowledge Transfer MnS Consumer triggered by the MLKTL Source - the ML Transfer Learning MnS Consumer acting as the MLKTL Source (the Source of the ML Knowledge) triggers the Training at the ML Knowledge Transfer MnS Consumer by providing the ML Knowledge to be used for the Training, the ML Transfer Learning MnS Consumer then undertakes the Training

# 3. 5G System use of AI/ML

## *Knowledge Sharing and Transfer Learning*

Specifically, the 3GPP Management System should provide means for an MnS Consumer to request and receive Sharable Knowledge as well as means for the Provider of MLKLT to share the Knowledge with the MnS Consumer or any stated Target ML Training Function.

Similarly, the 3GPP Management System should provide means for an MnS Consumer to manage and control the MLKLT Process and the related requests associated with Transfer Learning between two (2) ML Entities or between the two (2) ML Entities and a Shared Knowledge Repository.

The two (2) Use Cases (UCs) should address the four (4) Scenarios described in the Figures.

***Note** that, the UC and Requirements focus on the Required Management Capabilities*.

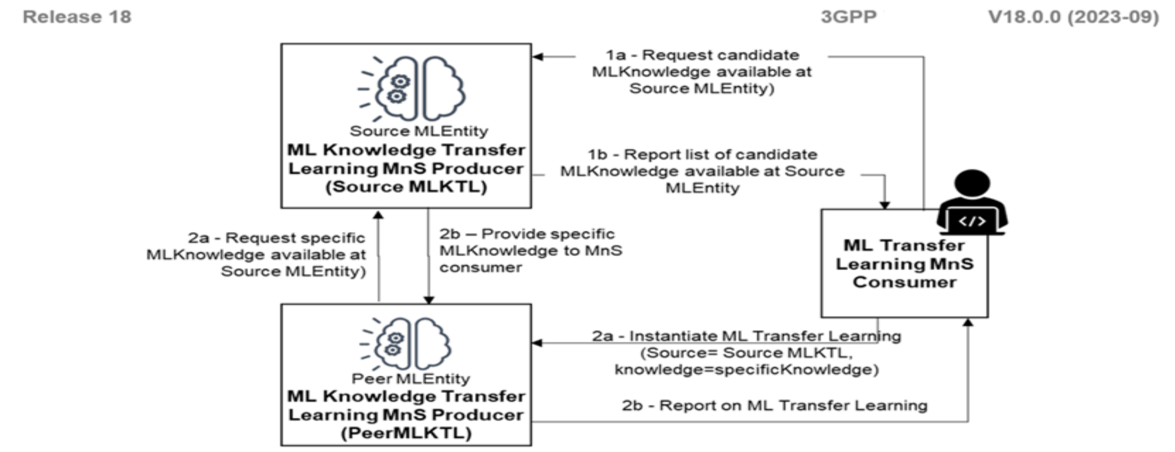



**Figure: Scenario 3 - Interactions for ML-Knowledge Transfer Learning (MLKLT) to support Training at the Peer ML Knowledge Transfer MnS Producer, that is different from the ML Knowledge Transfer MnS Consumer - the ML Knowledge Transfer MnS Consumer triggers Training at the MLKLT peer MnS Producer. The MLKLT MnS Consumer then obtains the ML Knowledge from the MLKLT Source MnS Producer and then uses the Knowledge for Training the New ML Entity based on Knowledge received from the MLKLT Source MnS Producer**

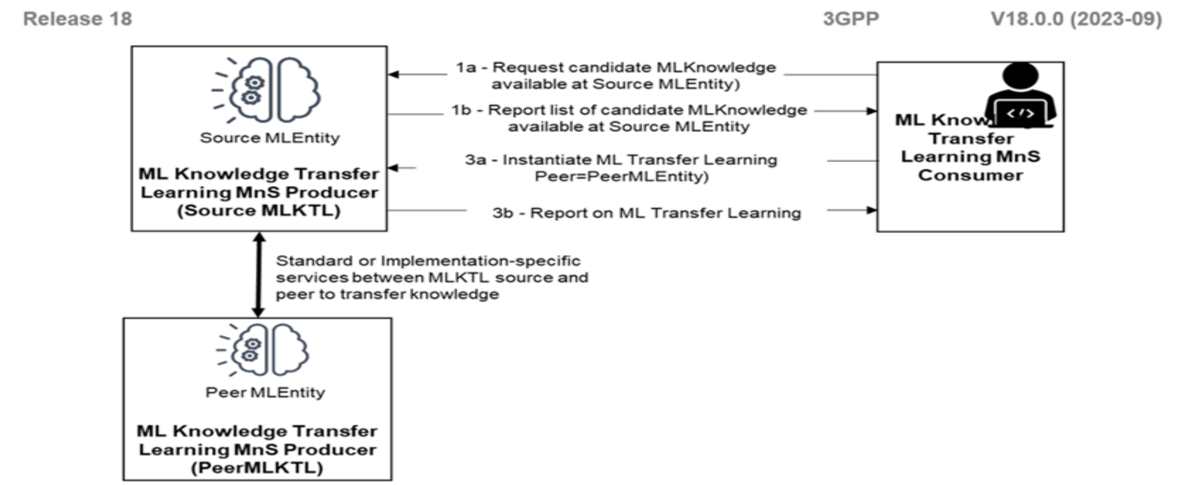



**Figure: Scenario 4 - interactions for ML-Knowledge Transfer Learning (MLKLT) to support Training at the Source ML knowledge Transfer MnS producer - the ML Knowledge Transfer MnS consumer triggers training at the MLKLT source MnS producer. The MLKLT MnS consumer then obtains the ML knowledge from the MLKLT Source MnS Producer and then uses the Knowledge for Training the new ML Entity based on Knowledge received from the MLKLT Source MnS Producer**

## 3. 5G System use of AI/ML

*Identifying Capabilities of ML entities*

Network Functions (NFs), especially Network Automation Functions, may need to rely on AI/ML Capabilities that are not internal to those Network Functions (NFs) to accomplish the desired Automation. E.g., "an MDA Function may optionally be deployed as one or more AI/ML inference function(s) in which the relevant models are used for inference per the corresponding MDA capability." Similarly, *owing to the differences in the kinds and complexity of intents that need to be fulfilled, an intent fulfilment solution may need to employ the capabilities of existing AI/ML to fulfil the Intents.*

In any such case, Management Services are required to identify the Capabilities of those existing ML Entities.

The Figure shows that the Consumer may wish to obtain Information about *AI/ML Capabilities to determine how to use them for the Consumer's needs, e.g. for fulfilment of Intent Targets or other Automation Targets.*

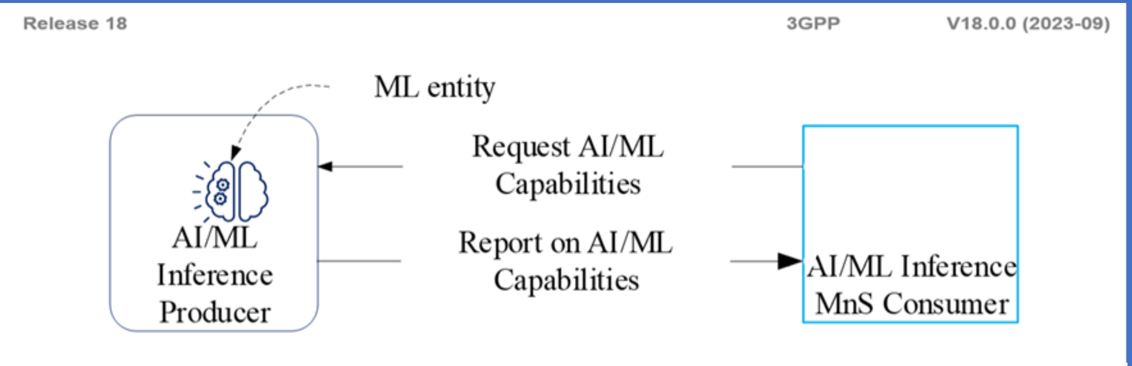



**Figure: 5GS MDA Request and Reporting AI/ML Capabilities**

## 3. 5G System use of AI/ML

*Mapping of the Capabilities of ML Entities*

Besides the discovery of the Capabilities of ML Entities, Services are needed for mapping the ML Entities and Capabilities.

Instead of the Consumer discovering Specific Capabilities, the Consumer may want to know the ML E'ntities than can be used to achieve a certain outcome.

For this, the Producer should be able to inform the Consumer of the set of ML Entities that together achieve the Consumer's Automation Needs.

In the case of Intents e.g., the complexity of the stated intents may significantly vary - from simple intents which may be fulfilled with a call to a single ML entity to complex intents that may require an intricate orchestration of multiple ML entities.

For simple Intents, it may be easy to map the execution logic to the one (1) or multiple ML Entities.

For complex intents, it may be required to employ multiple ML Entities along with a corresponding functionality that manages their inter-related execution. The usage of the ML entities requires the awareness of the capabilities of their capabilities and interrelations.

Moreover, given the complexity of the required mapping to the multiple ML entities, services should be supported to provide the mapping of ML Entities and Capabilities.



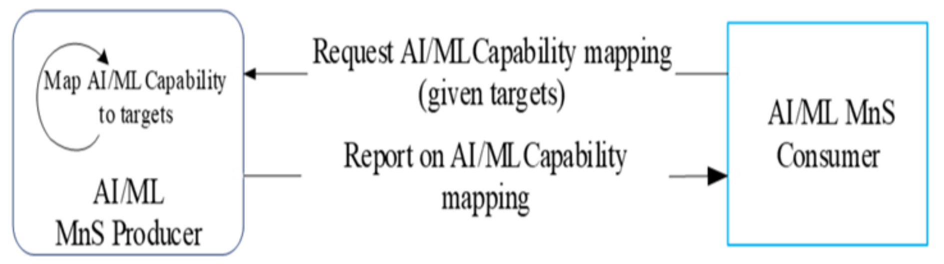



**Figure: 5GS MDA Mapping Execution Logic to AI/ML Capabilities**

**NOTE:** *The Figure shows that the Consumer may wish to obtain the Mapping of AI/ML Capabilities to some Management Tasks to determine how to use them for the Consumer's needs, e.g. for its Intent targets or other Automation targets. The Management Tasks may include specific metrics to be optimized, but the candidate tasks to be considered are to be agreed at the normative phase.*

# 3. 5G System use of AI/ML

5GS AI/ML Management Functionality and Service Framework for ML Training

An ML Training Function playing the Role of ML Training *MnS Producer*, may consume various Data for ML Training purpose.

As illustrated in the Figur,  the ML Training Capability is provided via *ML Training MnS* in the *context of SBMA* to the authorized *Consumer(s) by ML Training MnS Producer*.

The Internal Business Logic of ML Training leverages the current and Historical relevant Data, including those listed below to monitor the Networks and/or Services where:
- Relevant to the ML Model,
- Prepare the Data,
- Trigger and Conduct the Training:
- Performance Measurements (PM) and Key Performance Indicators (KPIs)
- Trace/MDT/RLF/RCEF Data,
- QoE and Service Experience Data .
- Analytics Data offered by NWDAF
- Alarm Information and Notifications
- CM Information and Notifications
- MDA Reports from *MDA MnS Producers*
- Management Data from Non-3GPP Systems.
- Other Data that can be used for training.

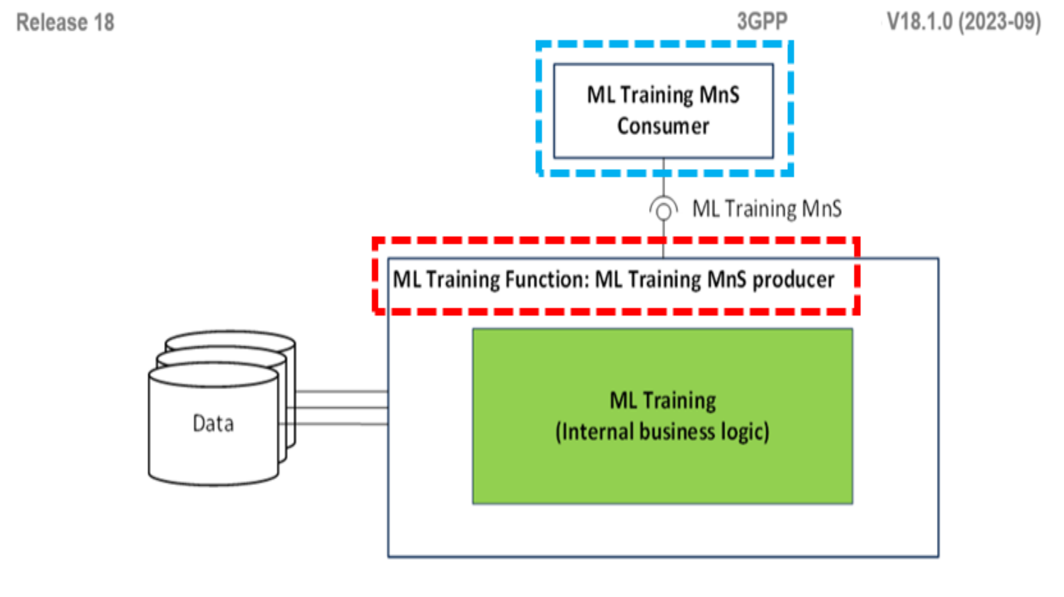



Figure: 5GS ML Functional Overview and Service Framework

# 3. 5G System use of AI/ML

*5GS ML Training requested by (MnS) Consumer*

The ML Training Capabilities are provided by an *MLT MnS Producer to one (1) or more (MnS) Consumer(s).*

The ML Training may be triggered by the request(s) from one (1) or more *MLT MnS Consumer(s).*

The "Consumer" , e.g., a Network Function (NF), a Management Function (MnF), an Operator (CSP), or another Functional Differentiation, to trigger an ML Training, the *MLT MnS Consumer* requests the *MLT MnS Producer* to train the *ML Model*.

In the *ML Training Request,* the "Consumer", should specify the *Inference Type,* which indicates the *Function or Purpose of the ML Entity,* e.g. Coverage Problem Analysis.
The *MLT MnS Producer* can perform the Training according to the designated Inference Type.

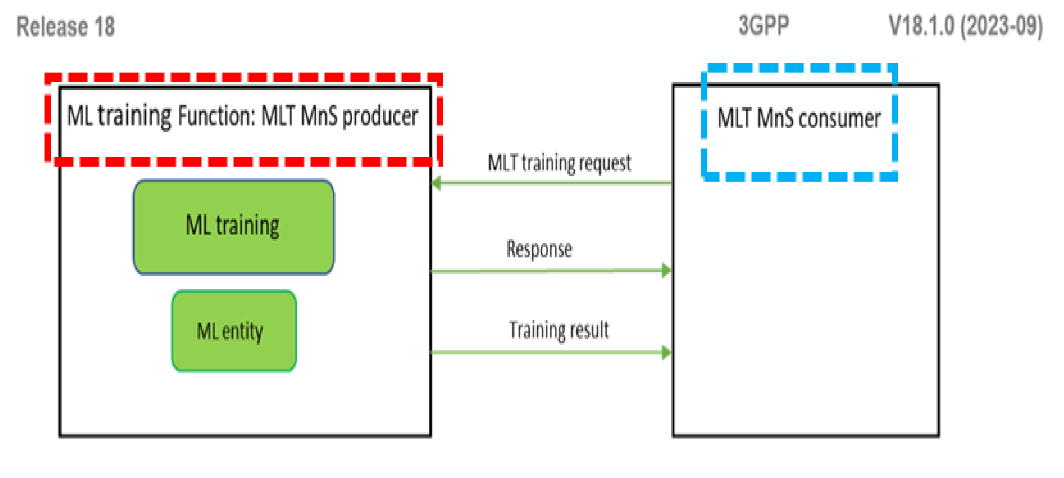


Figure: 5GS ML Training requested by ML MnS Consumer

The "*Consumer"* may provide the *Data Source(*s) that contain(s) the *Training Data*, which are considered as inputs candidates for training.
To obtain the *Valid Training Outcomes*, *Consumers* may also designate their *Requirements for Model Performance (e.g. Accuracy, etc.) in the Training Request.*

The *MLT MnS Producer* provides a response to the *Consumer* indicating whether the request was accepted.
If the request is accepted, the *MLT MnS Producer* decides when to start the ML Training with consideration of the request(s) from the Consumer(s).

Once the Training is decided, *the Producer performs the followings:*
- *selects the Training Data,* with consideration of the *Consumer provided Candidate Training Data*. Since the *Training Data directly influences the Algorithm and Performance of the trained ML Entity,* the *MLT MnS Producer* may examine the Consumer's provided Training Data and decide to select none, some or all of them. In addition, the *MLT MnS Producer* may select some *other Training Data that are available*;
- *Trains the ML Entity* using the *Selected Training Data*;
- provides the Training Results to the MLT MnS Consumer(s).

The *MLT MnS Producer* provides a response to the Consumer indicating whether the Request was accepted. If the request is accepted, the *MLT MnS Producer* decides when to start the ML Training with consideration of the Request(s) from the Consumer(s). Once the Training is decided, the Producer performs the followings:

# 3. 5G System use of AI/ML

## Handling Errors in Data and ML Decisions

Traditionally, the ML Models/Entities (e.g. , ML Entity1 and ML Entity2 in the Figure) are trained on "good quality" Data, i.e. , Data that were collected correctly and reflected the Real Network Status to represent the expected Context in which the ML Entity is meant to operate.

"Good Quality Data" is void of Errors, such as:

- Imprecise Measurements, with added Noise (such as RSRP, SINR, or QoE Estimations).

- Missing Values or Entire Records, e.g. , because of Communication Link failures.

- Records which are communicated with a significant delay (in case of online measurements).

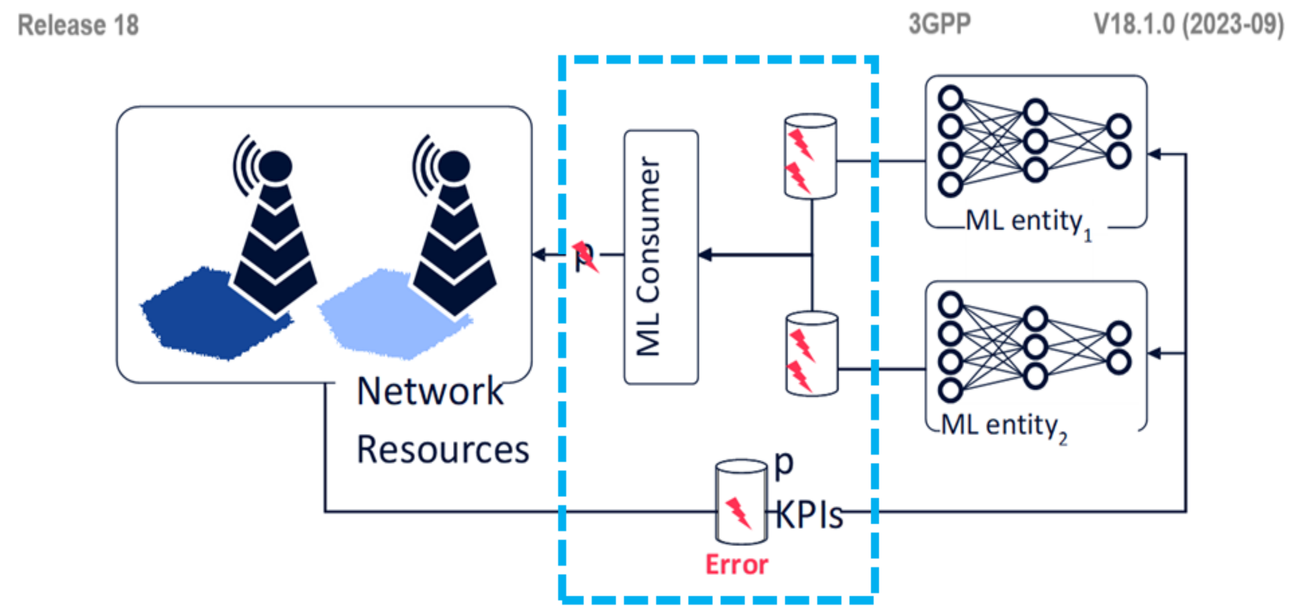



Figure: 5G System AI/ML Management example of Network Resource Propagation of Erroneous Information

Without Errors, an ML Entity can depend on a few precise Inputs, and does not need to exploit the Redundancy present in the Training Data.

However, during Inference, the ML Entity is very likely to come across these inconsistencies. When this happens, the ML Entity shows High Error in the Inference Outputs, even if Redundant and Uncorrupted Data are available from other Sources.

As such the System needs to account for Errors and Inconsistencies in the Input Data and the Consumers should deal with Decisions that are made based on such Erroneous and Inconsistent Data. The System should:

1) Enable Functions to undertake the Training in a way that prepares the ML Entities to deal with the Errors in the Training Data, i.e. , to identify the Errors in the Data during Training;
2) Enable the MLT MnS Consumers to be aware of the possibility of Erroneous Input Data that are used by the ML Entity.

## 3. 5G System use of AI/ML

Information Model Definitions for AI/ML Operational Phases

Information Model Definitions for ML Training for the Set of Classes (e.g. IOCs) that encapsulates the Information relevant to ML Model Training for NRM (using the UML Semantics).



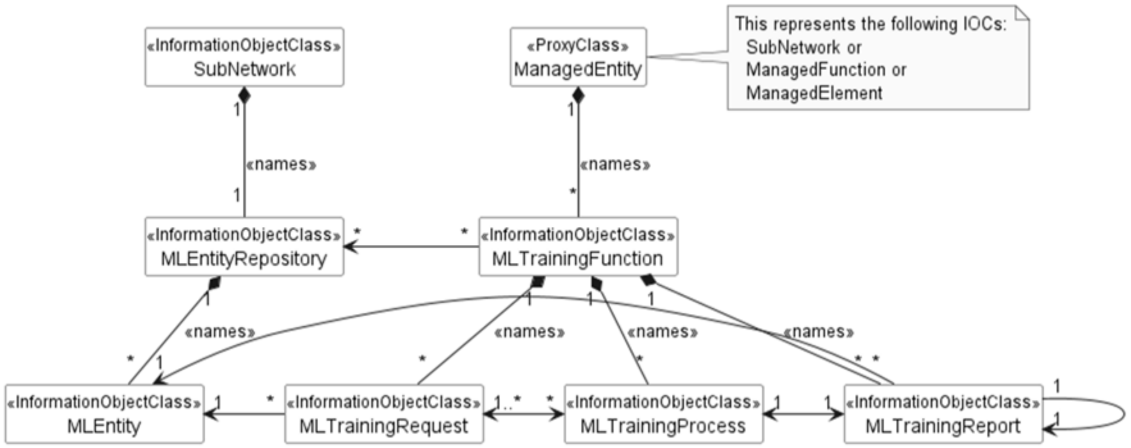



**Figure: 5G System AI/ML Management NRM fragment for ML Training**



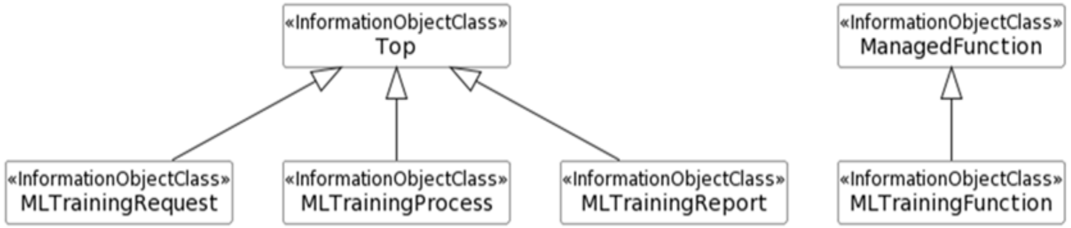



**Figure: 5G System AI/ML Management Inheritance Hierarchy for ML Training related NRM**



OpenAPI definition of the AI/ML NRM

OpenAPI document AiMlNrm.yaml


```
openapi: 3.0.1
info:
  title: AI/ML NRM
  version: 18.1.0
  description: >-
    OAS 3.0.1 specification of the AI/ML NRM
    © 2023, 3GPP Organizational Partners (ARIB, ATIS, CCSA, ETSI, TSDSI, TTA, TTC).
    All rights reserved.
externalDocs:
  description: 3GPP          AI/ML Management
```

### 3. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - ***use of AI/ML*** - 7

The 5G System Architecture allows NWDAF containing ***Analytics Logical Function (AnLF) to use trained Machine Learning (ML) Model Provisioning Services*** **from another NWDAF containing** ***Model Training Logical Function (MTLF).***

NOTE 2: Analytics Logical Function (AnLF) and Model Training Logical Function (MTLF) are described further below.

The *NWDAF* provides *Analytics* to *5G Core (5GC) NFs* and *OAM* as defined.

An ***NWDAF*** may contain the following ***Logical Functions***:

- ***Analytics logical function (AnLF)***: A *Logical Function in NWDAF*, which performs *inference,* derives analytics information (i.e. *derives statistics and/or predictions* based on *Analytics "Consumer" Request*) and exposes Analytics Service i.e. *Nnwdaf_AnalyticsSubscription* or *Nnwdaf_AnalyticsInfo.*

- *Model Training Logical Function* (***MTLF***): A *Logical Function in NWDAF,* which trains *Machine Learning (**ML**) Models* and exposes New Training Services (e.g. providing Trained ML Model) as defined in this Architecture specification.

**NOTE 1:** *NWDAF* can contain an *MTLF or an AnLF or both Logical Functions (LFs).*

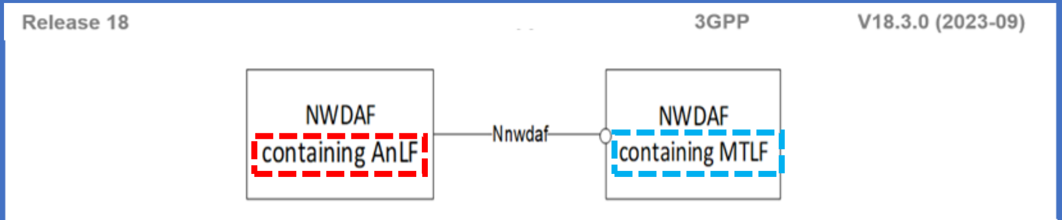


**Figure: 5G System Data Analytics Collection and Reporting Architecture using Trained Machine Learning (ML) Analytics Logical Function (LF) and Model Training Logical Function (MTLF)**

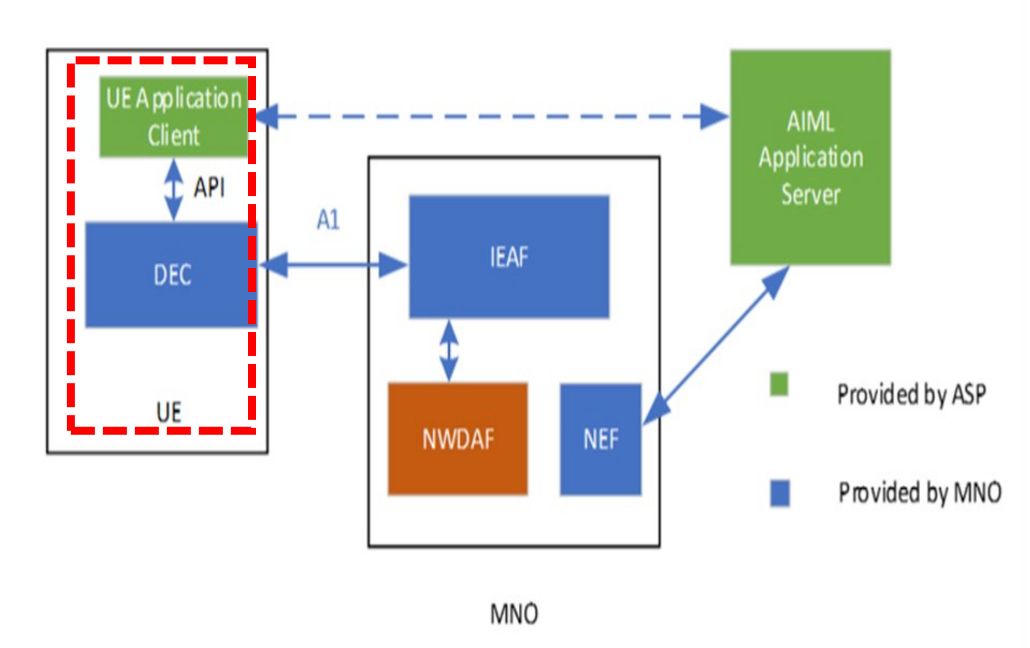


**Figure: 5G Data Analytics, Collection and Reporting Architecture evolved IEAF Data Information Collection Function**

 5G System Data Collection and Analytics Reference Architecture - ***use of AI/ML***

*UE ID retrieval - IEAF based solution*

The following information may be *requested by UE application Client from 5GC to assist the Application layer AIML operation:*

*- QoS Sustainability Analytics.*
*- User Data Congestion Analytics.*

Note: Whether and how the UE can use 5GC information (e.g. as above) for AI/ML operations is FFS and needs to be described with valid justification before solution can be adopted, considering also that the same information will be used by the AI/ML application server as well.

NOTE x:Support for analytics IDs that only support any UE as the target of analytics reporting is subject to SA WG3 evaluation on how to address security and privacy concerns when sharing analytics generated from other UEs to an individual UE.
The UE Data Exposure Client (DEC) is responsible for sending data request to the Data Information AF (IEAF) to collect data from NWDAF as an input for application layer AIML operation. The IEAF is always in the MNO domain and the DEC is based on 3GPP defined procedures and security and therefore is also under the control of MNO. The data collection request from UE Application may trigger the IEAF to collect Data from NWDAF.

NOTE 1: Both IEAF and DEC are controlled and managed by the MNO e.g. with 3GPP defined procedures.

The IEAF is configured based on the SLA above for each AI/ML Application. NWDAF follows existing Service User Consent checks as specified in 5G and Network Consent checks for the IEAF (as a NWDAF Service Consumer).
The IEAF may be also configured by the operator to do some data processing before sending the exposure data to DEC.
The following information are pre-configured in the UE by MNO or provisioned (via PCF) to the UE as part of AIML policy by using the procedure as defined in 5GS Procedures and used in the communication with IEAF:

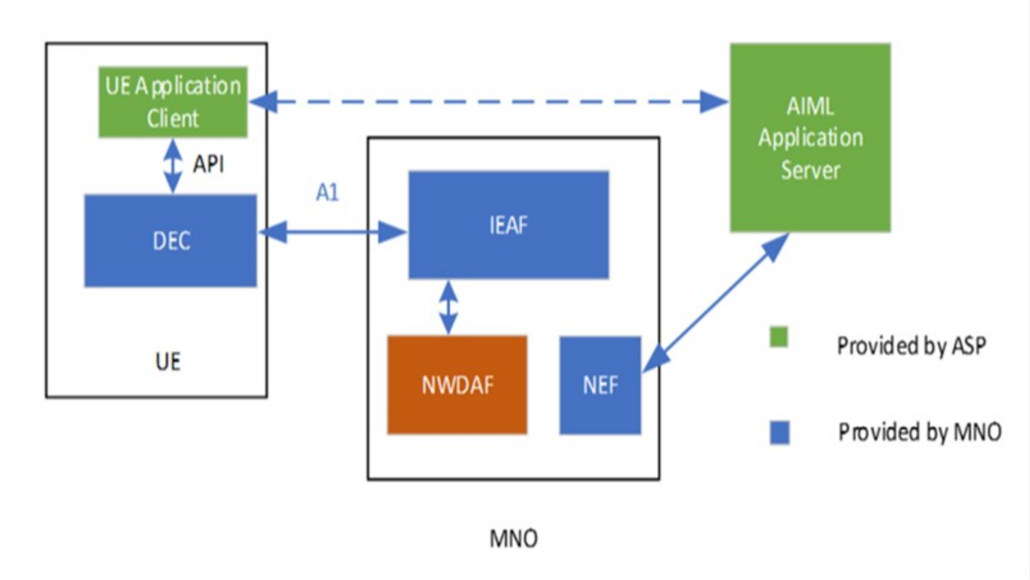



**Figure: 5G Data Analytics, Collection and Reporting Architecture evolved IEAF Data Information Collection Function**

The **DEC** communicates to the IEAF over User Plane (UP) via a PDU session established by the UE.

*NOTE 2:* ***The DEC is deployed per Application in this Release.***

The SLA between the Operator and the AIML Application Service Provider (SP) determines per Application ID in use by the ASP:

- **The Analytics ID(s) that the 5GC is allowed to expose, subject to User Consent and Network Consent.**

- The S-NSSAI for the AIML Application Service Provider (SP).
- The Authentication information that enable the IEAF to verify the authenticity of the DEC that collects data.

# 3. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - 9

*5GS Analytics and Data Reporting Reference Architecture Determining ML Model drift for improving Analytics accuracy*

*The Accuracy of Analytic Output from an NWDAF depends very much on the Accuracy of the ML Model provided by the MTLF NWDAF.*

The Training Data that are used to train an *ML Model are usually Historical Data (Data stored in the Analytics Data Repository Function (ADRF))*.

The **Validity/Accuracy of the ML Model** depends on *whether the Training Data used are "up to date" with the Real-Time Network configuration/ behaviour.*

E.g. Compared to When the Training Data were collected the Network Operator may configure *additional Network Resources to a Network Slice*, or the *Number of Users Accessing Services* via the *Core Network (**CN**)* may considerably increase *(e.g. Tourist Season in the Summer).*

Such UC may cause a "*Model drift"* given that ML Model was not trained with *Up-to-Date Data*.

There are many reasons that "*ML Model drift*" can occur but the *main cause is a change of the Data with time.*

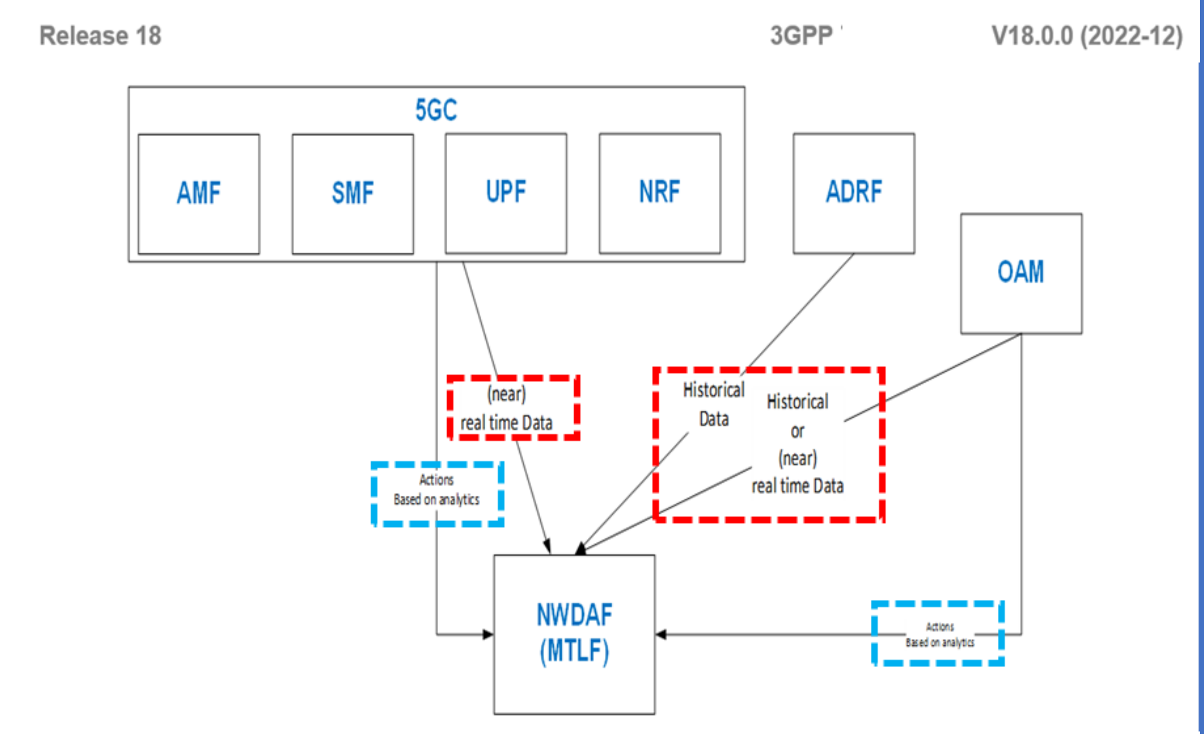



**Figure: 5G System Data Analytics Collection and Reporting Architecture Model drift detected at Network Data Analytics Function (NWDAF) Model Training Logical Function (MTLF)**

A "simple" Solution to this problem is to *Re-Train an ML Model Periodically*. Such approach will ensure that the *NWDAF always uses an "Up-to-Date Training Data" for an ML Model.* However, such approach requires *"considerable" Resources and is not energy efficient*.

Hence a Solution is required to allow the *Network (i.e. NWDAF*) to determine when an *ML Model requires Re-Training.*

The Solution proposed hereby focuses on the *NWDAF* to evaluate if an action taken by a "*Consumer"* would result in a Model drift and then evaluate if the *Training Data* are *"Up-to-Date".*

## Summary-1 of 5G Advanced implementation of AI/ML Applications and ML Model Transfer Capabilities

In 5G, AI/ML is specified to be used in a range of Application Domains across Industry sectors. In 5G Mobile Communications Systems, Mobile Devices (e.g. Smartphones, Automotive, Robots) are increasingly replacing conventional algorithms (e.g. Speech Recognition, Image Recognition, Video Processing) with AI/ML Models to enable Applications. **The 5G System (5GS) can at least support three (3) types of AI/ML operations**: ***1. The UE Data Exposure Client (DEC)*** is responsible for sending *Data request to the Data Information AF* (***IEAF,*** evolved Rel. 17 ***DCAF/AF)*** to collect Data from **NWDAF** as an input for **Application Layer AIML operation.** The **IEAF** is always in the MNO Domain & the **DEC** is based on **3GPP defined Procedures & Security &** *therefore is also under the control of MNO.* The Data Collection Request from UE Application may trigger the **IEAF** to collect Data from **NWDAF (IEAF deployment shown below).** ***2. AI/ML Model/Data Distribution & Sharing over 5GS*** (the Model Performance at the UE needs to be monitored constantly). ***3. Distributed/Federated Learning (FL) over 5GS*** (The Cloud Server trains a Global Model by aggregating Local Models partially-trained by each End Device via 5G UL). The Server aggregates the Interim Training results from the UEs & updates the Global Model. The Updated Global Model is then distributed back to the UEs & the UEs can perform the Training for the Next Iteration. Based on Operator Policy, 5GS shall be able to provide means to predict & expose predicted Network Condition changes (i.e. Bitrate, Latency, Reliability) per UE, to an Authorized 3rd Party. **Subject to User Consent, Operator Policy & Regulatory Constraints**, the 5GS shall be able to support **a Mechanism** to expose Monitoring & Status Information of an AI-ML Session to a 3rd Party AI/ML Application & be able to expose information (e.g. candidate UEs) to an Authorized 3rd Party to assist the 3rd Party to determine Member(s) of a Group of UEs (e.g. UEs of a FL Group). ***Depending on Local Policy or Regulations, to protect the Privacy of User Data, the Data Collection, ML Model Training & Analytics generation for a Subscriber/User id, Internal or External_Group_Id or "any UE" may be subject to User Consent*** **bound to a Purpose, such as Analytics or ML Model Training. The User Consent is "Subscription Information**" stored in the 5G CN, which includes: **A)** whether the User authorizes the Collection & Usage of its Data for a Particular Purpose; **B)** the **Purpose** for Data Collection, e.g. **Analytic or Model Training.**

**5GS (System)** proposes a Common **Solution Framework** to assist various Application AI/ML Operations with Assistance Info & Procedures from 5GC. In this Framework, the similar **Service Requirements & Operational behaviours** are organized into various *Application AI/ML Assistance* (**AaaML**) *Service Profiles* where *Each Profile defines specific AaaML Service*. The **AaaML Services** are a Set of Collective Extensions to the existing 5GC Services & the new 5GC Services which are defined specifically to assist the Application Layer AI/ML Service Operation. An **AaaML Service Profile** is composed of 3 main parts of information: A) **Objective** of Target AaaML Operation; **B) Input of Provisioned Service Parameter(s)** ( e.g. Minimum One Way Delay, Predicted QoS Performance within the next 5 min.; **C) Output** (*e.g. List of Candidate UEs, Event Report for the Group of UE's Bandwidth Consumption.*

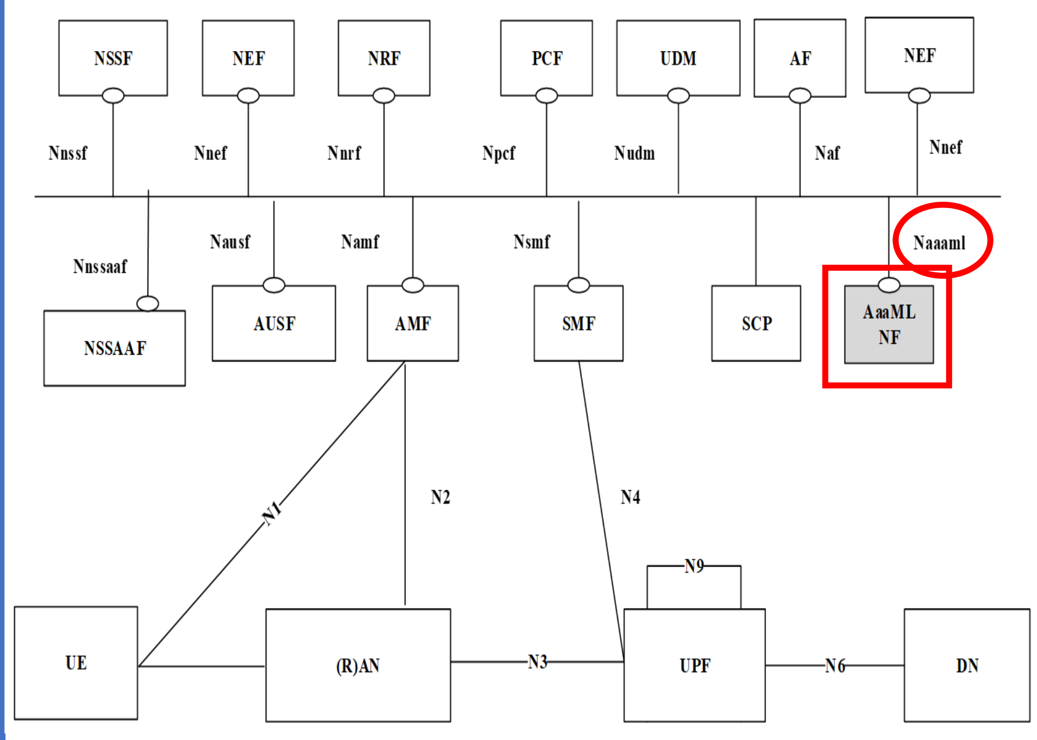



**Figure: 5G System Service Architecture with AaaML NF**

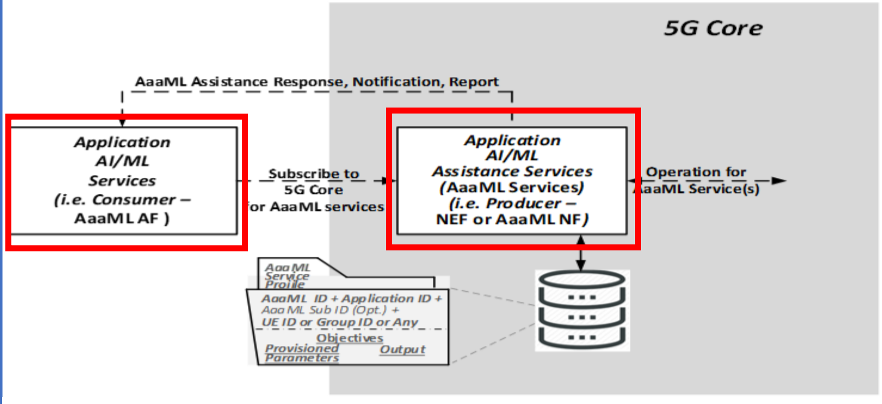



**Figure: 5G Application AI/ML Service Assistance Framework**

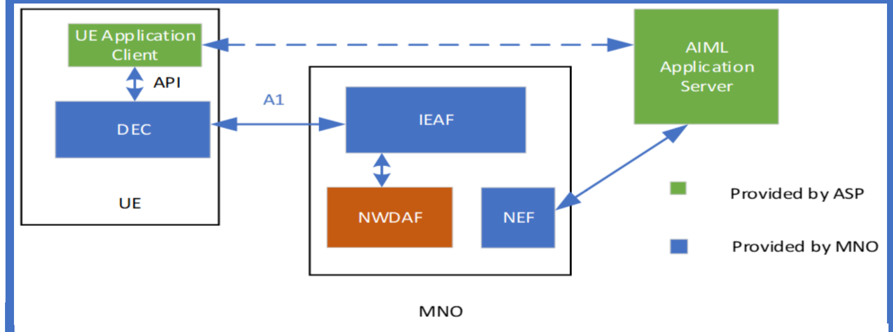



**Figure: 5G IEAF (Data Information AF)**

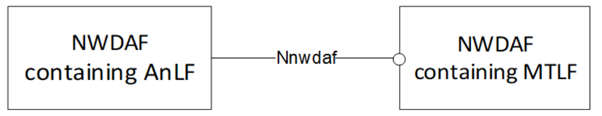



**Figure: 5G NWDAF containing AnLF (Analytics Logical Function) for request and subscribe to ML Model Provisioning Architecture**

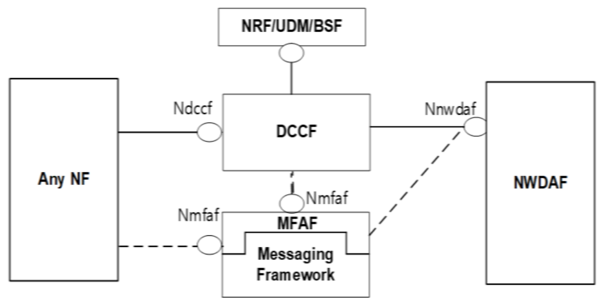



**Figure: 5G Network Data Analytics Exposure Architecture using DCCF**

**Table: 5G NFs Consumed by DCCF or NWDAF to determine which NF instances are serving UE**

| Type of NF instance (serving the UE) to determine | NF to be contacted by DCCF | Service |
|---|---|---|
| UDM | NRF | Nnrf_NFDiscovery |
| AMF | UDM | Nudm_UECM |
| SMF | UDM | Nudm_UECM |
| BSF | NRF | Nnrf_NFDiscovery |
| PCF | BSF | Nbsf_Management |
| NEF | NRF | Nnrf_NFDiscovery |
| NWDAF | UDM | Nudm_UECM |

## Summary-2: 5G Advanced UE ID retrieval IEAF Data Information Collection based Solution with UE DEC (Data Exposure Client)

In 5G, ***UE DEC (Data Exposure Client) Application Client*** may request from 5GC to assist the *Application Layer AI/ML Operation* with information about ***QoS Sustainability Analytics & User Data*** Congestion Analytics. The UE Data Exposure Client (**DEC**) is responsible for sending Data request to *the Data Information AF* (***IEAF***) to collect Data from NWDAF as an input for Application Layer AIML Operation. The IEAF is always in the MNO Domain & the **DEC** is based on 3GPP defined Procedures & Security & therefore is also under the control of MNO. The Data collection request from UE Application may trigger the IEAF to collect Data from NWDAF. Both IEAF & DEC are controlled and managed by the MNO e.g. with 3GPP defined procedures. The DEC communicates to the IEAF over User Plane (UP) via a PDU session established by the UE. The DEC is deployed per Application. The ***SLA** between the Operator & the AIML Application Service Provider (**ASP***) determines per Application ID in use by the ASP such as 1) the Analytics ID(s) that the 5GC is allowed to expose, subject to User Consent & Network Consent, 2) the S-NSSAI for the AIML Application Service Provider (ASP), 3***) the Authentication*** information that *enable the IEAF to verify the Authenticity of the DEC that collects Data.* The 5G System Architecture allows ***ADRF (Analytics Data Repository Function***) to store and retrieve the Collected Data & Analytics.

Based on the NF Request or Configuration on the ***DCCF***, the ***DCCF*** may determine the ***ADRF*** & interact directly or indirectly with the ADRF to request or store Data. A ***Consumer NF*** may specify in requests to a ***DCCF*** that ***Data** provided by a Data Source needs to be stored in the ADRF*. The ***ADRF*** checks if the ***Data Consumer*** is authorized to access ADRF Services & provides the requested Data using the Procedures 5G System specified Procedures.

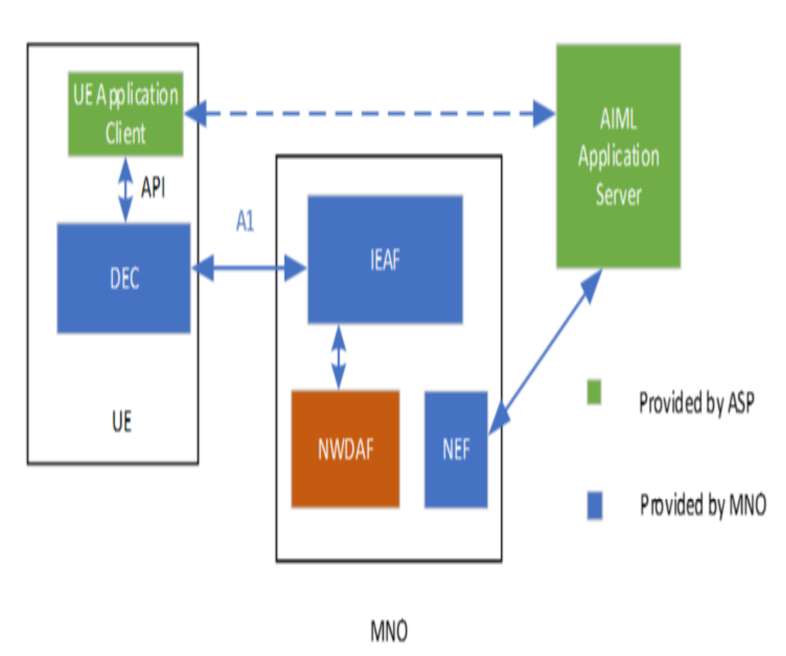



**Figure: 5G IEAF Data Information Collection**

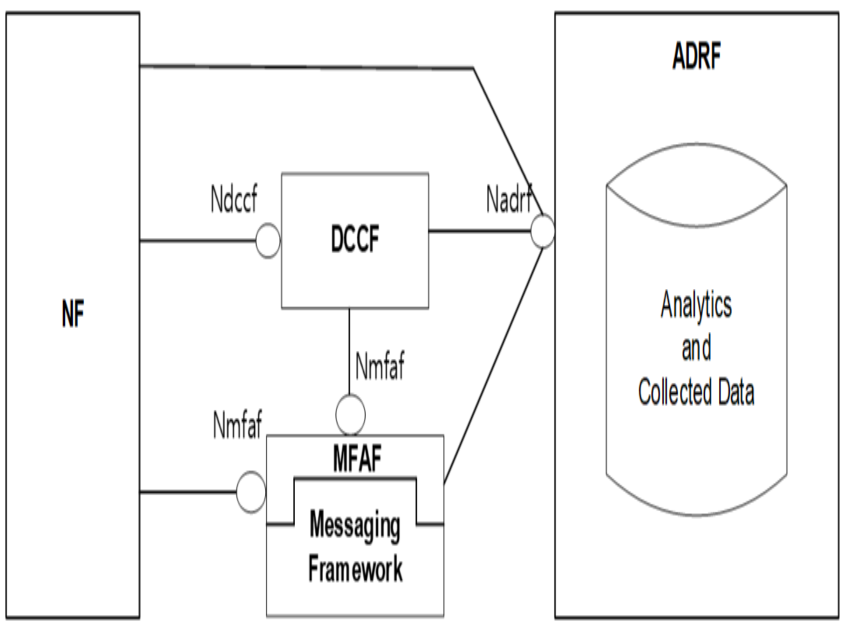



**Figure: 5G Data Storage for Analytics and Collected Data**

**Table: 5G KPI Table of AI/ML Inference Split between UE and Network Server/AF**

| Uplink KPI | | | | | Downlink KPI | | | | |
|---|---|---|---|---|---|---|---|---|---|
| Max allowed UL end-to-end latency | Experienced data rate | Payload size | Communication service availability | Reliability | Max allowed DL end-to-end latency | Experienced data rate | Payload size | Reliability | Remarks |
| 2 ms | 1.08 Gbit/s | 0.27 MByte | 99.999 % | 99.9 % | | | | 99.999 % | Split AI/ML image recognition |
| 100 ms | 1.5 Mbit/s | | | | 100 ms | 150 Mbit/s | 1.5 MByte/ frame | | Enhanced media recognition |
| | 4.7 Mbit/s | | | | 12 ms | 320 Mbit/s | 40 kByte | | Split control for robotics |
| NOTE 1: | Communication service availability relates to the service interfaces, and reliability relates to a given system entity. One or more retransmissions of network layer packets can take place in order to satisfy the reliability requirement. | | | | | | | | |

**Table: 5G KPI Table of Federated Learning (FL) between UE and Network Server/AF**

| Max allowed DL or UL end-to-end latency | DL experienced data rate | UL experienced data rate | DL packet size | UL packet size | Communication service availability | Remarks |
|---|---|---|---|---|---|---|
| 1s | 1.0Gbit/s | 1.0Gbit/s | 132MByte | 132MByte | | Uncompressed Federated Learning for image recognition |
| 1s | 80.88Mbit/s | 80.88Mbit/s | 10Mbyte | 10Mbyte | TBD | Compressed Federated Learning for image/video processing |
| 1s | TBD | TBD | 10MByte | 10MByte | | Data Transfer Disturbance in Multi-agent multi-device ML Operations |

# 3. 5G System use of AI/ML

Artificial Intelligence/Machine Learning (AI/ML) Techniques are being embraced by Telecommunication Service Providers (SPs) around the World to facilitate enabling the existing and the new challenging UCs that 5G offers.

AI/ML Capabilities are being increasingly adopted in Mobile Networks as a Key Enabler for wide range of Features and Functionalities that maximise efficiency and bring Intelligence and Automation in various Domains of the 5GS such as:

- Management Data Analytics (MDA) in the Management & Orchestration Domain
- Network Data Analytics Function (NWDAF) in the 5G Core Network (CN) Domain
- NG-RAN (e.g. RAN Intelligence) defined in 3GPP NG-RAN & NR Domain Specifications)

*The AI/ML Inference Functions in the 5GS use the ML Model for Inference and in order to enable and facilitate the AI/ML adoption, the ML Model needs to be*
*- Created,*
*- Trained and then*
*- Managed during its entire Lifecycle.*

To enable, facilitate and support AI/ML-Capabilities in the 5GS, the following Management Capabilities are in focus under evolvement:

- Validation of ML Model or Entity.
- Testing of ML Model or Entity (before deployment).
- Deployment of ML Model or Entity (New or Updated Model/Entity).
- Configuration of ML Training and AI/ML Inference.
- Performance Evaluation of ML Training and AI/ML Inference.

***NOTE:*** *The ML Model Training Capability is specified in 3GPP AI/ML Management*

AI/ML Techniques Generic Workflow of the Operational Steps in the Lifecycle of an ML Model or Entity, is depicted in the Figure.

Release 18 3GPP V18.0.0 (2023-09)

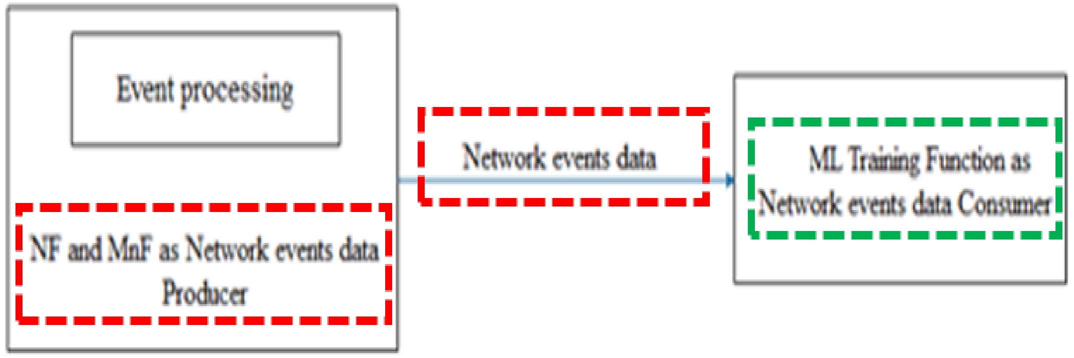



Figure: 5GS (including 5GC, NG-RAN and Management System) Exposing ND Storing Network Events Data

Release 18 3GPP V18.0.0 (2023-09)

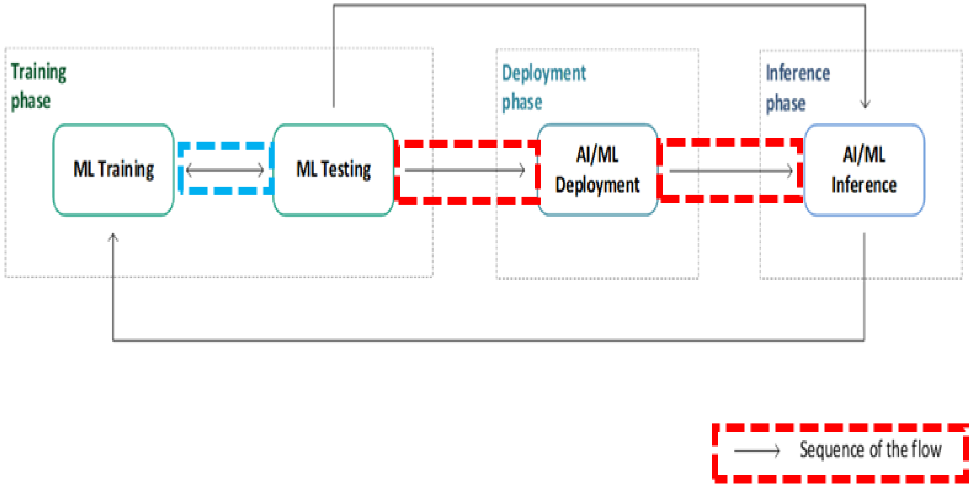



Figure: 5GS (including 5GC, NG-RAN and Management System), and the generic AI/ML Operational Workflow of the Operational Steps in the Lifecycle of an ML Model or Entity

Annex 1: Mobile Networks to evolve from:

# a Design that offers "Best-effort Services

# to

# a Design that offers Performance and User Experience Guarantees

**Capabilities** related to e.g.:

When a *Multi-access* (**MA**) **PDU Session** is established, the Network may provide the UE with *Measurement Assistance Information* to enable the UE in determining which measurements shall be performed over both Accesses, as well as whether measurement reports need to be sent to the Network.

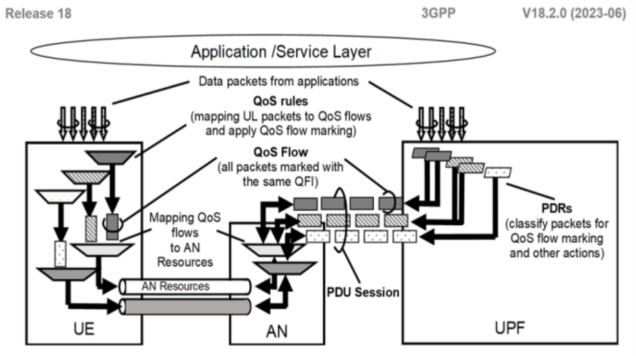


Figure: The 5G Principle for Classification and User Plane (UP) marking for QoS Flows and mapping to AN Resources

Measurement Assistance Information shall include the addressing information of *a* ***Performance Measurement Function*** (**PMF**) **in the UPF, the UE can send PMF protocol messages** incl.:

- Messages to allow for *Round Trip Time* (**RTT**) Measurements: the "*Smallest Delay*" steering mode is used or when either "*Priority-based*", "*Load-Balancing*" or "**Redundant**" steering mode is used with RTT threshold value being applied;
- Messages to allow for *Packet Loss Rate* (**PLR**) measurements, i.e. when steering mode is used either "*Priority-based*", "*Load-Balancing*" or "*Redundant*" steering mode is used with **PLR** threshold value being applied;

- Messages for reporting Access Availability/Un-availability by the UE to the UPF.
- Messages for sending **UE-assistance Data** to **UPF.**
- Messages for sending "*Suspend Traffic Duplication*" and "*Resume Traffic Duplication*" from **UPF** to **UE** to "suspend" or "resume" traffic duplication as defined in **5GS Architecture**.

=>

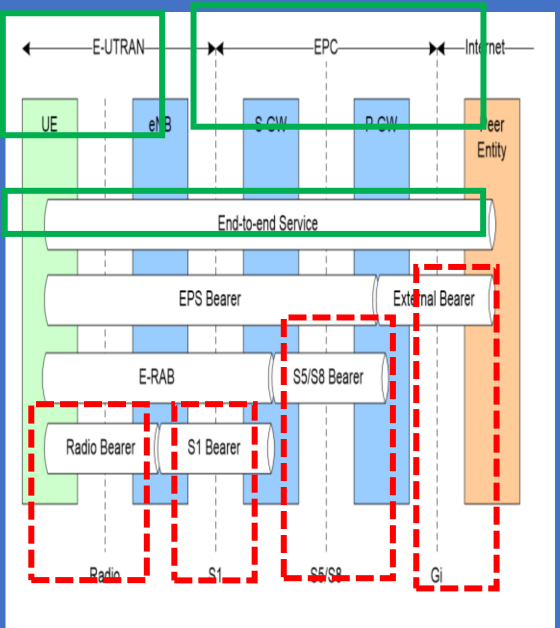


Figure : EPS Bearer Service Architecture

=>

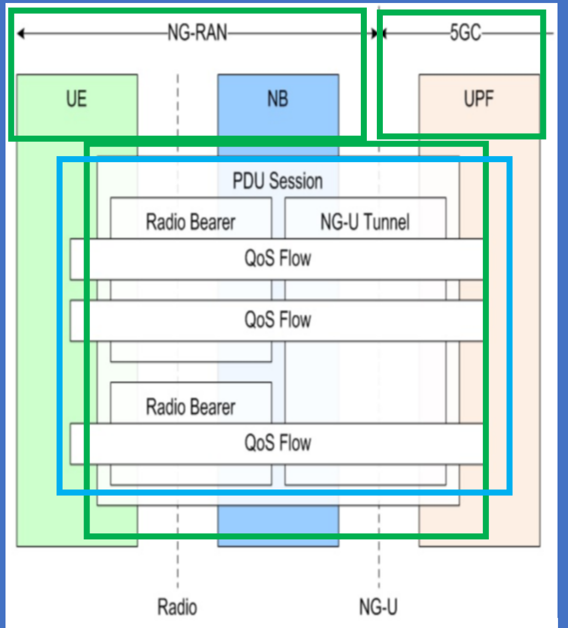


**Figure: 5G CN NG-RAN Bearer Services QoS Architecture**

**Annex 2: 5G Architecture for Hybrid and Multi-Cloud Environments with Telecom "aaS" and DevOps "SaaS" Business Models Difference**

**The Main Challenges to overcome in a Hybrid & Multi-Cloud Strategy** are:

*1. Maintaining Portability; 2. Controlling the Total Cost of Ownership (TCO); 3. Optimizing Productivity & Time to Market (TTM).*

**DevOps** – *a Set of Practices* that brings together *SW Development & IT operations* with the Goal of Shortening the Development & Delivery Cycle & increasing SW Quality **- is** often thought of and discussed **in the Context of a Single Company or Organization. The Company usually Develops the SW, Operates it & Provides it as a Service to Customers,** according to the **SW-as-a-Service (SaaS) Model. Within this context**, it is easier to have **Full Control over the Entire Flow**, including **Full Knowledge of the Target Deployment Environment.**

In the **Telecom Space**, by contrast, we typically follow the **"as-a-Product (aaP) Business model**, in which **SW is developed by Network SW Vendors** e.g. as Ericsson (Nokia, Huawei, ZTE) & provided to Communication Service Providers (CSPs) that Deploy & Operate it within their Network. This **Business Model requires the consideration of additional aspects**.

**The most important contrasts between the Standard DevOps SaaS Model & the Telecom aaP Model** are the **Multiplicity of Deployment Environments & the fact the Network SW Vendor Development Teams cannot know upfront exactly what the Target Environment looks like**. Although a SaaS Company is likely to Deploy & Manage its SW on two (2) or more different Cloud Environments, **this is inevitable within Telco**, as each CSP creates &/or selects its own Cloud infrastructure (Fig. 1 below).

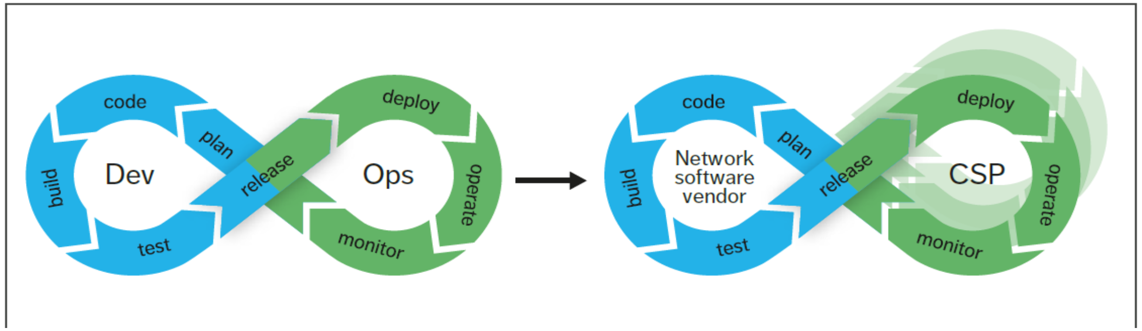



**Figure 1: The DevOps and (Telecom) aaP Business Models**

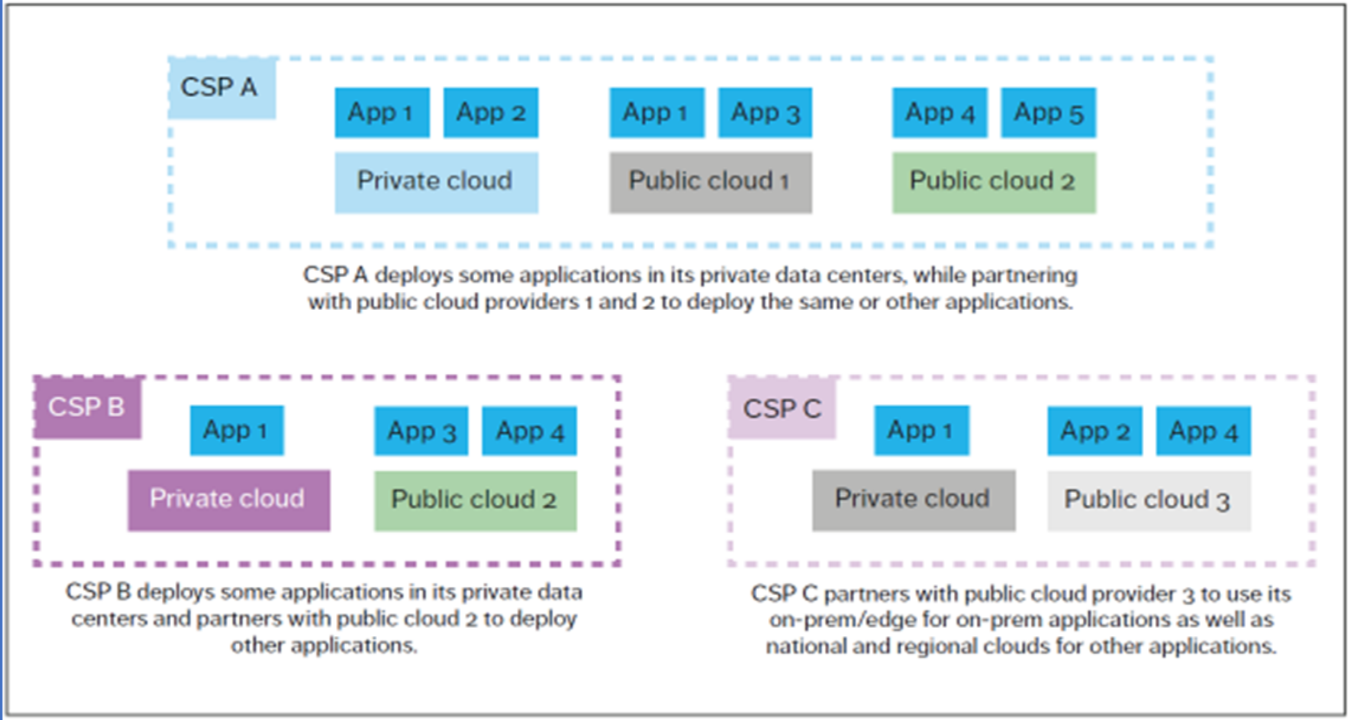



**Figure 2: Examples of Hybrid and Multi-Cloud Deployment Scenarios that Applications must be able to support**

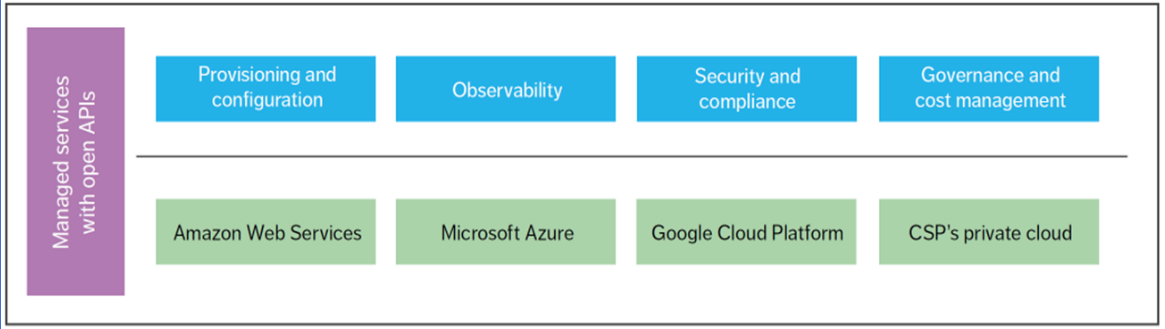



**Figure 3: Key Enablers for a Multi-Cloud Native Application**

# Annex 3 - 5G Personal IoT Networks (PINs) - 1: Definition

Personal IoT Network: A configured and managed group of PIN Element that are able to communicate each other directly or via PIN Elements with Gateway Capability (PEGC), communicate with 5G network via at least one PEGC, and managed by at least one PIN Element with Management Capability (PEMC).

**PIN Element (PINE): A UE or Non-3GPP device** that can communicate within a PIN (via PIN "direct" connection, via PEGC, or via PEGC and 5GC), or outside the PIN via a PEGC and 5GC.

**PIN Element with Gateway Capability:** A PIN Element with the ability to provide connectivity to & from the 5G Network for other PIN Elements, or to provide "relay" for the communication between PIN Elements.

**PIN Element with Management Capability:** A PIN Element with capability to manage the PIN.

**NOTE: A PIN Element can have both PIN Management Capability and Gateway Capability.**

**PINE-to-PINE communication:** communication between two PINEs which may use PINE-to-PINE direct communication or PINE-to-PINE indirect connection.

***PINE-to-PINE direct connection***: the connection between two PIN Elements (PINEs) without PEGC, any 3GPP RAN or core network entity in the middle.

***PINE-to-PINE indirect connection***: the connection between two PIN Elements (PINEs) via PEGC or via UPF.

**PINE-to-PINE routing:** the traffic is routed by a PEGC between two PINEs, the two PINEs direct connect with the PEGC via non-3GPP access.

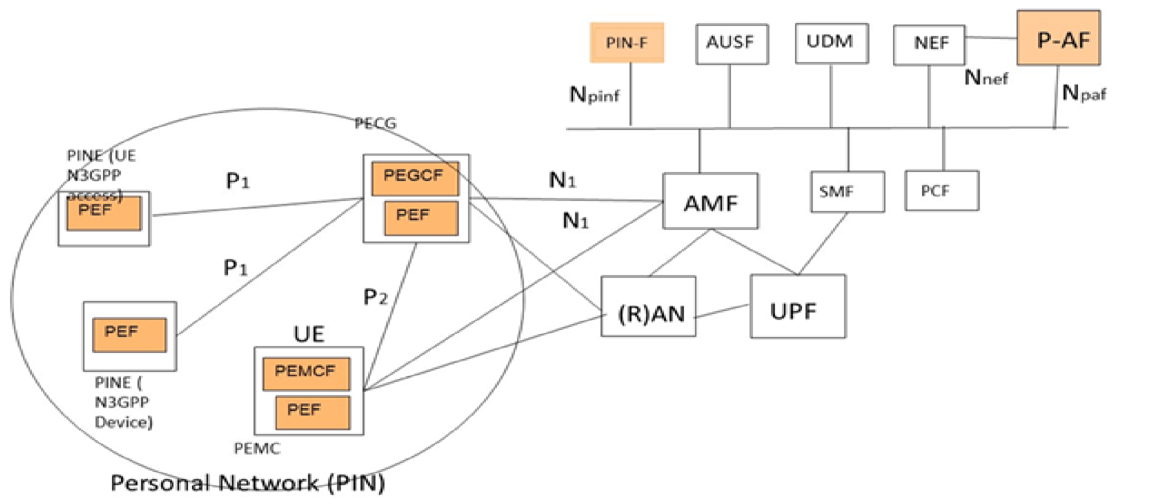
**PINE-to-Network routing:** the traffic is routed by a PEGC between PINE and 5GS, the PINE direct connects with the PEGC via non-3GPP access separately.

**Network local switch for PIN:** the traffic is routed by UPF(s) between two PINEs, the two PINEs direct connect with two PEGCs via non-3GPP access separately.

**Abbreviations**

| | |
|---|---|
| PIN | Personal IoT Networks |
| PINE | PIN Element |
| PEGC | PIN Elements with Gateway Capability |
| PEMC | PIN Elements with Management Capability |
| P2P | PINE-to-PINE |
| P2N | PINE-to-Network |
| NLSP | Network Local Switch for PIN |

*Note 1: The AF relies on PIN signaling between the PINE/PEGC/PEMC and the PIN AF, which is transferred via UP transparently to the 5G System, to determine the need for a QoS modification.*




**5G System PIN Solution Reference Architecture**

## **Annex 3 - 5G System (5GS) enhancements** to support **Personal IoT Networks (PINs) -2.**

- Management of PIN,
- Access of PIN via PIN Element (PINE) with Gateway Capability (PEGC), and
- Communication of PIN (e.g. PINE (e.g. a UE) communicates with
  - other PINE (UE) "directly" or
  - via PEGC or
  - via PEGC and 5GS.

- Security related when identifying PIN and the PINE when:
  - How to identify PIN and the PINEs in the PIN at 5GC level to serve for Authentication& Authorization
  - Management as well as Policy and Routing Control enforcement:

- Management of a PIN.
- PIN & PINE Discovery

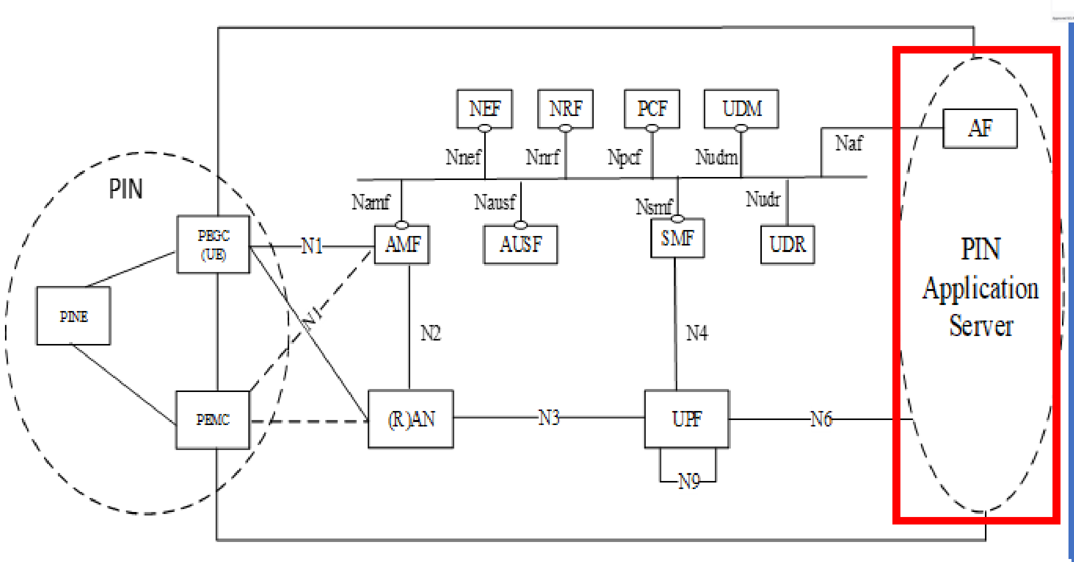



**Figure: 5GS PIN Personal IoT Network Reference Architecture**

A **Personal IoT Network (PIN**) in **5GC** consists of:

- 1 (one) or more Devices providing Gateway/Routing Functionality known as **the PIN Element with Gateway Capability (PEGC**), and
- 1 (one) or more Devices providing PIN Management Functionality known as the **PIN Element with Management Capability (PEMC**) to manage the Personal IoT Network; and
- Device(s) called the PIN Elements (PINE). A PINE can be a non-3GPP Device.

*The PIN can also have a PIN Application Server (AS) that includes an AF (Application Function) functionality*.

The AF can be deployed by Mobile Operator or by an Authorized Third (3rd) Party.

When the AF is deployed by 3rd Party, the interworking with 5GS is performed via the NEF.

The PEMC and PEGC communicates with the PIN Application Server (AS) at the Application Layer over the User Plane.

*The PEGC and PEMC can communicate with each other via "Direct" Communication"*

**Only a 3GPP UE can act as PEGC and/or PEMC.**

# Annex 3: 5G Personal IoT Networks(PINs) and 5G Customer Premises Networks (CPNs)

Personal IoT Networks (PINs) and Customer Premises Networks (CPNs) provide local connectivity between UEs and/or Non-3GPP Devices.

The CPN via an eRG, or in 5G PINs with PIN Elements (PINEs) via a PIN Element with Gateway Capability (PEGC) can provide access to 5G Network Services for the UEs and/or Non-3GPP Devices on the CPN or PIN.

CPNs and PINs have in common that, in general, they are:
- owned, Installed and/or (at least partially) Configured by a Customer of a Public Network Operator.

**A Customer Premises Network (CPN**) is a Network located within
- a Premises (e.g. a Residence, Office or Shop).
- via an evolved Residential Gateway (eRG), the CPN provides connectivity to the 5G Network. The eRG can be connected to the 5G Core Network via wireline, wireless, or hybrid access.
- A *Premises Radio Access Station* (**PRAS**) is a Base Station installed in a CPN. Through the PRAS, UEs can get Access to the CPN and/or 5G Network Services.

The **PRAS** can be configured to use
- Licensed,
- Unlicensed, or
- Both Frequency bands.

Connectivity between the **eRG** and the **UE**, **non-3GPP Device**, or **PRAS** can use any suitable **Non-3GPP Technology** (e.g. **Ethernet, optical, WLAN).**

***A Personal IoT Network (PIN****) consists of* ***PIN Elements (PINEs)*** *that communicate using PIN*
- *"Direct Connection" or*
- *"Direct Network Connection*

*and is managed locally using a PIN Element (PINE) with Management Capability (PEMC).*

Examples of PINs include Networks of Wearables and Smart Home / Smart Office Equipment.

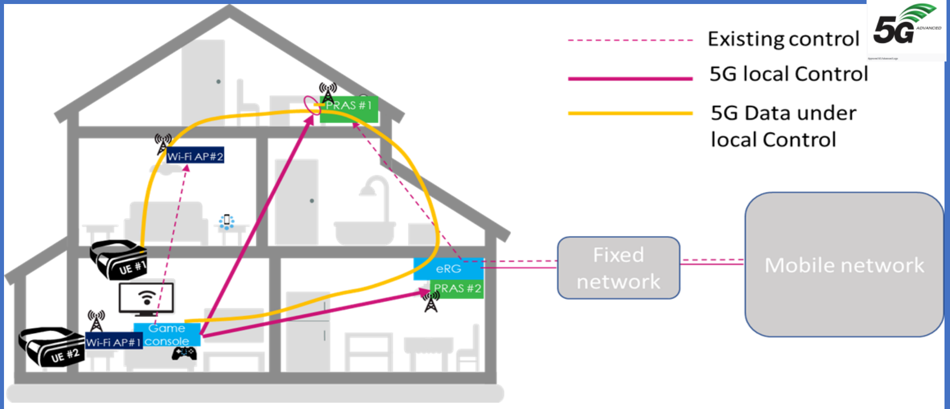



**Figure: 5G Local Control of Premise Radio Access Stations (PRASs) for UE to access CPN Device**

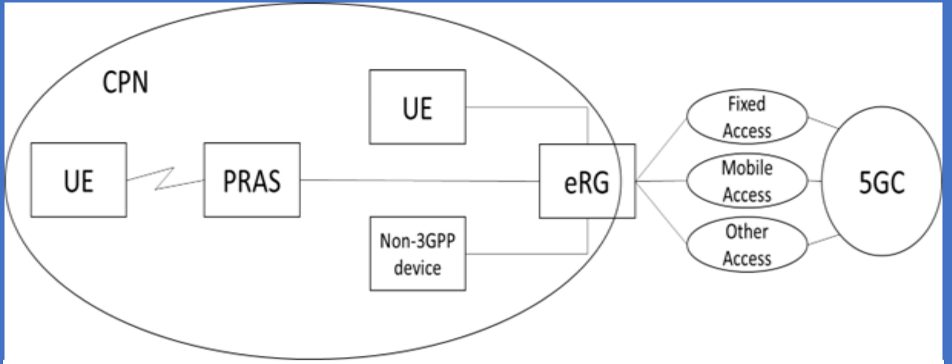



**Figure: Customer Premises Network (CPN) connected to 5GC**

Vodafone unveils Open RAN 5G network-in-a-box

Feb 17, 2023

Vodafone's Yago Tenorio shows off the operator's 5G network-in-a-box.

- Vodafone has unveiled a new mini 5G network the size of a Wi-Fi router
- It has a core and radio software, a mini computer and a software-defined radio chipset
- It is just a prototype currently
- But if offered as a product could revolutionise the 5G private network sector


Remarks & Questions?