5G Advanced

5G Advanced use of AI ML

in

5G Network and Application Layers

with

specified (AI/ML) Management Data Analytics, AI/ML Model Transfer and AI/ML Services KPIs

Ike Alisson

2023 - 11 - 04 Rev PA06

Table of Contents

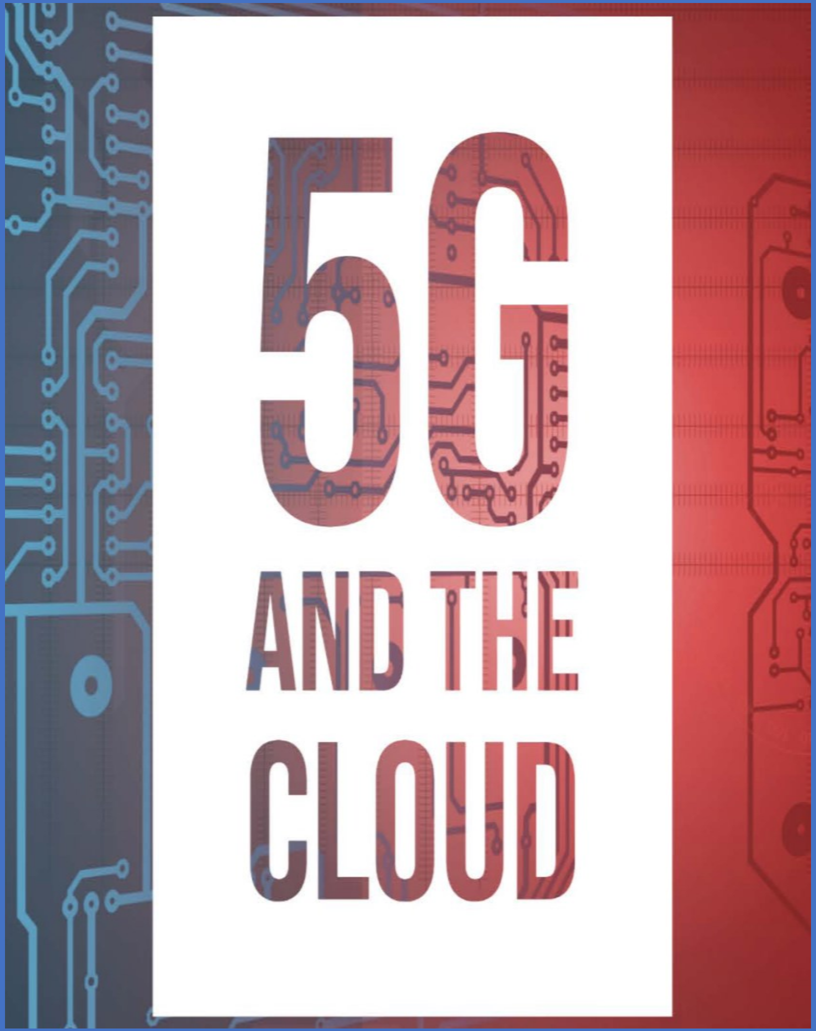
1. Introduction - Cloud and Communication Systems Challenges and Issues

2. 5G System use of AI/M

Annex

1. Shift from 2G/3G/4G "Best-effort" Services to 5G Services with Performance and User Experience Guarantees
2. 5G Architecture for Hybrid and Multi-Cloud Environments with Telecom "aaS" and DevOps "SaaS" Business Models Difference
3. 5G Personal IoT Networks (PINs)

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: - Development, - Delivery and Test, - In-Service Upgrades - Improved Version Management	The Context in which *Cloud Native* was designed is being *misrepresented or abused in two (2) senses:* 1. Cloud Native was *designed for People who write & operate the Applications*. In today's Cellular Network, this clearly is not the case 2. ***Cloud Native* was designed for *Applications in which long interruptions*** *are tolerable*, therefore, *good Reliability is measured* ***in minutes of outage per month.*** 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

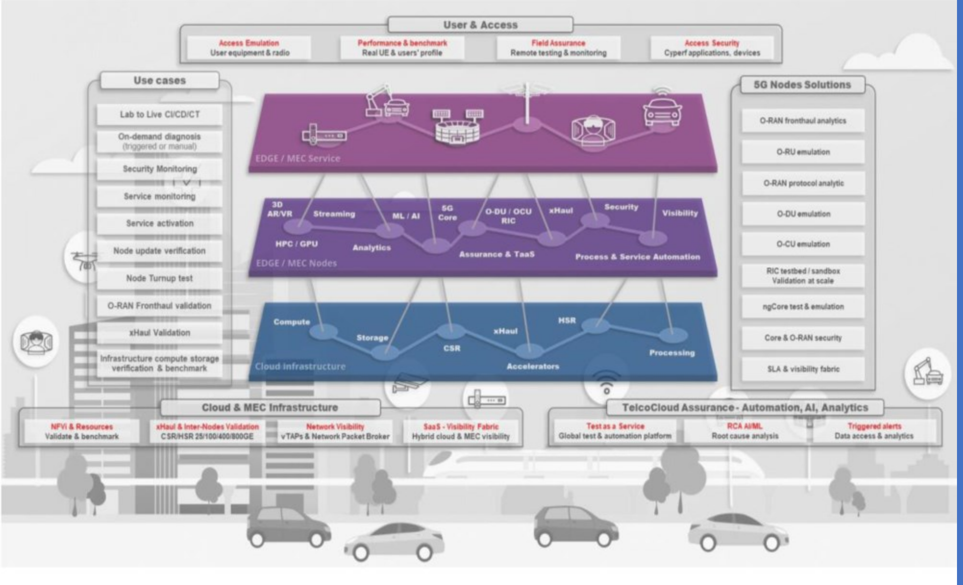


Figure : Telco Edge Cloud, Next-Gen Service Assurance at Scale

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*

Ref.: 5GA, DCC, Nov., 2022: 4

1. The Cloud is "Changing"

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

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.

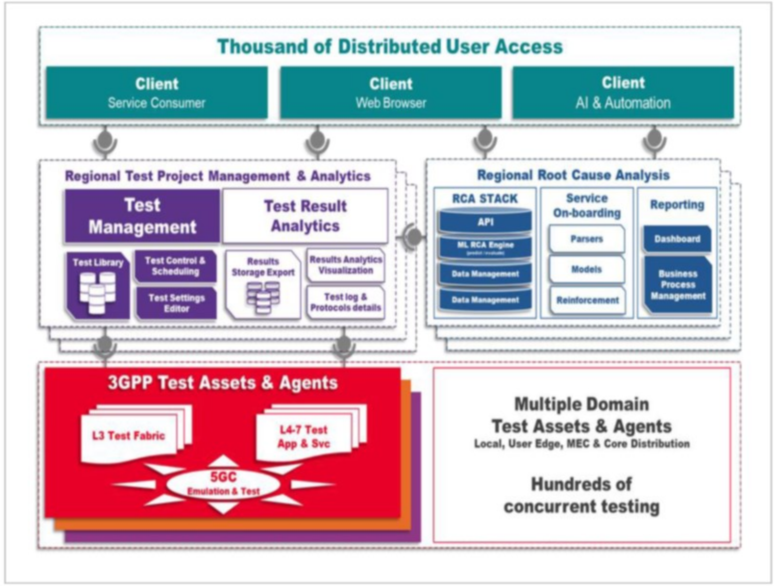

Thousand of Distributed User Access
Client
Service Consumer
Client
Web Browser
Client
AI & Automation
Regional Test Project Management & Analytics
Test Management
Test Result Analytics
Test Library
Test Control & Scheduling
Test Settings Editor
Results Storage Export
Results Analytics Visualization
Test log & Protocols details
Regional Root Cause Analysis
RCA STACK
API
ML RCA Engine
Data Management
Data Management
Service On-boarding
Parsers
Models
Reinforcement
Reporting
Dashboard
Business Process Management
3GPP Test Assets & Agents
L3 Test Fabric
L4-7 Test App & Svc
5GC
Emulation & Test
Multiple Domain Test Assets & Agents
Local, User Edge, MEC & Core Distribution
Hundreds of concurrent testing

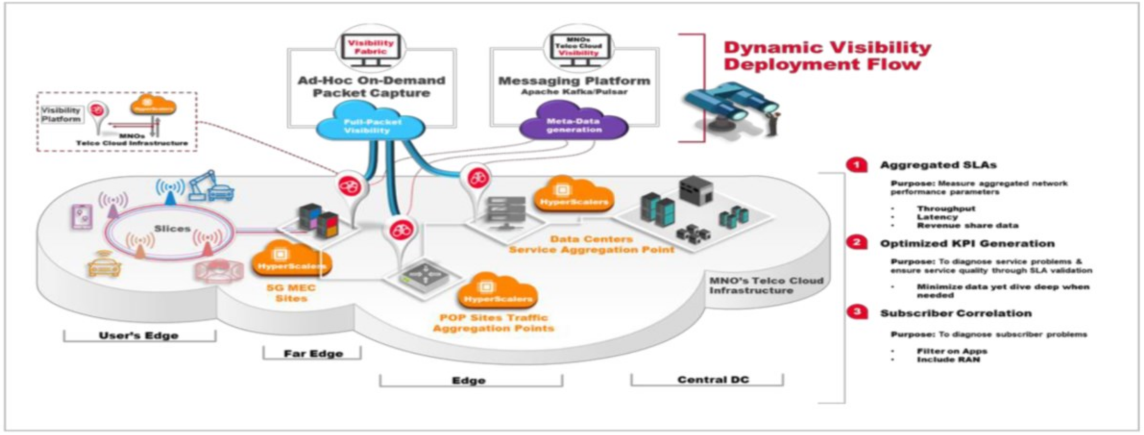
Fig – Nationwide Distributed Cloud based Network Test Bed Orchestration

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

**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.*


Dynamic Visibility Deployment Flow
Visibility Fabric
Ad-Hoc On-Demand Packet Capture
Full-Packet Visibility
MNOs Telco Cloud Visibility
Messaging Platform
Apache Kafka/Pulsar
Meta-Data generation
Visibility Platform
HyperScalers
MNOs Telco Cloud Infrastructure
Slices
HyperScalers
5G MEC Sites
HyperScalers
Data Centers Service Aggregation Point
HyperScalers
POP Sites Traffic Aggregation Points
MNO's Telco Cloud Infrastructure
User's Edge
Far Edge
Edge
Central DC

1 **Aggregated SLAs**

Purpose: Measure aggregated network performance parameters

- **Throughput**
- **Latency**
- **Revenue share data**

2 **Optimized KPI Generation**

Purpose: To diagnose service problems & ensure service quality through SLA validation

- **Minimize data yet dive deep when needed**

3 **Subscriber Correlation**

Purpose: To diagnose subscriber problems

- **Filter on Apps**
- **Include RAN**

Figure : Hybrid Network Visibility Platform architecture

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.

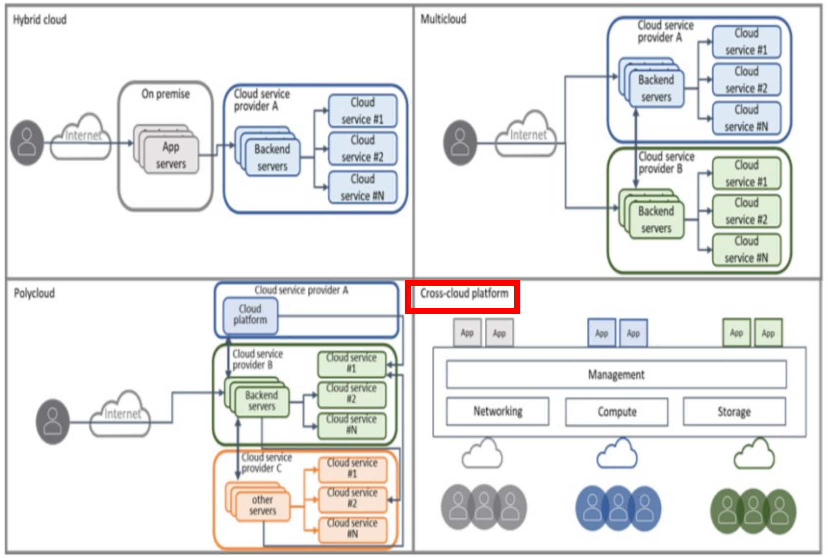
Most existing Compute Systems today, including Edge Computing Systems, rely **on "Static Provisioning"**.

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



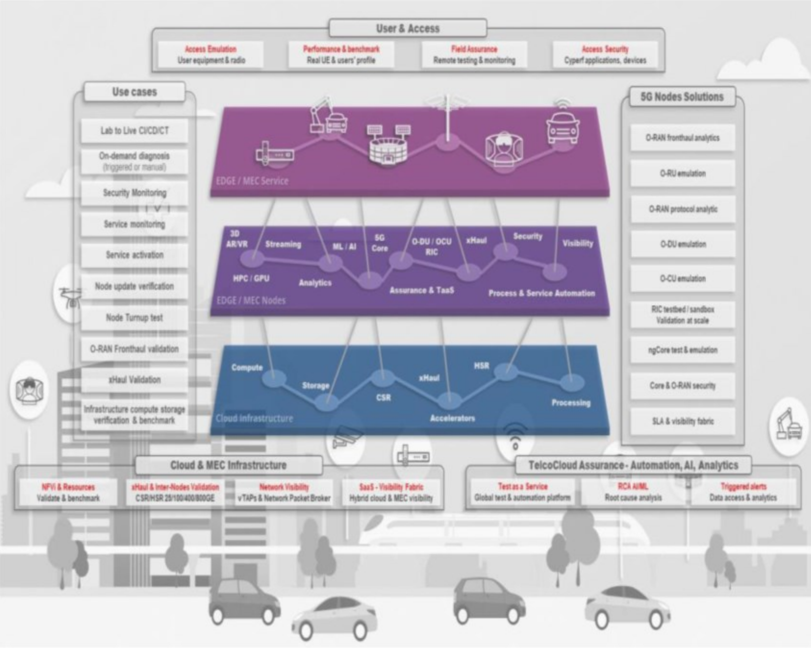


Figure :Telco Edge Cloud, Next-Gen Service Assurance at Scale

2. 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.

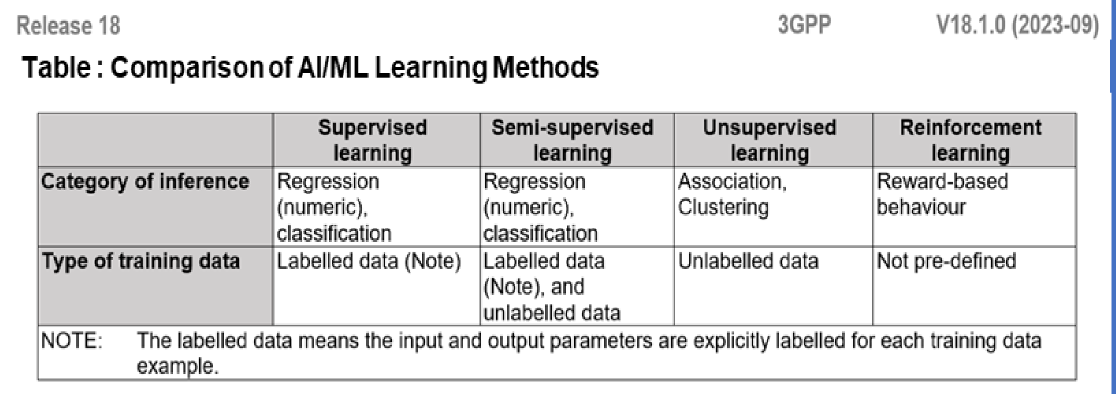
The AI/ML Techniques can be generally characterized from different perspectives including the followings:

- ***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.

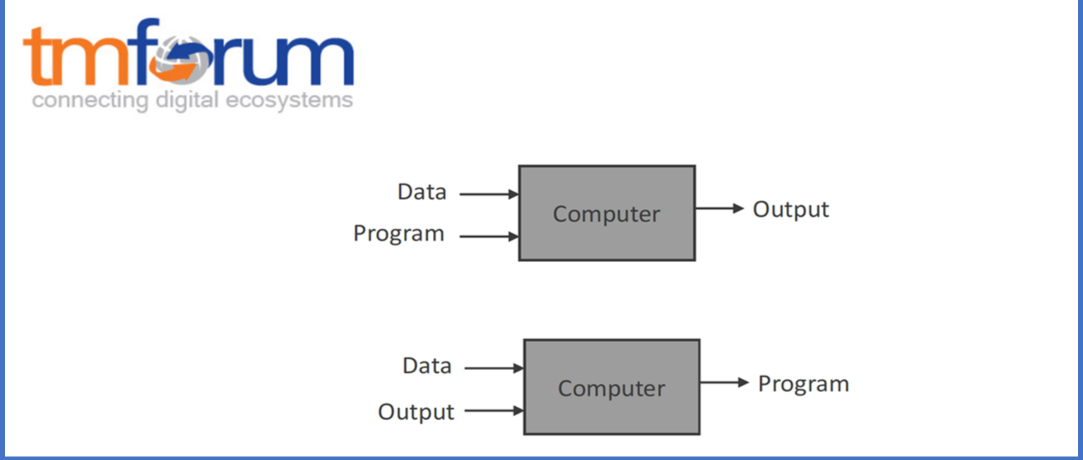
2. 5G System use of AI/ML

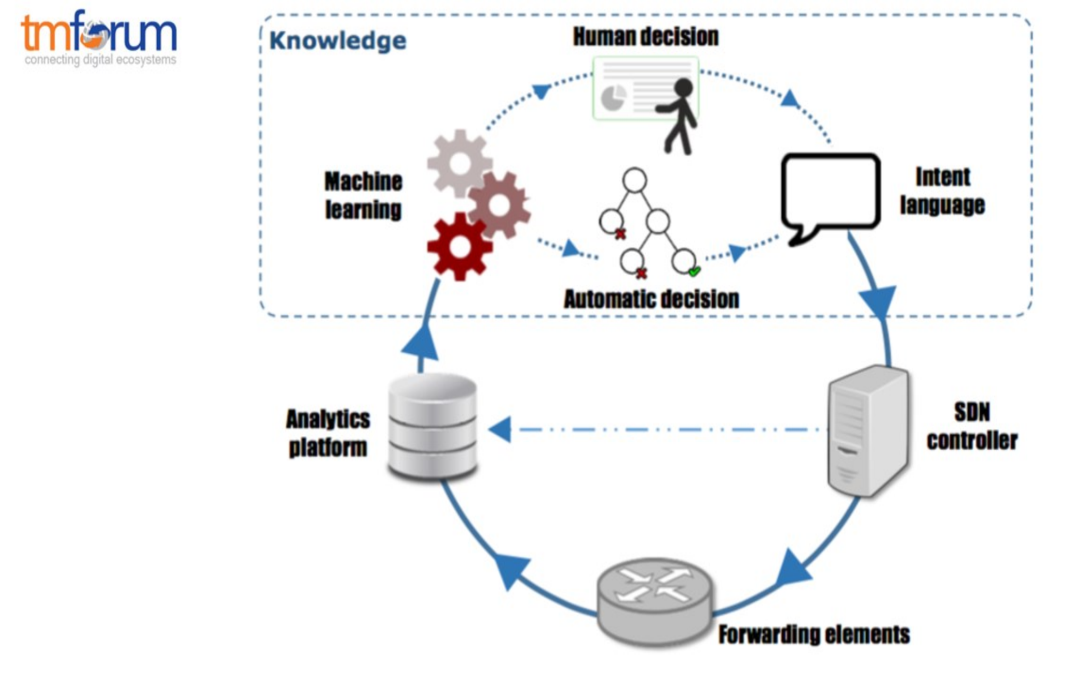
In **Programming**, a Human writes a Computer Program and provides the Data, which the Computer processes to create the Output.

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).

2. 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.

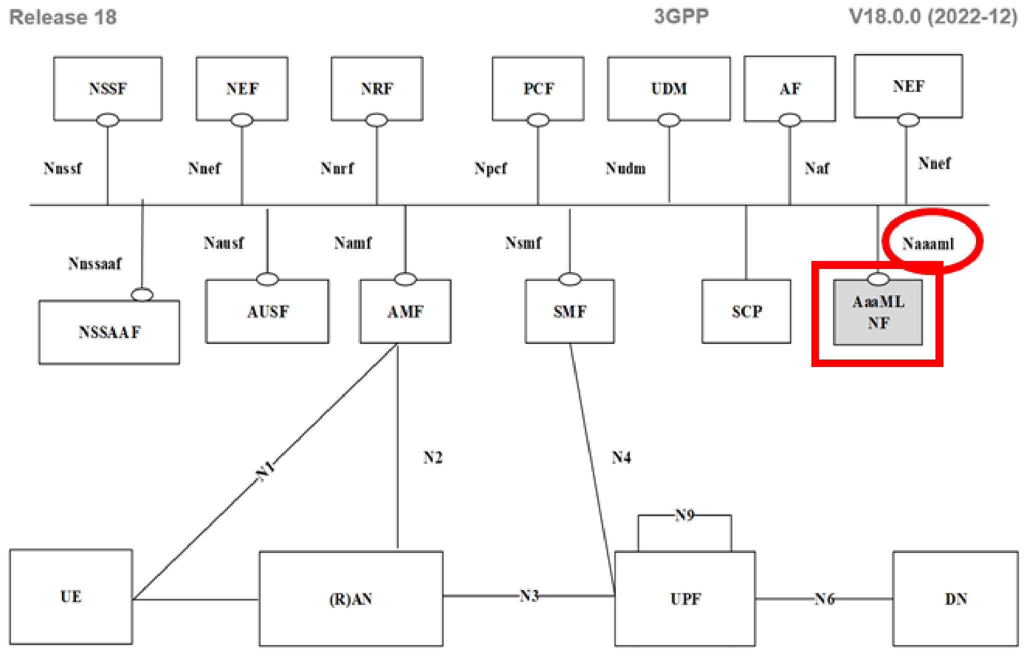


Figure: 5G System Service Architecture with AaaML NF

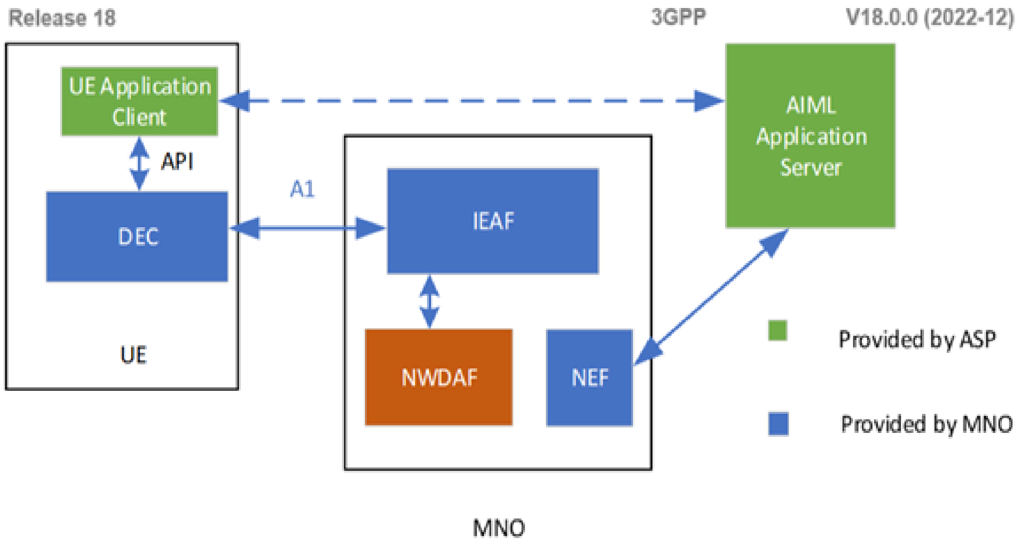


Figure: 5G IEAF (Data Information AF)

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.*

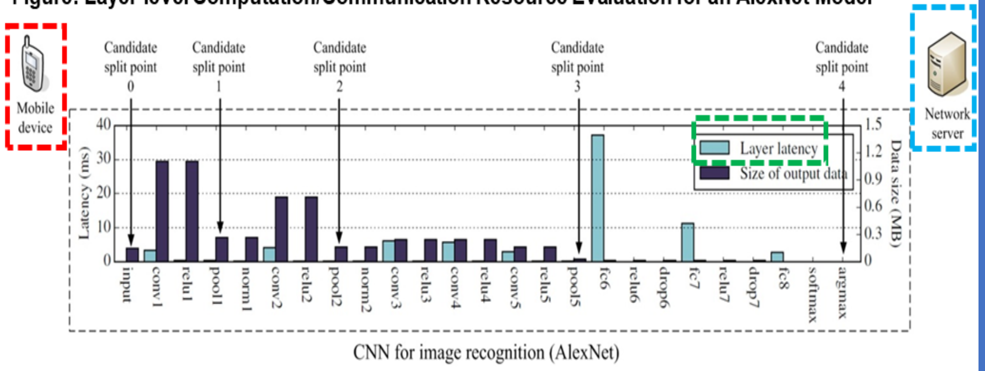
However, sometimes the *Data Rate cannot be increased due to Radio Resource Limitation,* in such circumstances, *UE with Low Computation Capacity* needs to *off-load the Computation Task to a Proximity UE* (likely a ***Relay UE***), but still keeping the Computation Service and let the ***Proximity UE*** 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



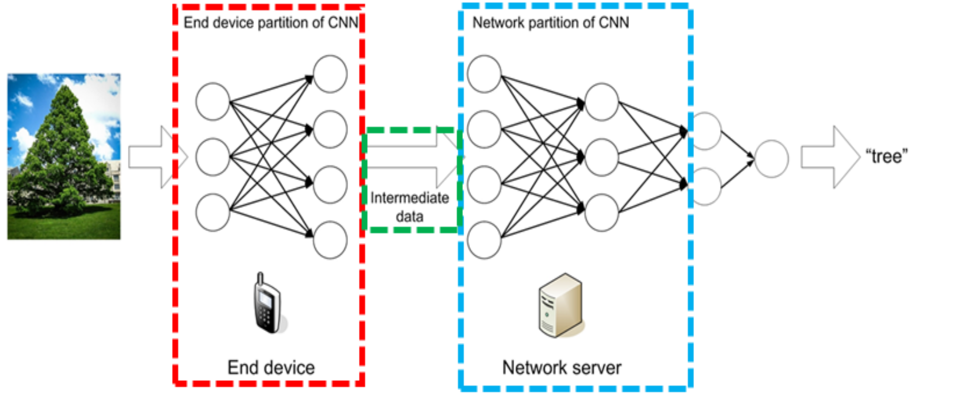
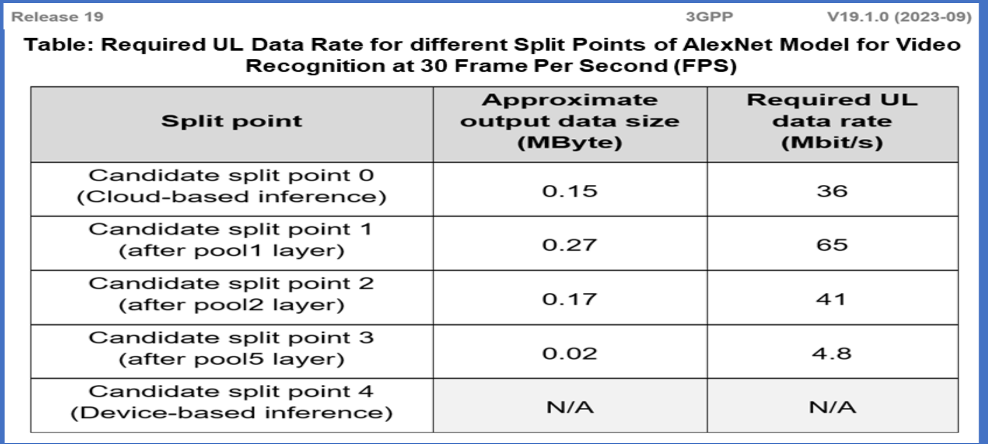


Figure: Example of split AI/ML Image Recognition



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

2. 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.*

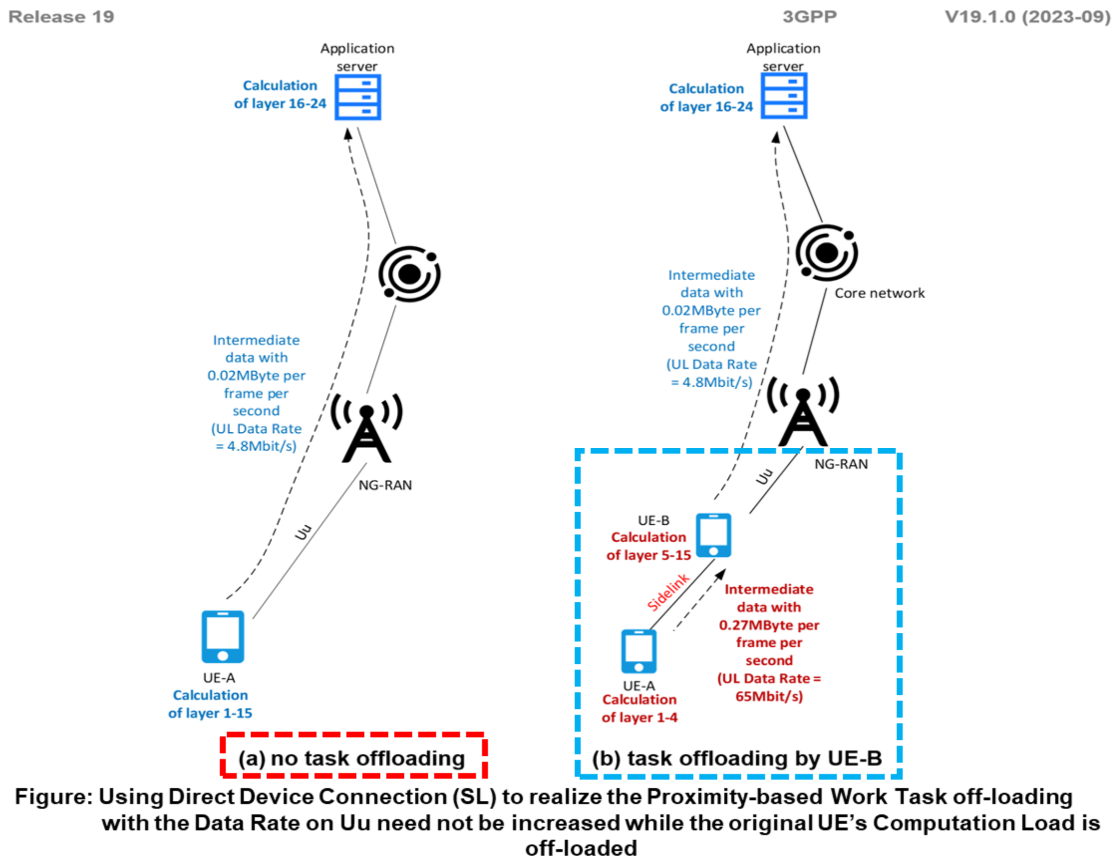
Being managed by 5G Network, the *UE-A discovers UE-B (a Customer Premise Equipment, CPE*) which has installed the same Model and is willing to *take the off.loading task from UE-A*.

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.

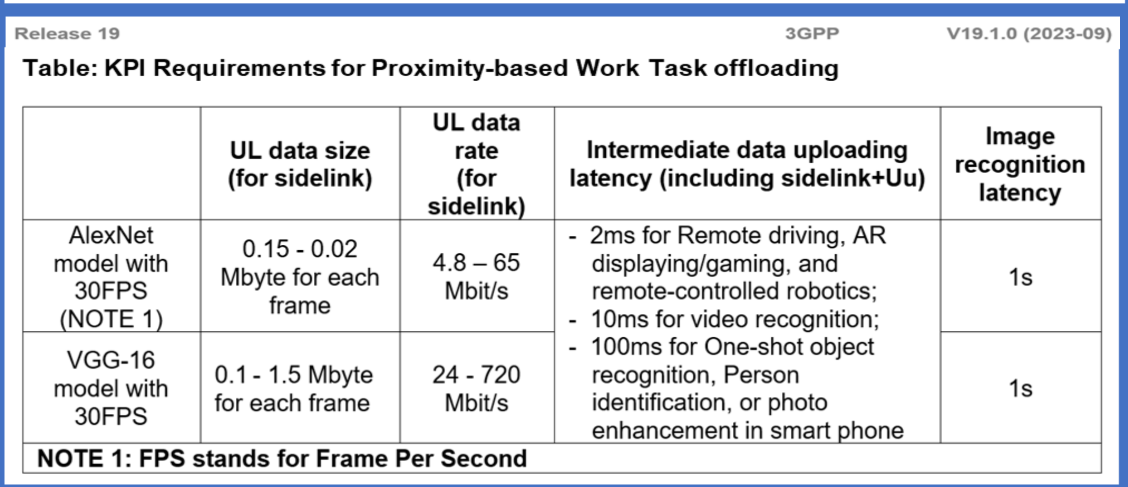
NOTE 2: *It is assumed that the UE-A and UE-B have the same Computation Capacity, i.e. the time used for computing the certain AlexNet Model Layers are the same for UE-A and UE-B. Otherwise, the Data Rate on Uu and Sidelink may be changed accordingly.*



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

Application server
Calculation of layer 16-24
Intermediate data with 0.02MByte per frame per second (UL Data Rate = 4.8Mbit/s)
NG-RAN
Uu
UE-A
Calculation of layer 1-15
(a) no task offloading
Application server
Calculation of layer 16-24
Core network
Intermediate data with 0.02MByte per frame per second (UL Data Rate = 4.8Mbit/s)
NG-RAN
Uu
UE-B
Calculation of layer 5-15
Sidelink
Intermediate data with 0.27MByte per frame per second (UL Data Rate = 65Mbit/s)
UE-A
Calculation of layer 1-4
(b) task offloading by UE-B

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



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

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; - 10ms for video recognition; - 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				

2. 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***.

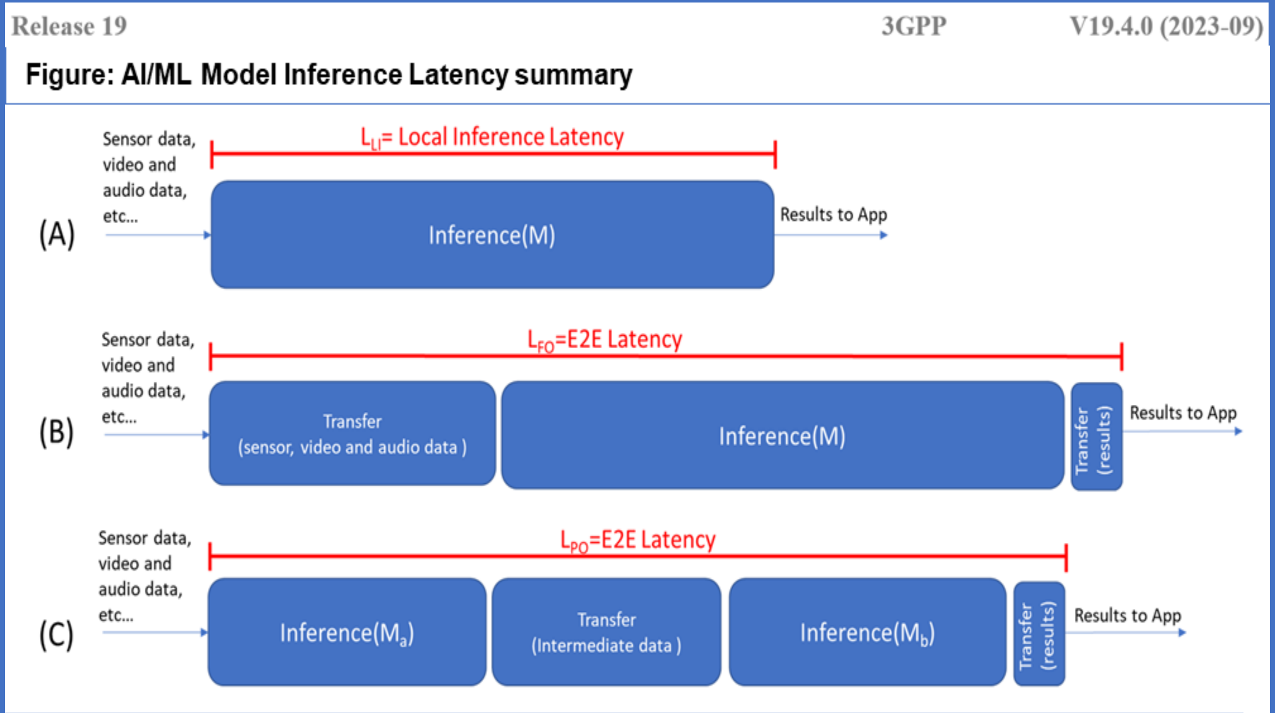

Release 19 3GPP V19.4.0 (2023-09)

Figure: AI/ML Model Inference Latency summary

(A)
Sensor data, video and audio data, etc...
L_{LI}= Local Inference Latency
Inference(M)
Results to App
(B)
Sensor data, video and audio data, etc...
L_{FO}=E2E Latency
Transfer (sensor, video and audio data)
Inference(M)
Transfer (results)
Results to App
(C)
Sensor data, video and audio data, etc...
L_{PO}=E2E Latency
Inference(M_a)
Transfer (Intermediate data)
Inference(M_b)
Transfer (results)
Results to App

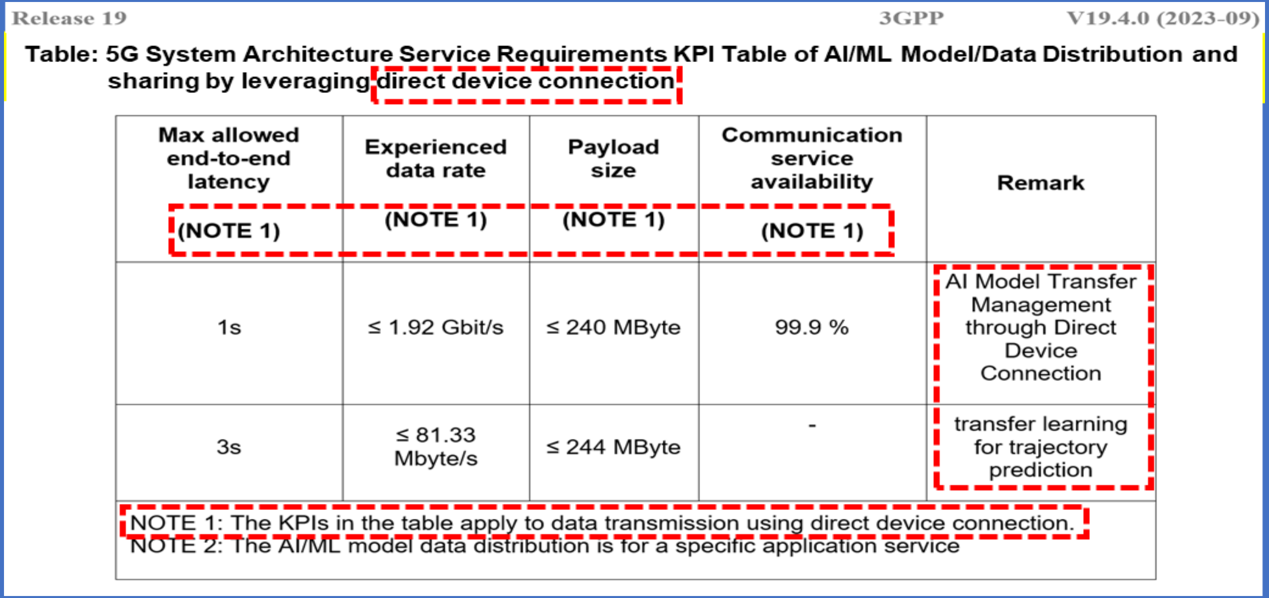

Release 19 3GPP V19.4.0 (2023-09)

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. NOTE 2: The AI/ML model data distribution is for a specific application service				

2. 5G System use of AI/ML

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.

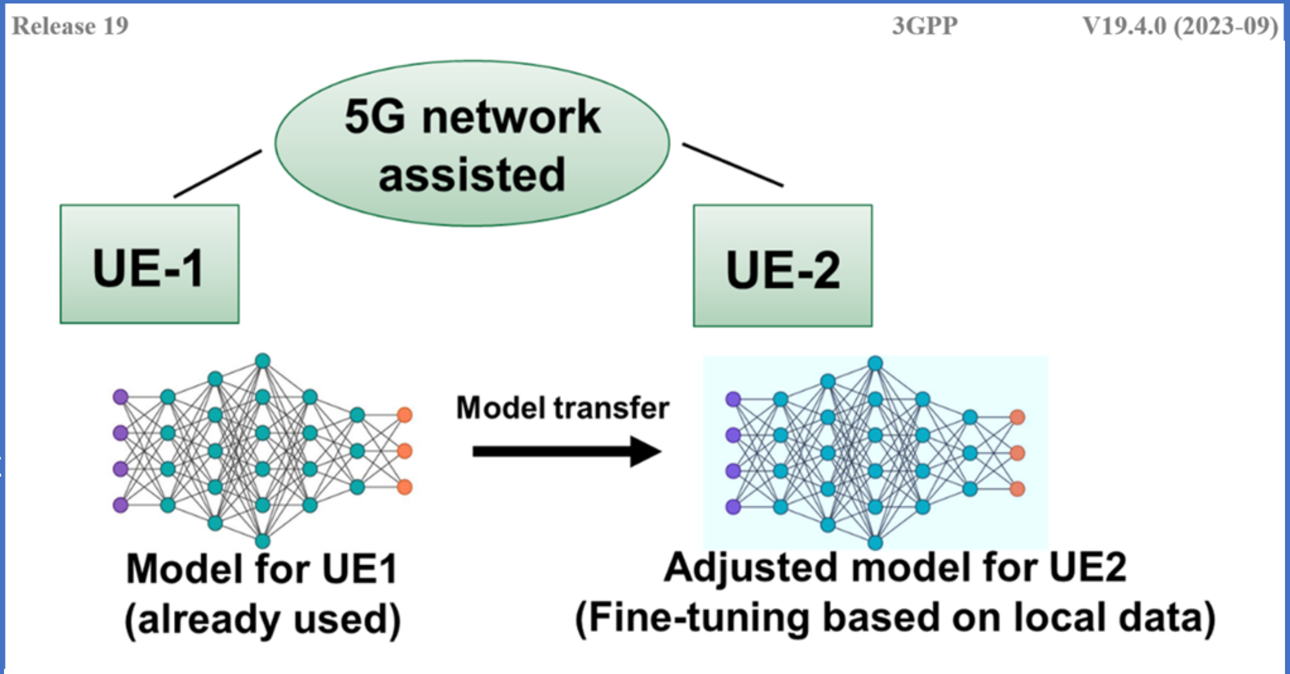

Release 19
3GPP
V19.4.0 (2023-09)
5G network assisted
UE-1
UE-2
Model transfer
Model for UE1 (already used)
Adjusted model for UE2 (Fine-tuning based on local data)

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

2. 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)

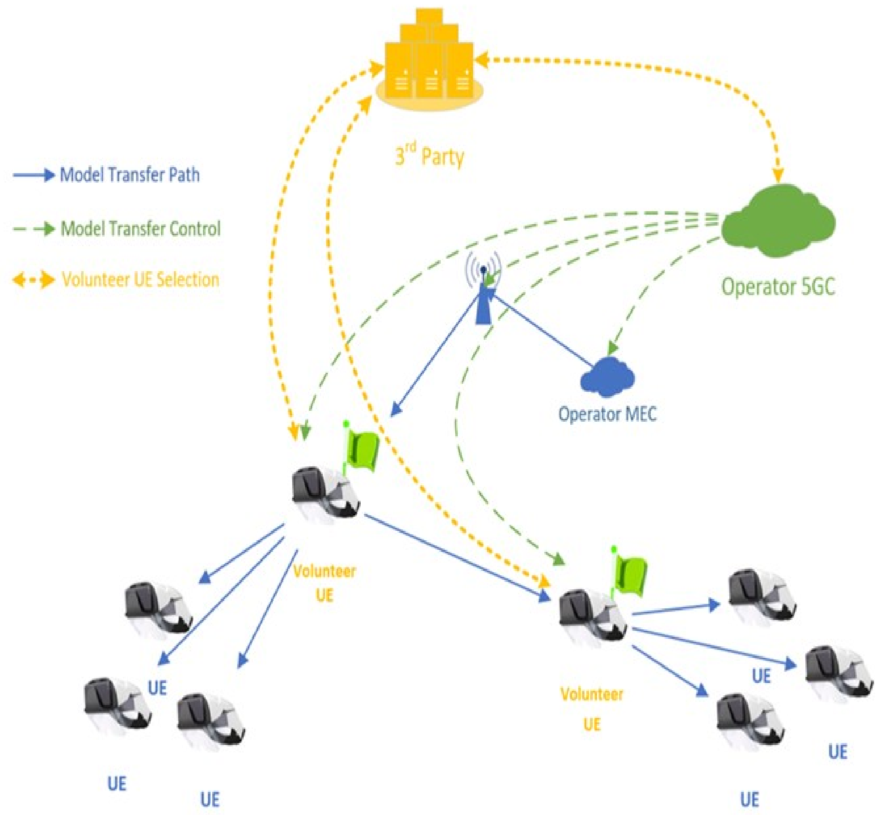


Figure: 5G System Architecture AI/ML Model Management through Direct Device Connection

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.

2. 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.

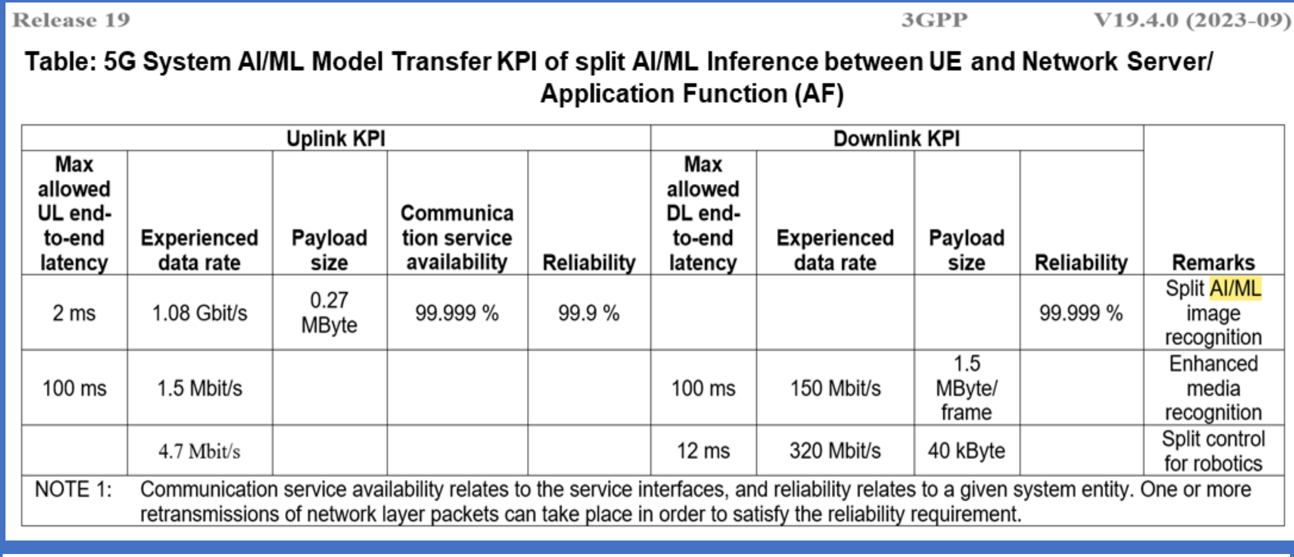


Table: 5G System AI/ML Model Transfer KPI of split AI/ML Inference between UE and Network Server/ Application Function (AF)

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:

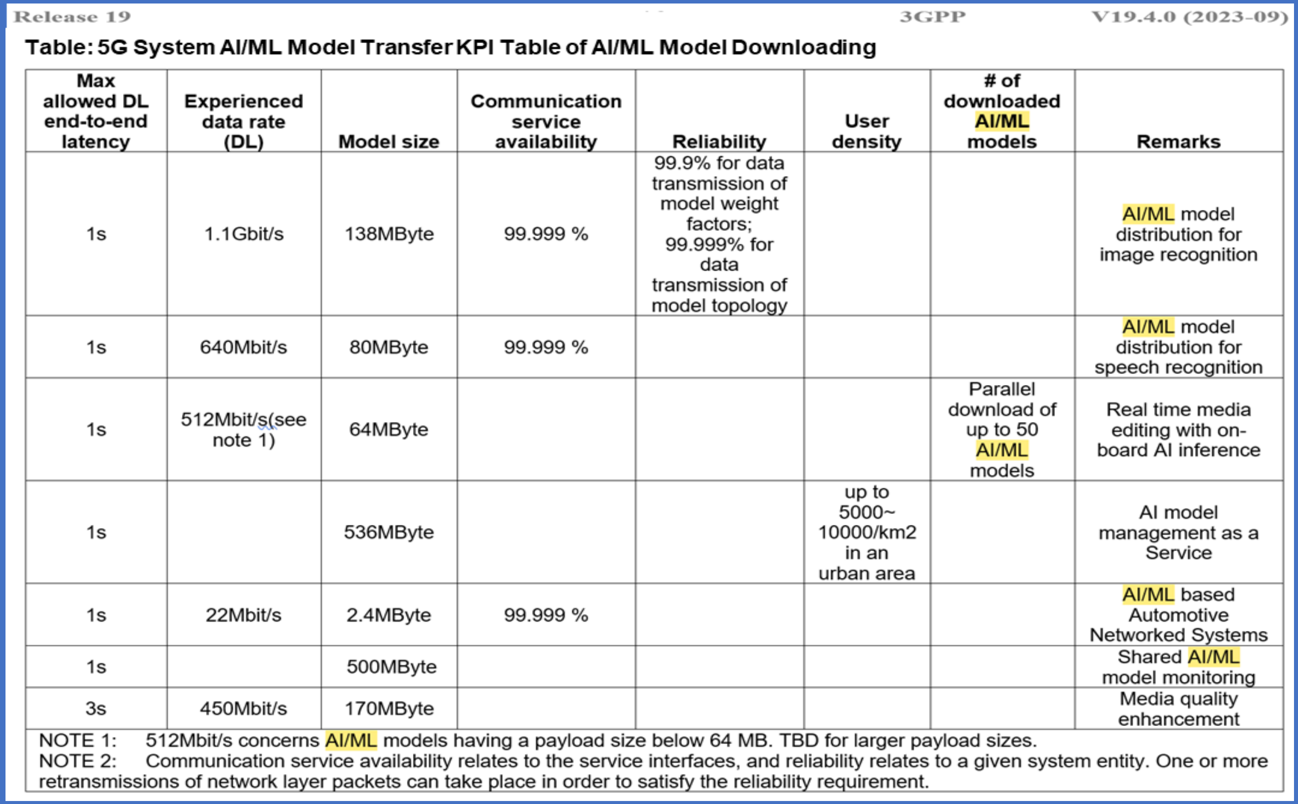


Table: 5G System AI/ML Model Transfer KPI Table of AI/ML Model Downloading

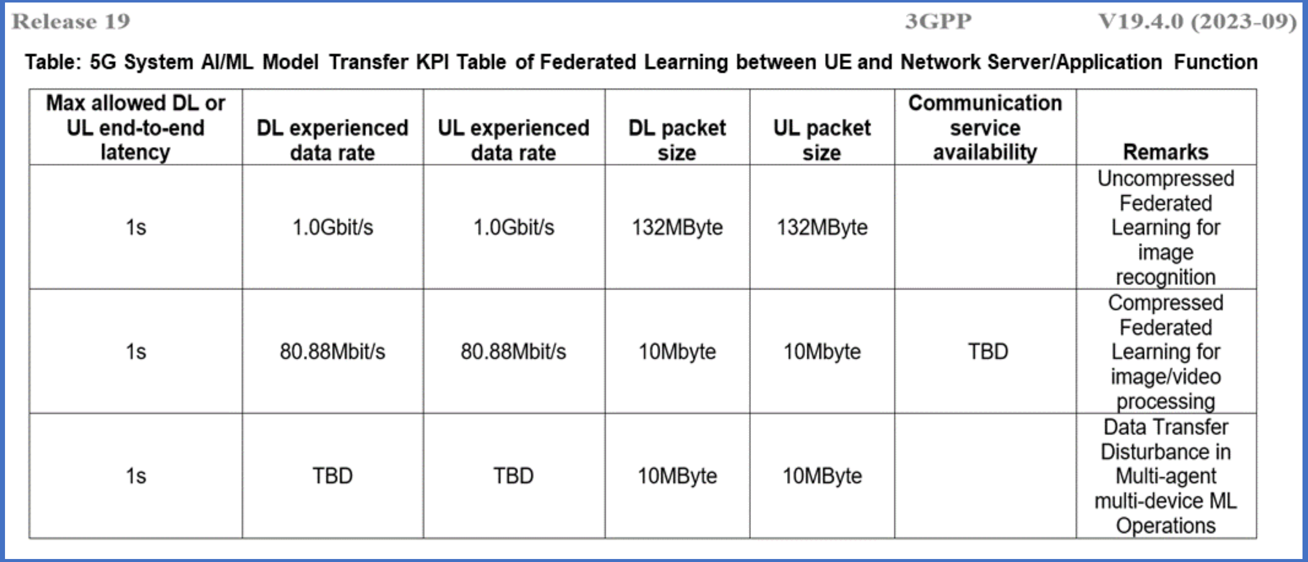
Max allowed DL end-to-end latency	Experienced data rate (DL)	Model size	Communication service availability	Reliability	User density	# of downloaded AI/ML models	Remarks
1s	1.1Gbit/s	138MByte	99.999 %	99.9% for data transmission of model weight factors; 99.999% for data transmission of model topology			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.

2. 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:

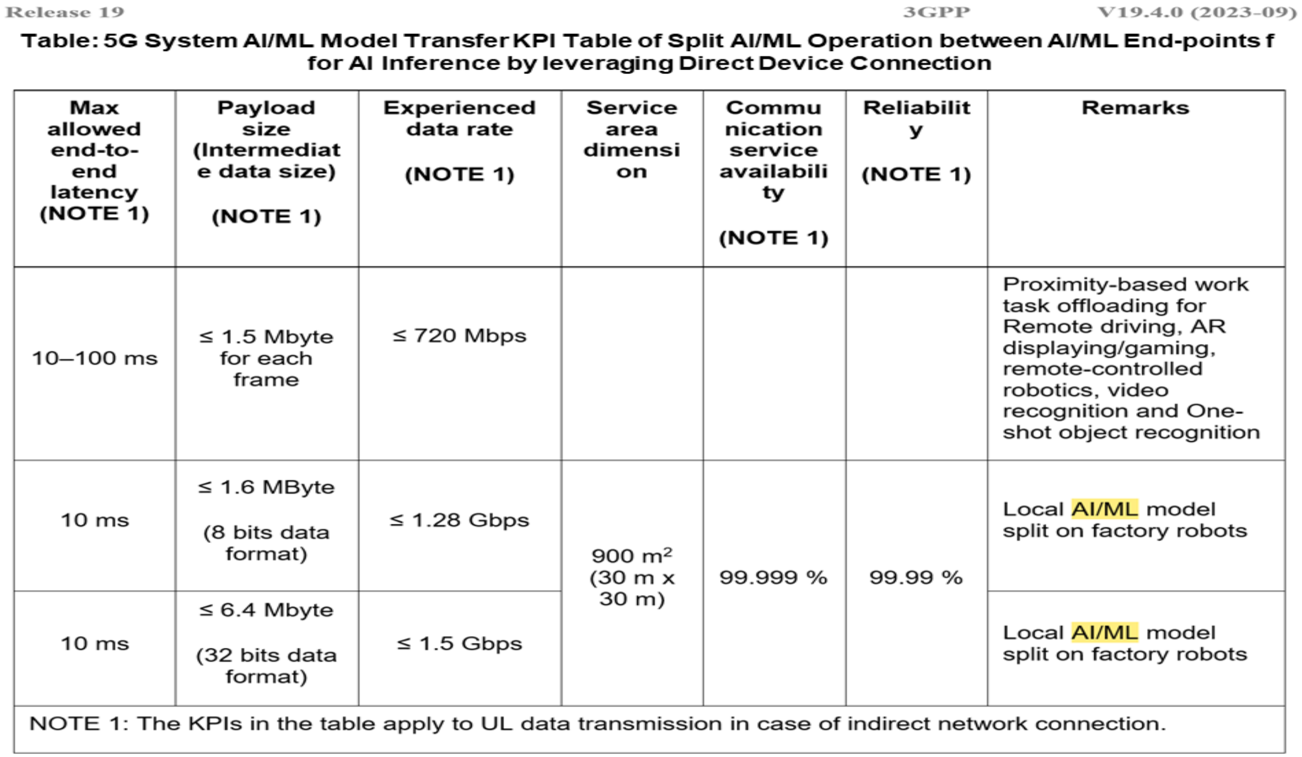


Release 19 3GPP V19.4.0 (2023-09)

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	TBD	Uncompressed Federated Learning for image recognition
1s	80.88Mbit/s	80.88Mbit/s	10Mbyte	10Mbyte		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

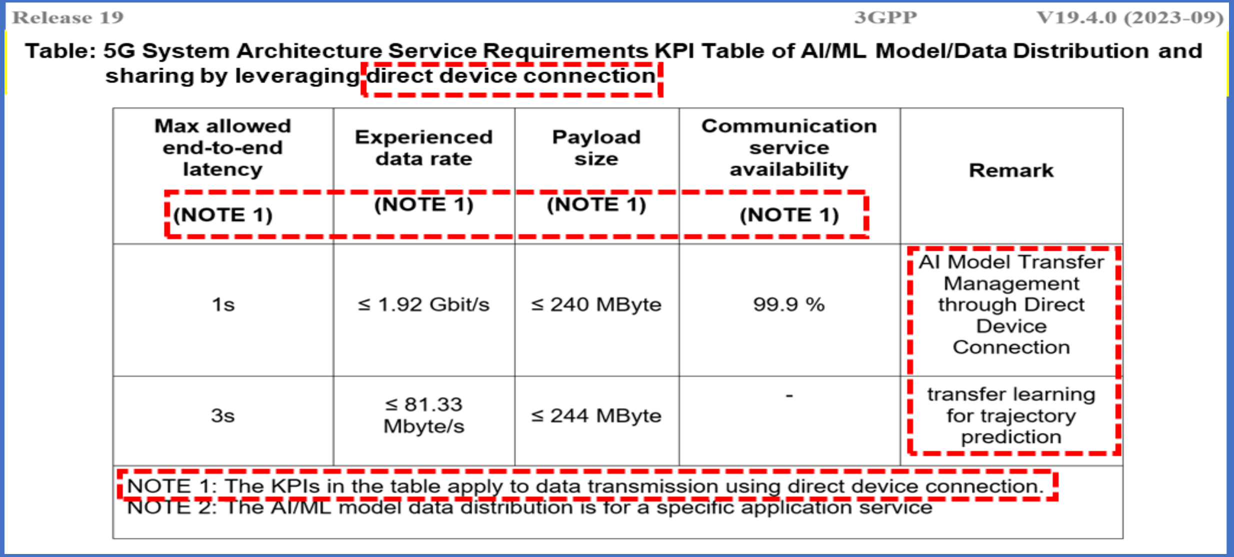
Max allowed end-to-end latency (NOTE 1)	Payload size (Intermediate data size) (NOTE 1)	Experienced data rate (NOTE 1)	Service area dimension	Communication service availability (NOTE 1)	Reliability (NOTE 1)	Remarks
10–100 ms	≤ 1.5 Mbyte for each frame	≤ 720 Mbps				Proximity-based work task offloading 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.

2. 5G System use of AI/ML

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:



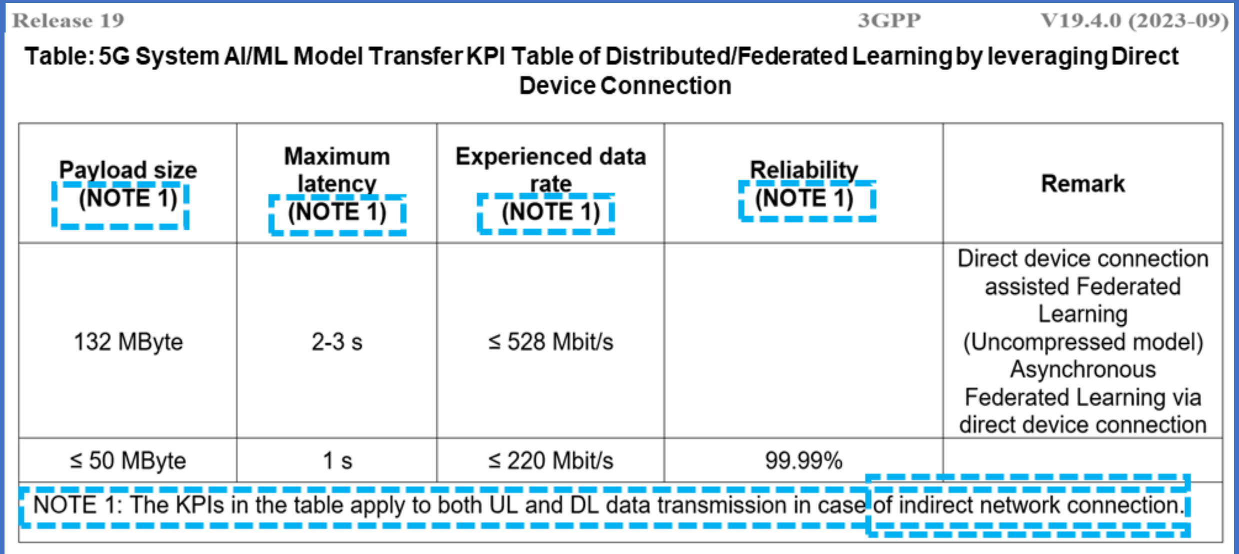
Release 19 3GPP V19.4.0 (2023-09)

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.
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.

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

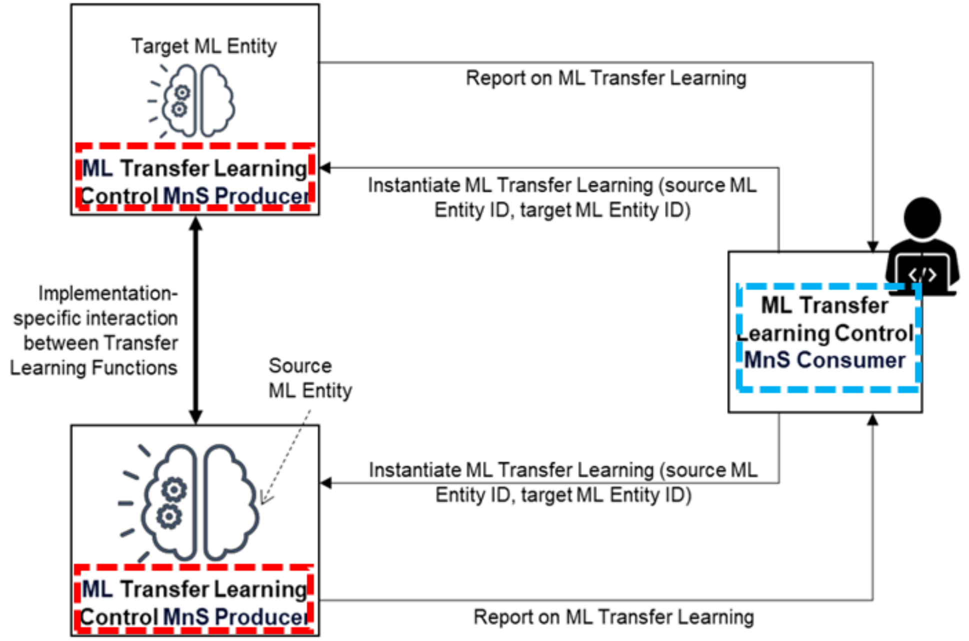

Target ML Entity
ML Transfer Learning Control MnS Producer
Report on ML Transfer Learning
Instantiate ML Transfer Learning (source ML Entity ID, target ML Entity ID)
Implementation-specific interaction between Transfer Learning Functions
Source ML Entity
ML Transfer Learning Control MnS Consumer
Instantiate ML Transfer Learning (source ML Entity ID, target ML Entity ID)
ML Transfer Learning Control MnS Producer
Report on ML Transfer Learning

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

2. 5G System use of AI/ML

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.*

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

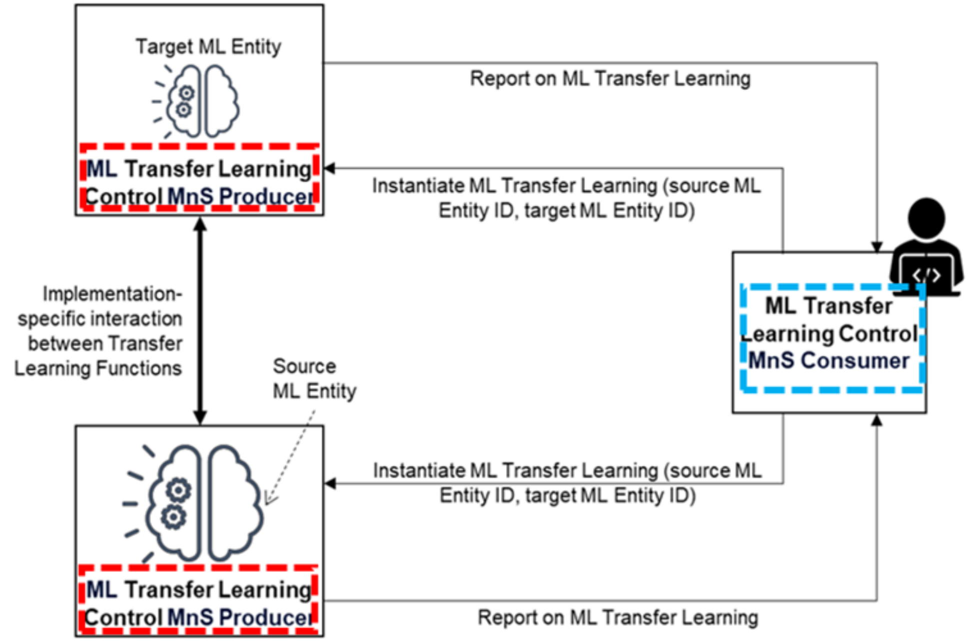

Target ML Entity
ML Transfer Learning Control MnS Producer
Report on ML Transfer Learning
Instantiate ML Transfer Learning (source ML Entity ID, target ML Entity ID)
Implementation-specific interaction between Transfer Learning Functions
Source ML Entity
ML Transfer Learning Control MnS Consumer
Instantiate ML Transfer Learning (source ML Entity ID, target ML Entity ID)
ML Transfer Learning Control MnS Producer
Report on ML Transfer Learning

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)

2. 5G System use of AI/ML

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)

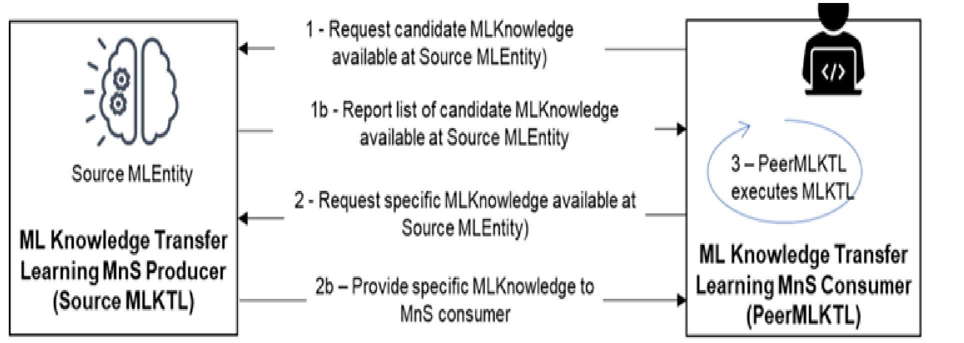

Source MLEntity
ML Knowledge Transfer Learning MnS Producer (Source MLKTL)
1 - Request candidate MLKnowledge available at Source MLEntity)
1b - Report list of candidate MLKnowledge available at Source MLEntity
2 - Request specific MLKnowledge available at Source MLEntity)
2b – Provide specific MLKnowledge to MnS consumer
3 – PeerMLKTL executes MLKTL
ML Knowledge Transfer Learning MnS Consumer (PeerMLKTL)

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)

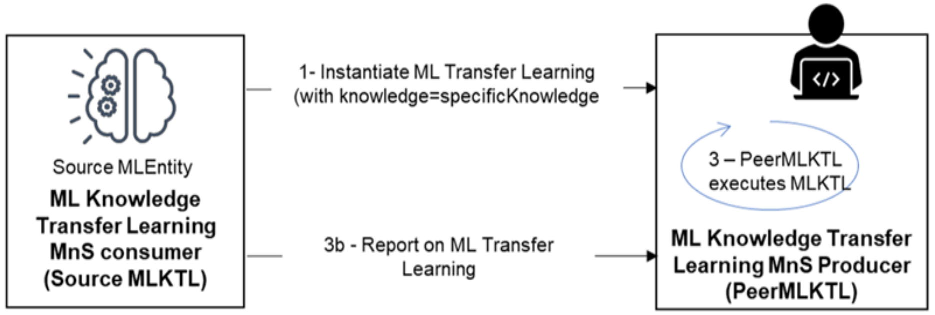

Source MLEntity
ML Knowledge Transfer Learning MnS consumer (Source MLKTL)
1- Instantiate ML Transfer Learning (with knowledge=specificKnowledge
3b - Report on ML Transfer Learning
3 – PeerMLKTL executes MLKTL
ML Knowledge Transfer Learning MnS Producer (PeerMLKTL)

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

2. 5G System use of AI/ML

Knowledge Sharing and Transfer Learning

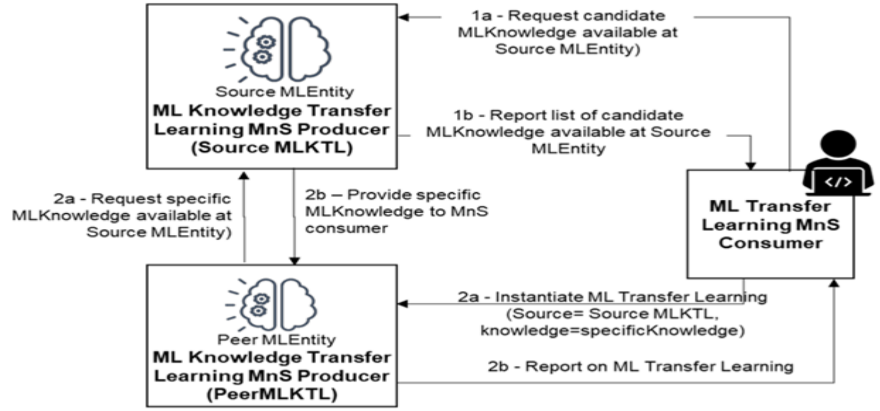
Specifically, the 3GPP Management System should provide means for an MnS Consumer to request and receive Sharable Knowledge as well as means for the Provider of MLKLT to share the Knowledge with the MnS Consumer or any stated Target ML Training Function.

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*.

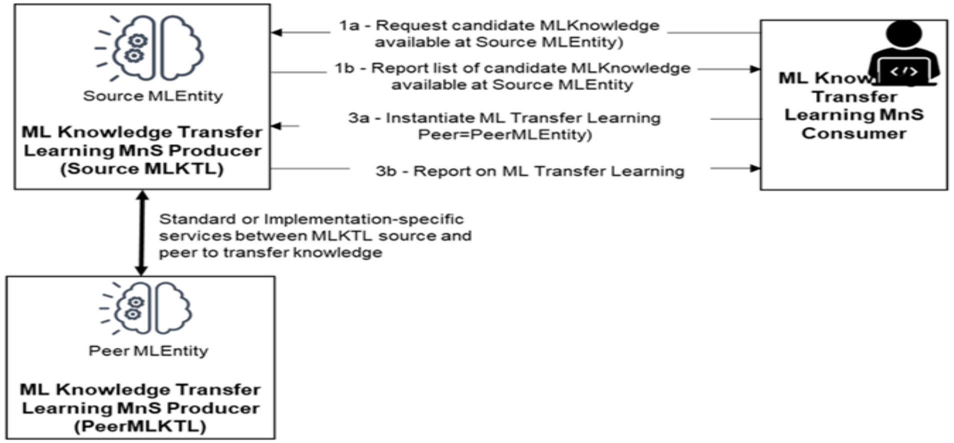
Release 18 3GPP V18.0.0 (2023-09)



Source MLEntity
ML Knowledge Transfer Learning MnS Producer (Source MLKTL)
1a - Request candidate MLKnowledge available at Source MLEntity)
1b - Report list of candidate MLKnowledge available at Source MLEntity
2a - Request specific MLKnowledge available at Source MLEntity)
2b – Provide specific MLKnowledge to MnS consumer
ML Transfer Learning MnS Consumer
</>
Peer MLEntity
ML Knowledge Transfer Learning MnS Producer (PeerMLKTL)
2a - Instantiate ML Transfer Learning (Source= Source MLKTL, knowledge=specificKnowledge)
2b - Report on ML Transfer Learning

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

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



Source MLEntity
ML Knowledge Transfer Learning MnS Producer (Source MLKTL)
1a - Request candidate MLKnowledge available at Source MLEntity)
1b - Report list of candidate MLKnowledge available at Source MLEntity
3a - Instantiate ML Transfer Learning Peer=PeerMLEntity)
3b - Report on ML Transfer Learning
ML Knowledge Transfer Learning MnS Consumer
</>
Standard or Implementation-specific services between MLKTL source and peer to transfer knowledge
Peer MLEntity
ML Knowledge Transfer Learning MnS Producer (PeerMLKTL)

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

5G System Management Data Analytics (MDA) AI/ML support

The MDA Process may utilize AI/ML Technologies.

5G MDA Function may optionally be deployed as one (1) or more AI/ML Inference Function(s) in, which the relevant ML entities are used for inference per the corresponding MDA Capability.

Specifications for MDA ML entity training to enable ML entity deployments are given in 3GPP 5G Management and Orchestration AI/ML Management.

Intelligence in Analytics, played by MDA, in the Management loop which can be Open Loop (Operator Controlled) or Closed Loop Autonomous) as shown in the Figure, generates value by processing and providing analysis of Management and Network Data, where AI and ML Techniques could be utilized.

The Management Services (MnSs) for a Mobile Network including Network Slicing (SST) may be produced by a Set of Functional Blocks, such as NSMF, NSSMF, NFMF and CSMF, that are "producing" and "consuming! various Management Services (MnSs).

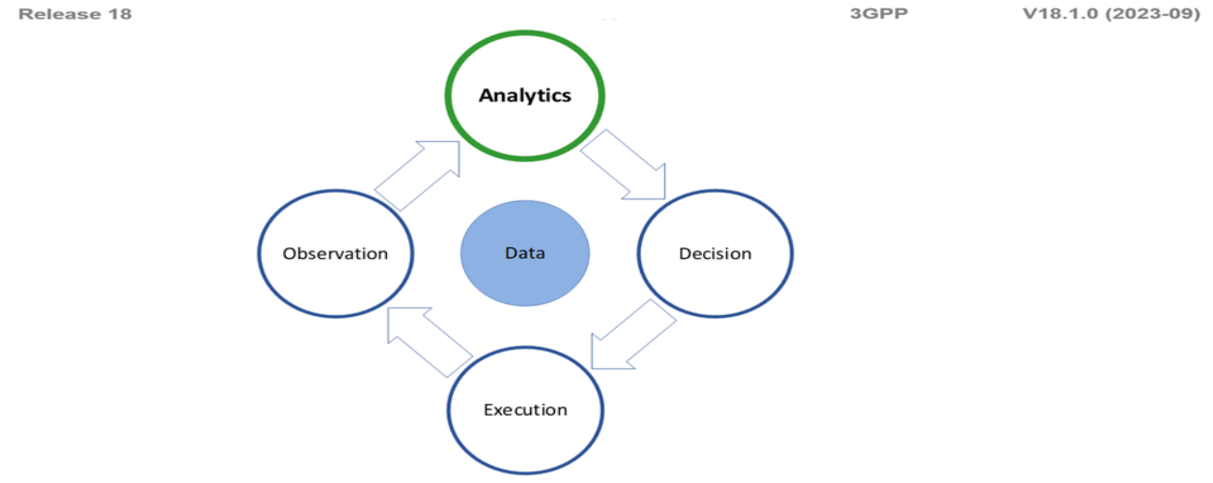


Figure: 5G System Intelligence in Analytics, played by 5G MDA in the Management Loop

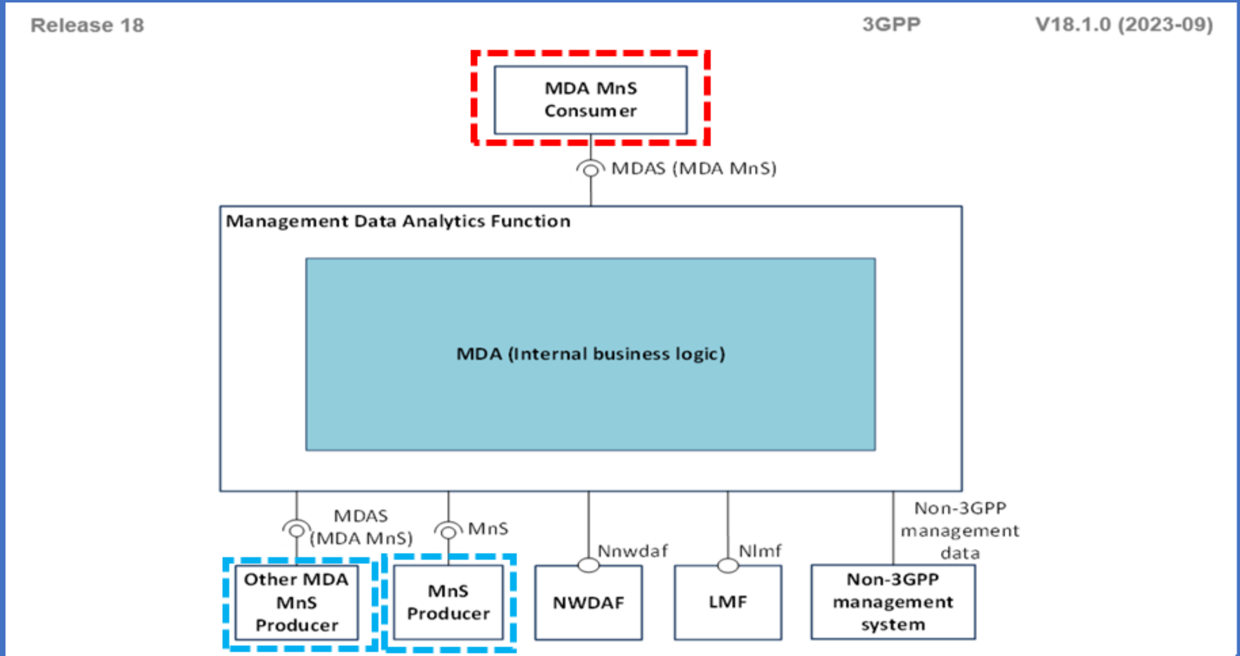


Figure: 5G System Architecture Management Data Analytics (MDA) Functional Overview and Service Framework

2. 5G System use of AI/ML

5G System Management Data Analytics (MDA) AI/ML support

Interaction with 5G CN and RAN Domains

The MDA MnS Producer provides Analytics Data for Management Purposes based on Input Data related to different types of NFs or Entities in the Network, e.g. Data reported from gNB and/or specific Core Network Function(s).

Depending on the Use Case (UC) and when needed, the MDA MnS Producer may use the Analytics results produced by NWDAF as Input.

Management Data Analytics Function (MDAF) may act as 3GPP Domain-specific (e.g. RAN or CN) or as 3GPP Cross-Domain MDA MnS Producer. The Figure illustrates the example of co-ordination between NWDAF, gNB and MDA MnS producer(s) for Data Analytics Purpose.

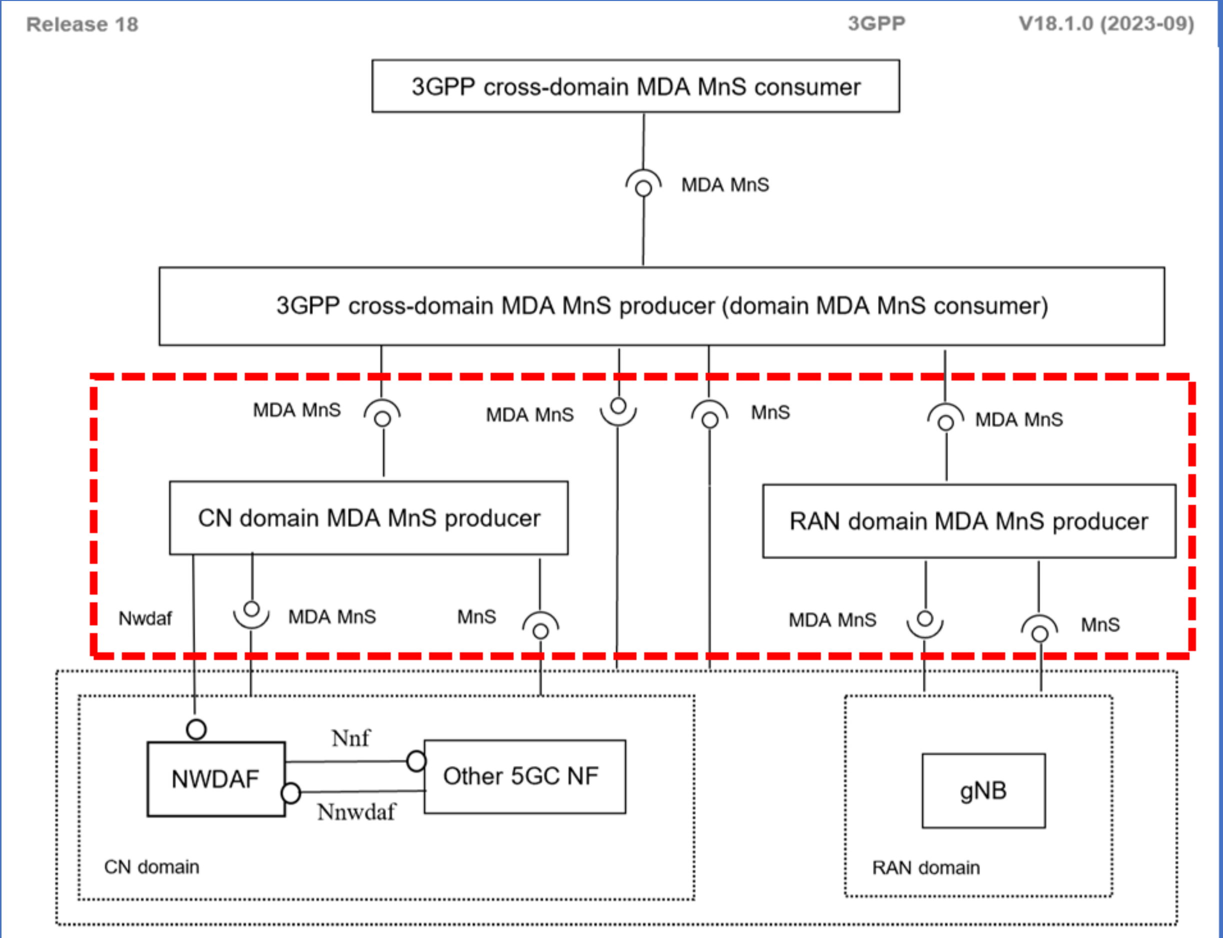

Release 18
3GPP
V18.1.0 (2023-09)
3GPP cross-domain MDA MnS consumer
MDA MnS
3GPP cross-domain MDA MnS producer (domain MDA MnS consumer)
MDA MnS
MDA MnS
MnS
MDA MnS
CN domain MDA MnS producer
RAN domain MDA MnS producer
Nwdaf
MDA MnS
MnS
MDA MnS
MnS
Nnf
NWDAF
Other 5GC NF
Nnwdaf
gNB
CN domain
RAN domain

Figure: 5G System Architecture Management Data Analytics (MDA) example of co-ordination between NWDAF, gNB and MDAS (MDA MnS) Producer

2. 5G System use of AI/ML

5G System Management Data Analytics (MDA) AI/ML support

Deployment of Multiple MDAs

The 3GPP Cross Domain Management may consume MDA MnS provided by RAN Management as shown in the Figure.

The Management Function (MDAF) playing the Role of *3GPP Cross Domain MDA MnS Producer* interacts with *RAN Domain MDA* per each MDA Use Case/Capability as follows:

- The Cross Domain MDA MnS Producer may consume the RAN Domain MDA MnS.

- The Cross Domain MDA MnS Producer may consume MnS provided by RAN Domains, and produce MDA MnS that may be consumed by 3GPP Cross-Domain MDA MnS Consumer(s).

The Management Function (MDAF) playing the Role of RAN Domain MDA MnS Producer interacts with MnS Producers per each Use Case/Capability as follows:

- The RAN Domain MDA MnS Producer may consume MnS provided by RAN Domain Management, other MDA MnS Producers, Management Data derived by Sub-Network Management Function(s), and Management Data derived by element Management Function(s).

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

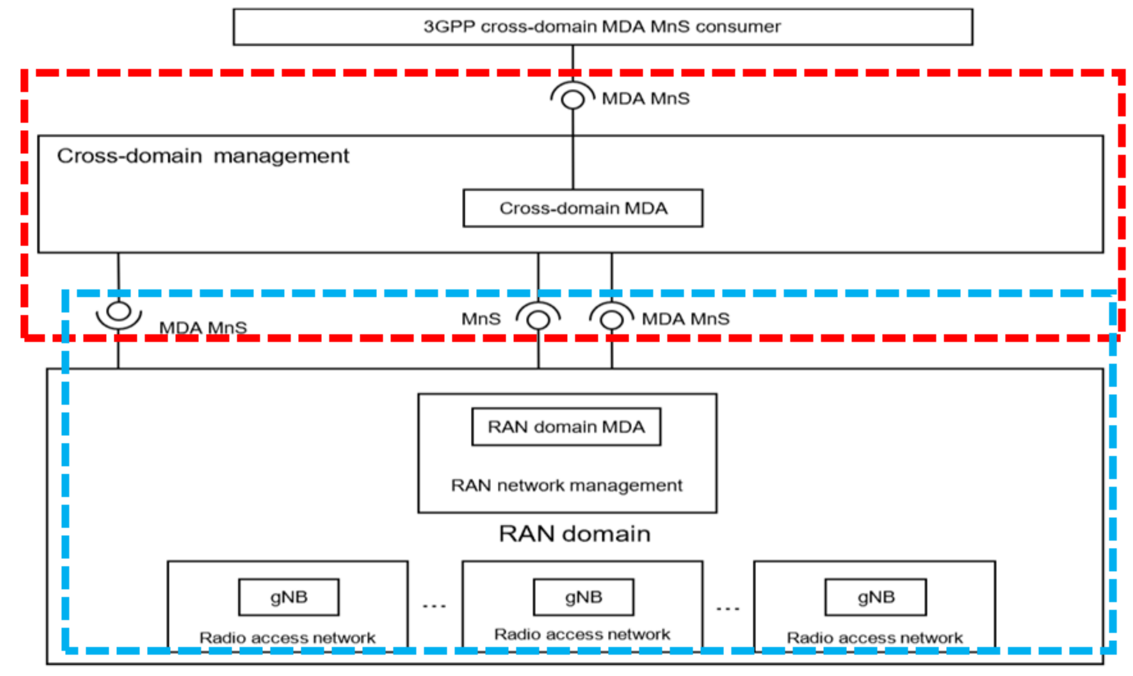


Figure: 5G System Architecture Management Data Analytics (MDA) example of coordination Cross-domain MDA and RAN Domain MDA

5G System Management Data Analytics (MDA) AI/ML support

Solution for Measurement Data Correlation Analytics

The Solution may enable the ML MnS Producer to support (with or without scheduled) Measurement Data Correlation Analytics via Configuration, request from Consumer.

The High-level Solution Description is illustrated in the Figure.

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

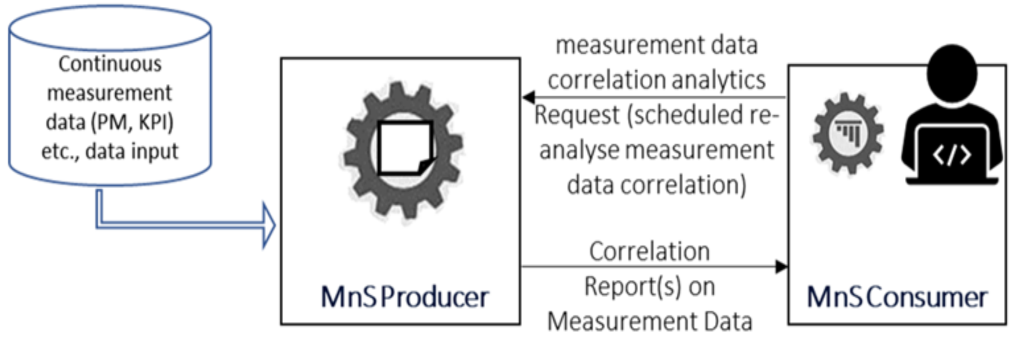

Continuous measurement data (PM, KPI) etc., data input
MnS Producer
measurement data correlation analytics Request (scheduled re-analyse measurement data correlation)
Correlation Report(s) on Measurement Data
MnS Consumer
</>

Figure: 5GS MDA High-level Solution description for Measurement Data Correlation Analytics

The steps in the figure are explained below:

1) *MnS Consumer* may request the *(MnS) Producer* to initiate the Measurement Data Correlation Analytics. The request includes an indication may be an *attribute (e.g. an information element*, it may be named as *MDCA-Measurement Data Correlation Analytics*, which may include necessary Configuration for Measurement Data Correlation Analytics).
2) MnS Producer upon receiving the request with configuration for measurement data correlation analytics, instantiates the MOI.
3) The instantiated MOI takes care of the input measurement data correlation analytics as part of request handling. E.g. prepare the pipeline of Data input, cleansing, measurement data correlation analytics, and report the results with correlation results, etc.
4) If the MnS may support regular re-new the Measurement Data Analytics, the *MnS Consumer* may:

- Suspend the MDCA by update the configuration of scheduling information in *MnS Producer*.
- Cancel the scheduled MDCA by stopping ongoing correlation analytics activity and delete related configuration in *MnS Producer*.
- Update the scheduled MDCA cycle by re-configure the ongoing scheduled correlation analytics activity in *MnS Producer*.

The configuration information Element (e.g. MeasurementDataCorrelationAnalytics) may contain the following attributes:

- An attribute may indicate the address(es) of the candidate correlated measurement data generated from MDCA activity.
- An attribute may indicate the MDCA results, it may be SUCCESSFUL WITH MDCA GENERATED, FAILED DUE TO PERFORMANCE IMPACT, or other failure results.
- An attribute may indicate the MDCA performance requirement. It can be a percentage which requires the performance impact of trained ML Entity with generated measurement data within the range of the Performance trained with full measurement data. E.g. 5 % means the Model performance for the ML Entity trained with generated measurement data shall be no worse than 5 % of the performance trained with Full Measurement Data.
- An attribute may indicate the actual MDCA performance impact. It can be a percentage which indicate the loss the model performance from trained ML Entity with generated measurement data comparison to the performance trained with full measurement data.
- If MnS support producer initiated regular MDCA, a scheduled MDCA activity may be enabled, a scheduling attribute may include the following attributes:
- An attribute may indicate how frequent the MDCA activity shall be performed, e.g. weekly, or monthly.
- Optionally MnS consumer may indicate when to start the scheduled MDCA.
- An attribute may indicate the status of current scheduled MDCA as: RUNNING, SUSPENDED.
- The scheduling Flag may indicate if a scheduled MDCA is enabled or not.

2. 5G System use of AI/ML

ML context in 5GS

ML Context attribute represents the *Status and Conditions related to the ML entity*. This may include the *Network Context* as defined in 3GPP as well as other conditions that may be *applicable to the ML Entity, but are not part of Network Characteristics e.g. the Time of Day, Season of the Year*.

As part of ML Model Performance Management, there is the identification of the problem that the ML model is meant to address or deal with. The differences in the Network Context, i.e. Network Status, under which Data is collected to produce Analytics, significantly affect the produced Analytics.

Similarly, the *changes in the ML Context, e.g. the Characteristics of the Data related to the Network Status and Conditions* used for ML Model Training, Testing and Deployment may affect the ML Entity Performance, thus may represent a problem for the ML Entity.

Thus, Management Capabilities are needed to enable awareness of the ML Context in terms of the Identification as well as Monitoring and Reporting of Changes in ML Context as part of the Identification of the Problem that the ML Entity is meant to address or deal with.

ML Context Monitoring and Reporting

ML Context related to ML Model Training, Testing and Deployment needs to be identified by characterizing the Input Data, used by the ML Model, is targeted to work. E.g., such Characterization may be done based on the Statistical Properties of Data.

Monitoring of such ML Context serves to detect the Changes and Anomalies in the ML Context. Some Anomalies may be considered as a Problem that ML Entity is facing as it may lead to its Performance Degradation. Therefore, the (MnS) Consumer of the related AI/ML Service needs to be informed about such observed ML Context Change.

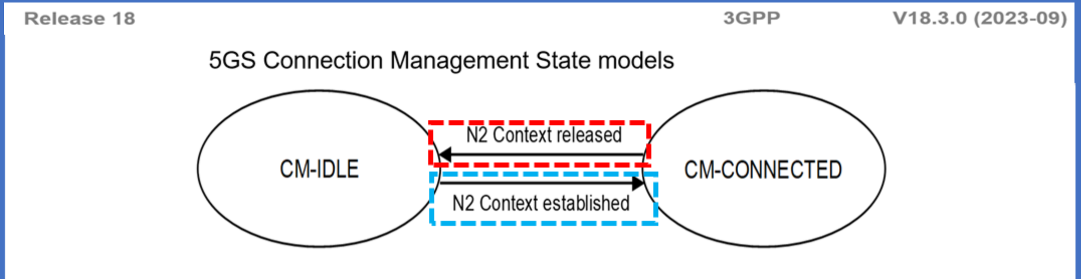


Figure: 5GS Architecture CM State Transition in AMF

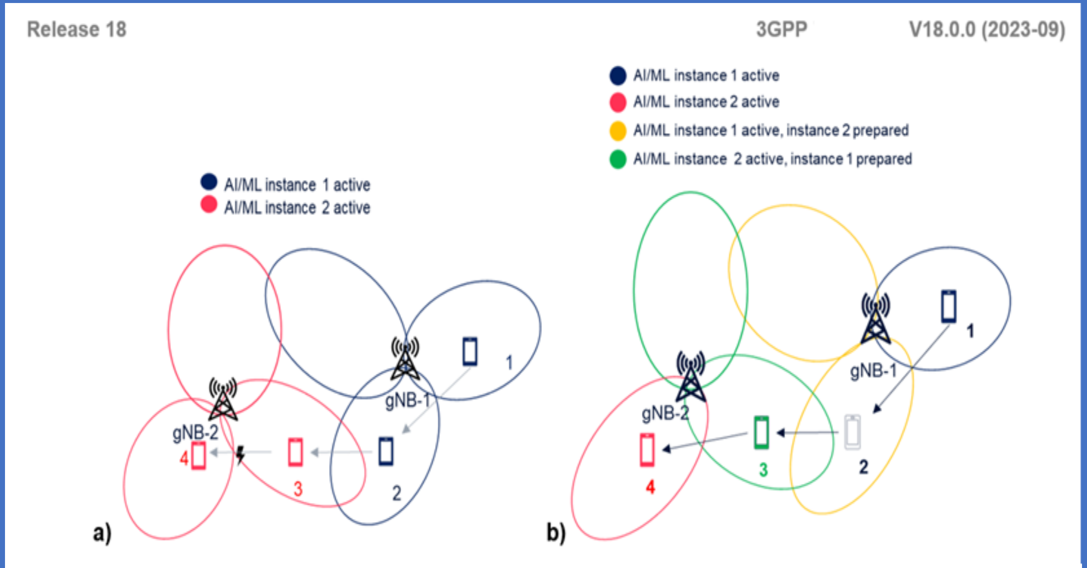


Figure: 5GS MDA Mobility ML Context
a) Validity Scopes; b) Validity and Standby Scope

2. 5G System use of AI/ML

Mobility of ML Context

In several Network Automation Use Cases (UCs), the respective AI/ML Inference Function cannot cover the Complete Network by employing Single ML Entity Instance.

An ML Entity may be trained for a Specific Local Context, and similarly, a different Context may be applicable for Inference, so the ML Entity may be characterized by different Training Context and an expected Inference Context.

However, the Network scopes where the Data used for Training and Inference is collected, does not always necessarily overlap with the Network scopes in which the Function makes decisions.

The Context of ML Entities or AI/ML Inference Function may need to distinguish between Context for Generating Decisions or Insights, the Context from which it generates Measurements or Data as well as the Context in which it is prepared before being Active for Inference.

So, the Characteristics of the respective AI/ML Inference Function need to be distinguished depending on the different Contexts of the AI/ML Inference Function.

As such besides the Validity scope defined by the Training Context and an expected Inference Context, the ML Entity should also be characterized by Specific Measurement Scopes, where the Input Measurements are collected.

And these may also be separately defined for the 2 Use Cases (UCs).

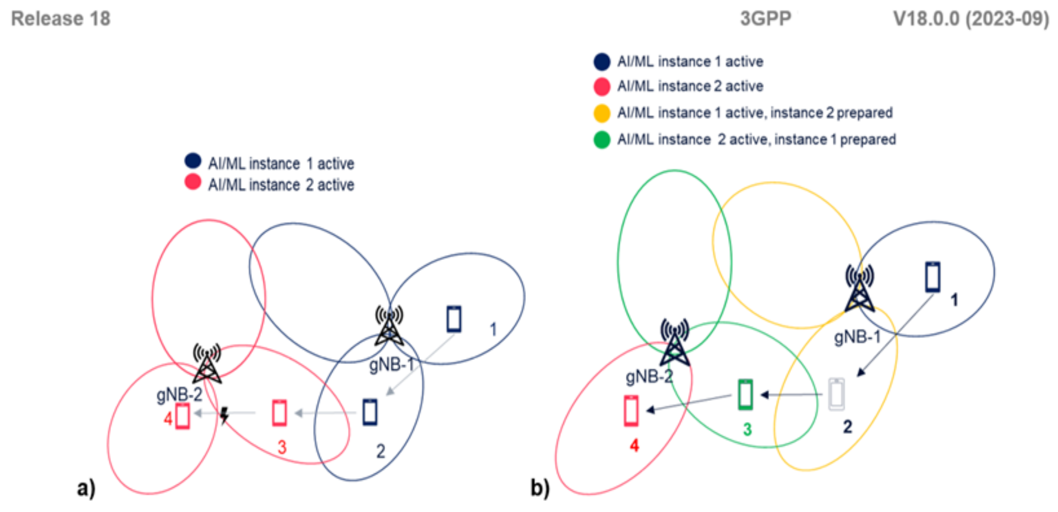


Figure: 5GS MDA Mobility ML Context
a) Validity Scopes; b) Validity and Standby Scope

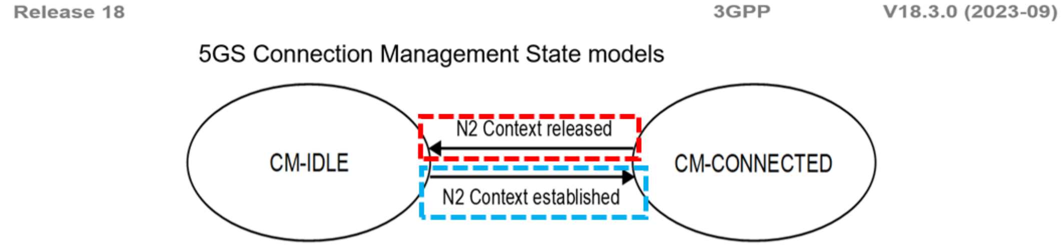


Figure: 5GS Architecture CM State Transition in AMF

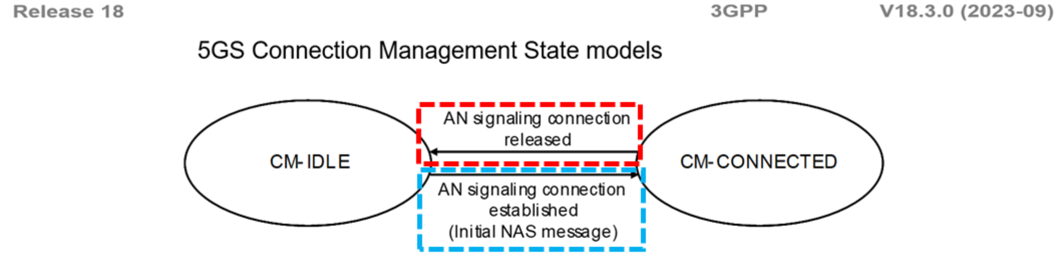

Release 18
3GPP
V18.3.0 (2023-09)
5GS Connection Management State models
CM-IDLE
AN signaling connection released
AN signaling connection established (Initial NAS message)
CM-CONNECTED

Figure: 5GS Architecture CM State Transition in UE

2. 5G System use of AI/ML

Network Context

An MDA MnS Producer provides Analytics with respect to a particular Network Context, i.e. Network Status, under which Data is collected to produce Analytics. E. g, a Prediction of Load in an Area of Interest (AoI) may differ when all gNBs and potential additional RATs are operating compared to Case where certain gNBs or other RATs are experiencing a fault or are powered off to save energy.

The Analytics conducted and produced by the *MDA MnS Producer* for these two example Scenarios would be different and directly affected by the specific Status of Network.

Although the Network Status (Context) affects the produced Analytics conducted by the MDA Producer, awareness of the Network Context would fall on the Consumer side to complement the obtained Analytics results. This Network Context, reflecting Network Status at the Time of Enabling Data Collection, is important for the MDA MnS Consumer to understand the Network Conditions related to the obtained Analytics and hence be able to use such Analytics more efficiently.

The MDA MnS Consumer cannot expect the MDA Producer to provide the Network Context, because the Network Context Interest of each MDA MnS Consumer may differ depending on the Usage and Purpose of Analytics.

The usage can include a proprietary Algorithm that assist a Decision-Making Process. E. g., a Load balancing Algorithm may require the Load and Mobility Information among neighboring gNB whereas other Load balancing Algorithms may also require Load and Mobility Information from a greater Geographical Area.

The selection of the Parameters and their combinations may prove to be impractical for the MDA MnS producer to prepare and provide information. Hence, it is efficient for the MDA MnS Producer to prepare only the MDA Output without including any network context and allow the MDA MnS consumer to obtain the required network context, to Complement to the obtained Analytics, using Conventional Configuration Management Procedures as described.

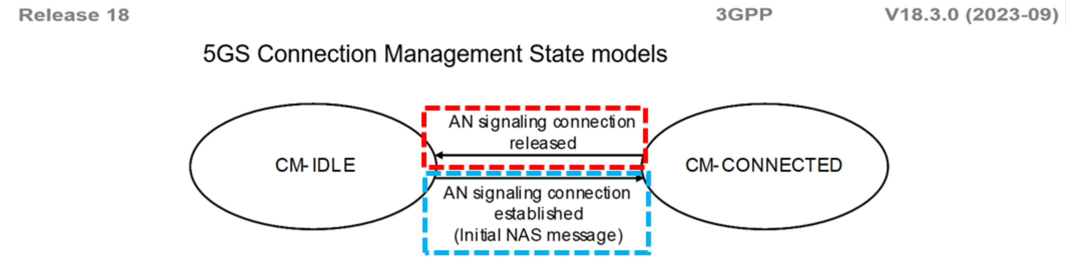

Release 18
3GPP
V18.3.0 (2023-09)
5GS Connection Management State models
CM-IDLE
CM-CONNECTED
AN signaling connection released
AN signaling connection established (Initial NAS message)

Figure: 5GS Architecture CM State Transition in UE

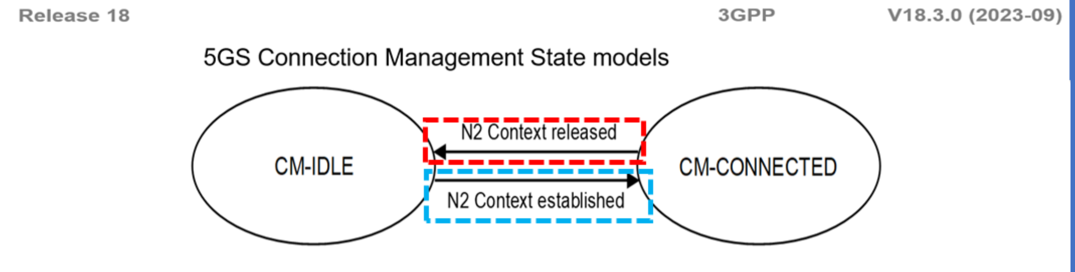


Figure: 5GS Architecture CM State Transition in AMF

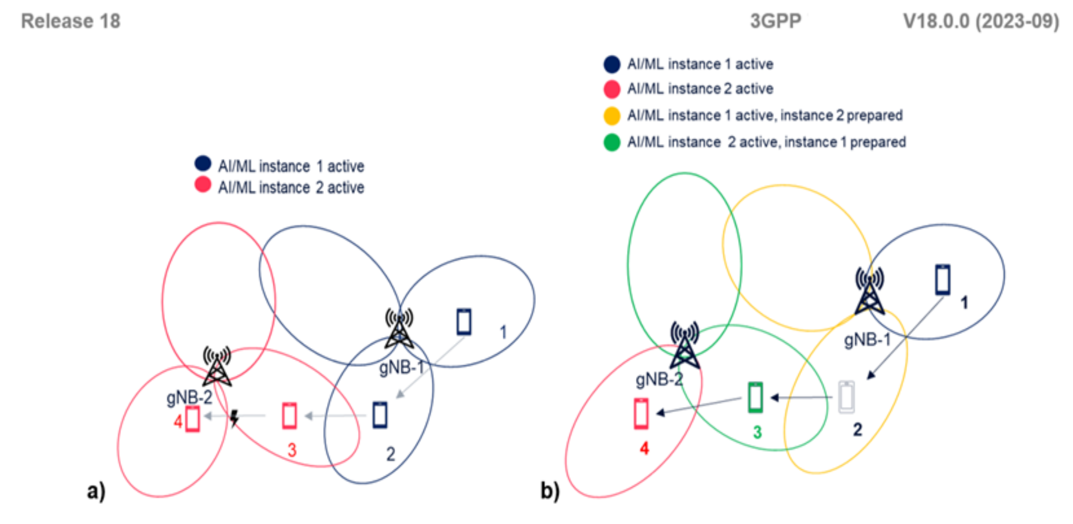


Figure: 5GS MDA Mobility ML Context
a) Validity Scopes; b) Validity and Standby Scope

2. 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.*

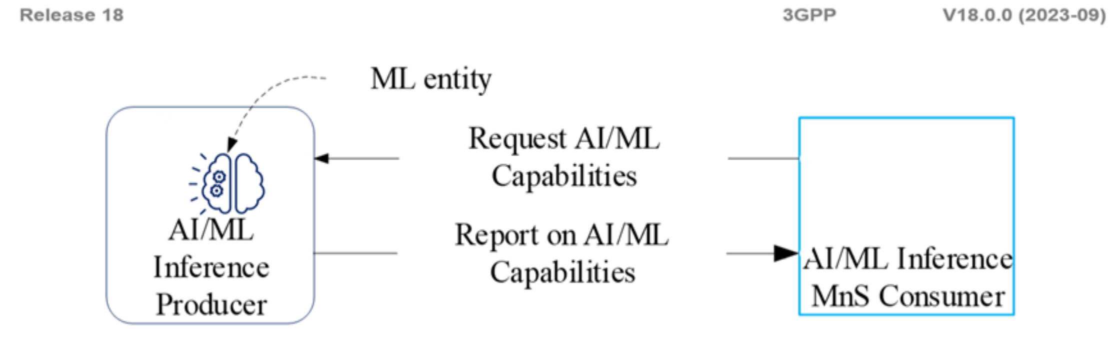

Release 18
3GPP
V18.0.0 (2023-09)
ML entity
AI/ML Inference Producer
Request AI/ML Capabilities
Report on AI/ML Capabilities
AI/ML Inference MnS Consumer

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

2. 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.

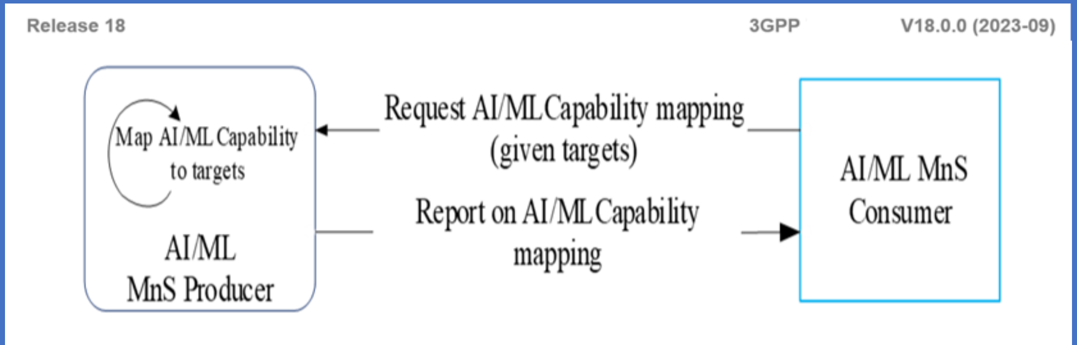

Release 18
3GPP
V18.0.0 (2023-09)
Map AI/ML Capability to targets
AI/ML MnS Producer
Request AI/ML Capability mapping (given targets)
Report on AI/ML Capability mapping
AI/ML MnS Consumer

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.*

2. 5G System use of AI/ML

5GS Service Based Management Architecture (SBMA) - Management Services (MnS)

The Fundamental Building Block of the Service Based Management Architecture (SBMA) is the Management Service (MnS).

A MnS is a set of offered Capabilities for Management and Orchestration of Network and Services.

The Entity producing an MnS is called *MnS Producer.*

The Entity consuming an MnS is called *MnS Consumer*.

An MnS provided by an MnS Producer can be consumed by any entity with appropriate Authorisation and Authentication.

An MnS Producer offers its services via a Standardized Service Interface composed of Individually specified MnS Components.

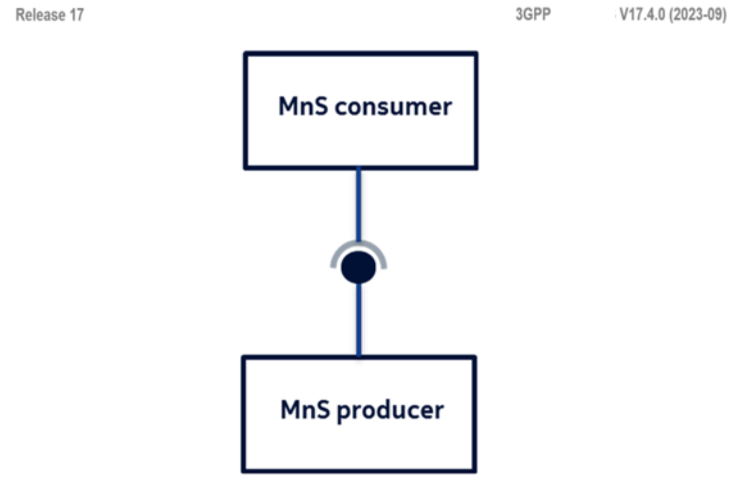
A MnS is specified using different independent components. A concrete MnS is composed of at least two (2) of these Components.

Three (3) different Component Types are defined, called

- MnS Component Type A,

- MnS component type B and

- MnS component type C.



Release 17 3GPP V17.4.0 (2023-09)

Figure: 5GS Service Based Management Architecture (SBMA) - Management Services (MnSs)

5GS Service Based Management Architecture (SBMA) - Management Service (MnS) and Management Function (MnF) concept

A *Management Function (MnF)* is a Logical Entity playing the *Roles of MnS Consumer and/or MnS Producer.*

A *Management Service (MnS)* produced by Management Function (MnF) may have *multiple Consumers*.

The *MnF* may consume multiple Management Services (*MnSs*) from one (1) or multiple Management Service Producers.

An example of a *MnF playing both roles (Management Service Producer and Management Service Consumer) is illustrated in the Figures.*

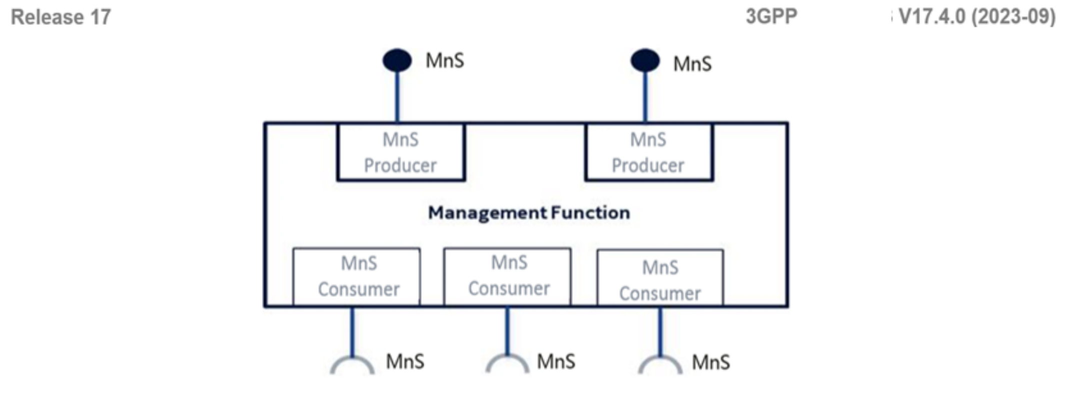


Figure: 5GS Service Based Management Architecture (SBMA) - Management Services (MnSs) Example of Management Function (MnF) and Management Services (MnSs)

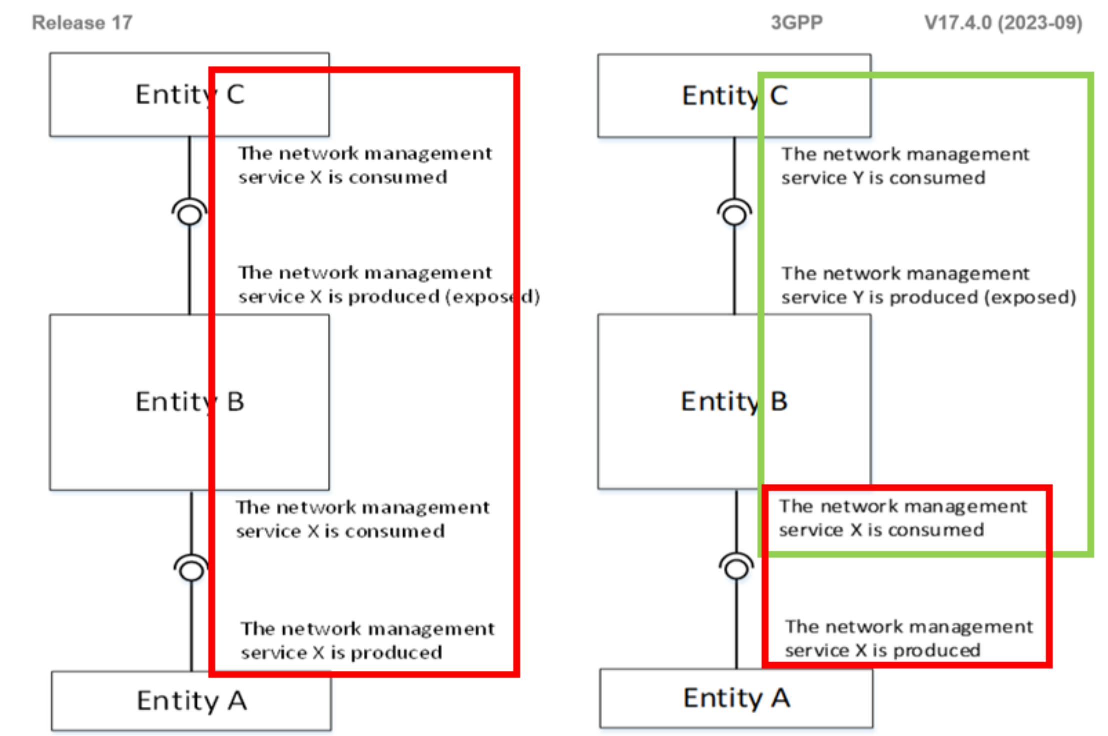
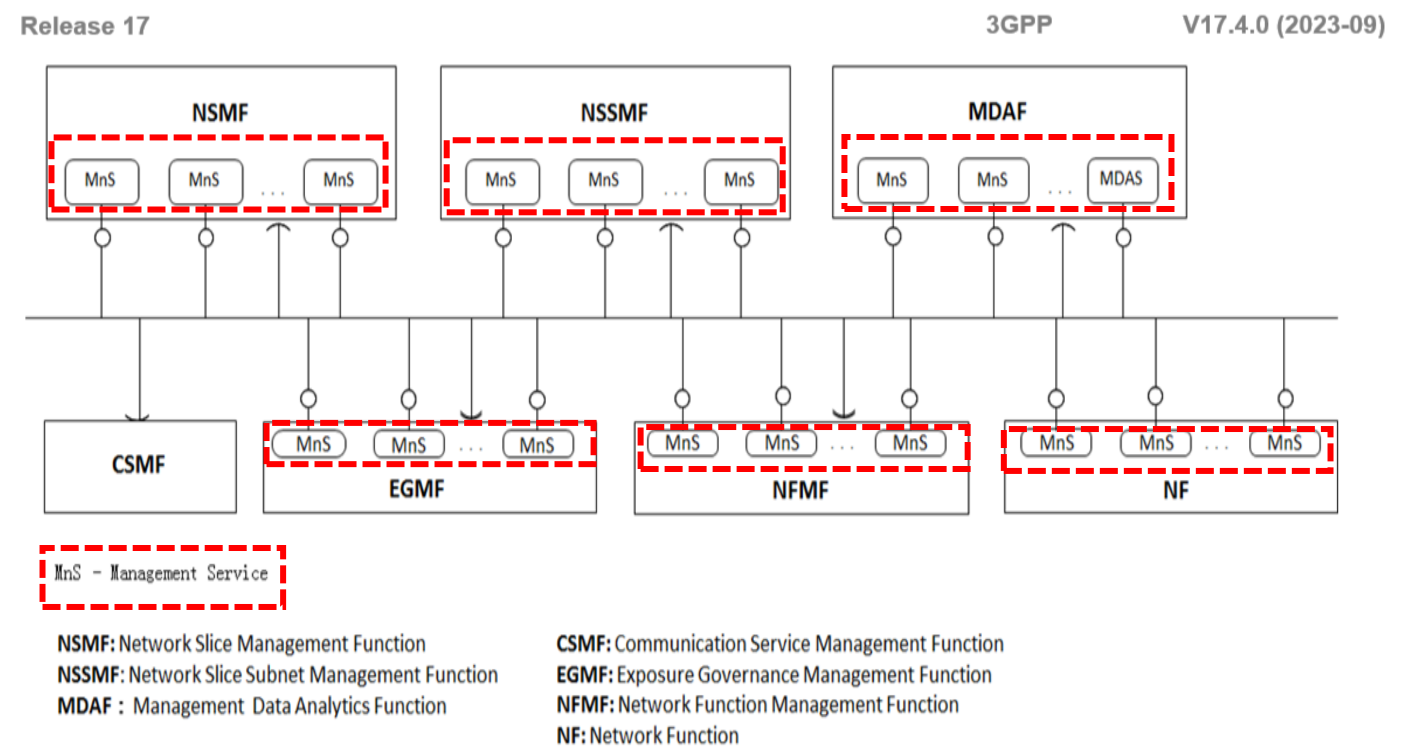

Release 17
3GPP
V17.4.0 (2023-09)
Entity C
The network management service X is consumed
The network management service X is produced (exposed)
Entity B
The network management service X is consumed
The network management service X is produced
Entity A
Entity C
The network management service Y is consumed
The network management service Y is produced (exposed)
Entity B
The network management service X is consumed
The network management service X is produced
Entity A

Figure: 5G System Management Architecture Framework Reference Model for Management Service (MnS) Producers, Consumers and Exposure

Utilization of Management Services (MnSs) in Functional Management Architecture

The Management Services (MnSs) for a Mobile Network including Network Slicing (SST) may be produced by a Set of Functional blocks.

An example of such Deployment Scenario where Functional Blocks (such as NSMF, NSSMF, NFMF and CSMF) are producing and consuming various Management Services is illustrated in the Figure.



NSMF: Network Slice Management Function
NSSMF: Network Slice Subnet Management Function
MDAF : Management Data Analytics Function
CSMF: Communication Service Management Function
EGMF: Exposure Governance Management Function
NFMF: Network Function Management Function
NF: Network Function

Figure: 5G System Functional Management Architecture Management Services example

2. 5G System use of AI/ML

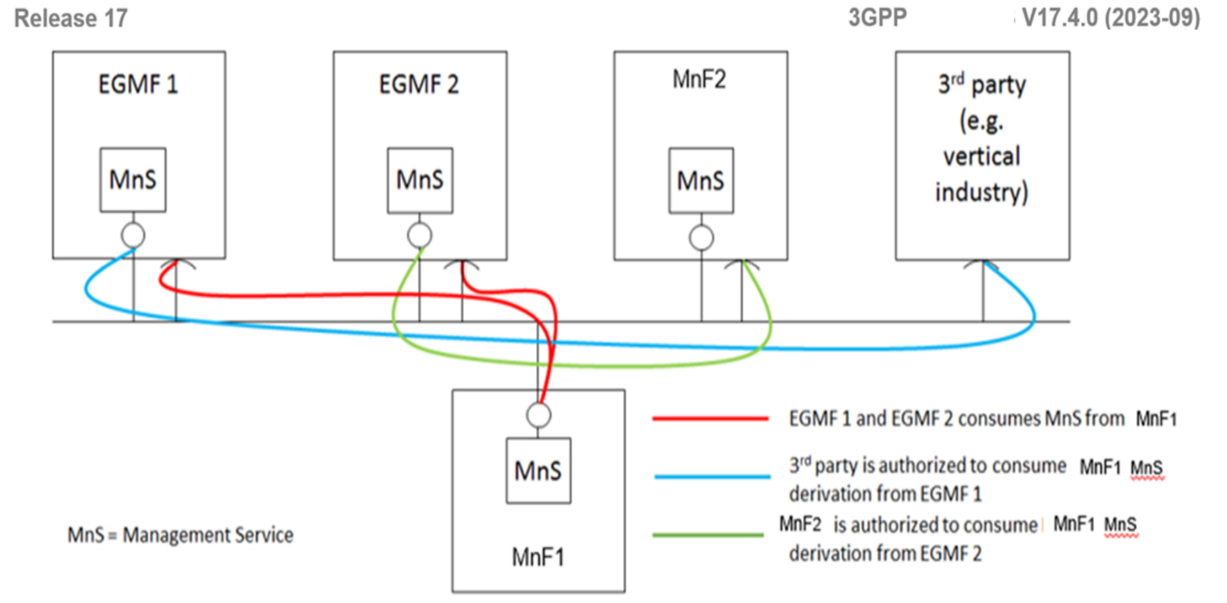
Utilization of *5GS Management Services (**MnSs**) by Exposure Governance Management Function (**EGMF**)*

*Exposure Governance Management Function (**EGMF**)* offers *Exposure Governance Management Capability* as illustrated in the Figure:

EGMF produces *Exposure Governance Management Capability* that Operators can apply on *Management Function (**MnF) 1 MnS*** for exposing different *derivation of MnF 1 MnS* to:

- *MnF 2 (e.g. from another Operator*) and
- *3rd party (e.g. from Vertical Industry*).

NOTE: Exposure Governance can be controlled by a Policy for different Type of MnF 1 MnS Consumers (e.g. other Operator, other Management System, 3rd Party, Other Administrative Domain, etc.)



Release 17
3GPP
V17.4.0 (2023-09)
EGMF 1
MnS
EGMF 2
MnS
MnF2
MnS
3rd party (e.g. vertical industry)
MnS
MnF1
MnS = Management Service
EGMF 1 and EGMF 2 consumes MnS from MnF1
3rd party is authorized to consume MnF1 MnS derivation from EGMF 1
MnF2 is authorized to consume MnF1 MnS derivation from EGMF 2

Figure: 5G System Functional Management Architecture Management Services MnF-1 Management Service (MnS) exposed through Exposure Governance Management Function 1 (EGMF 1) and through Exposure Governance Management Function 2 (EGMF 2)

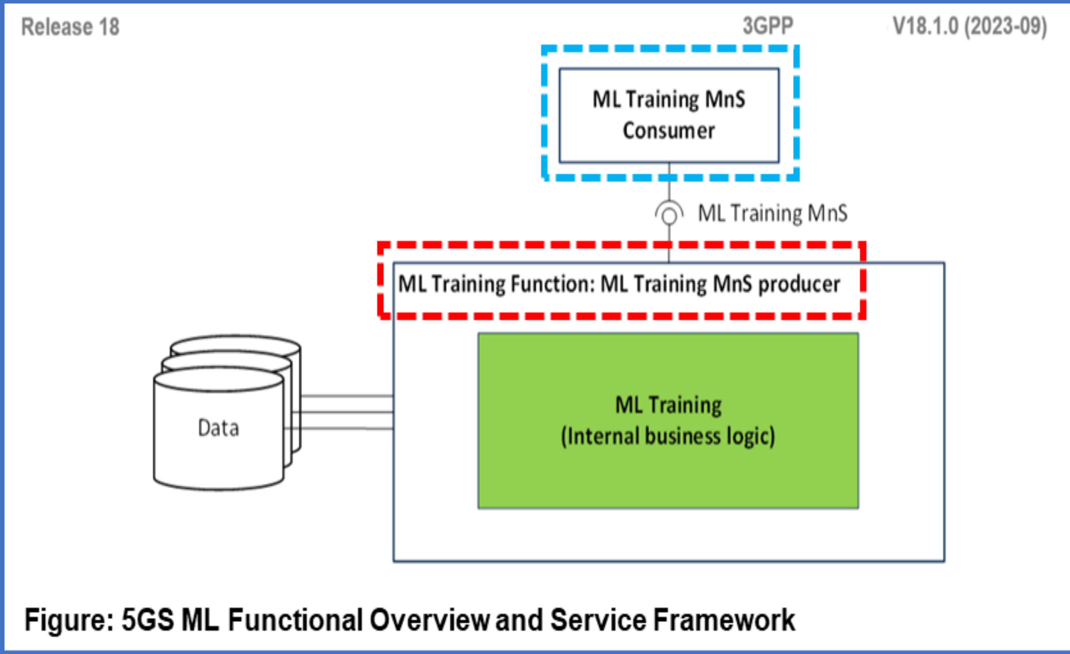
2. 5G System use of AI/ML

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.



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

ML Training MnS Consumer
ML Training MnS
ML Training Function: ML Training MnS producer
Data
ML Training
(Internal business logic)

Figure: 5GS ML Functional Overview and Service Framework

2. 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.

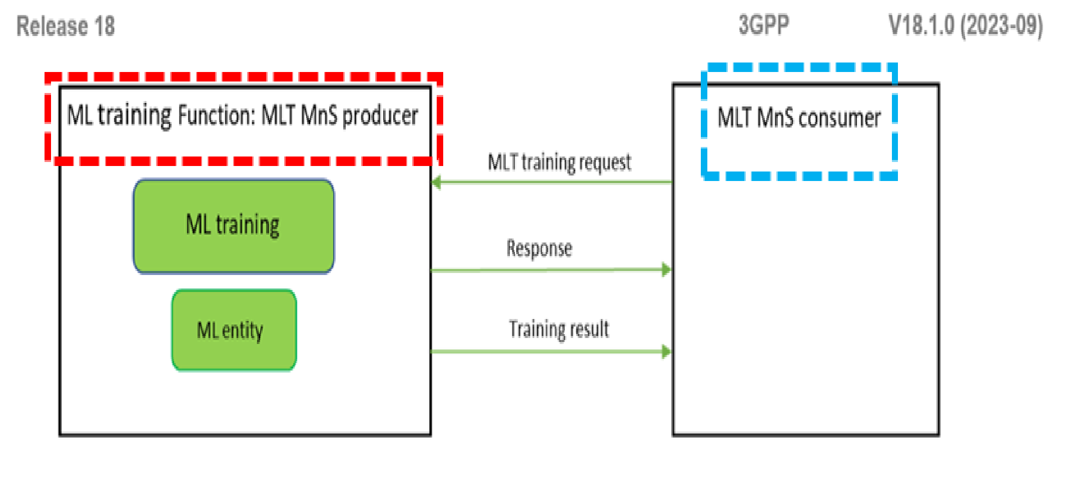


Figure: 5GS ML Training requested by ML MnS Consumer

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:

2. 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).

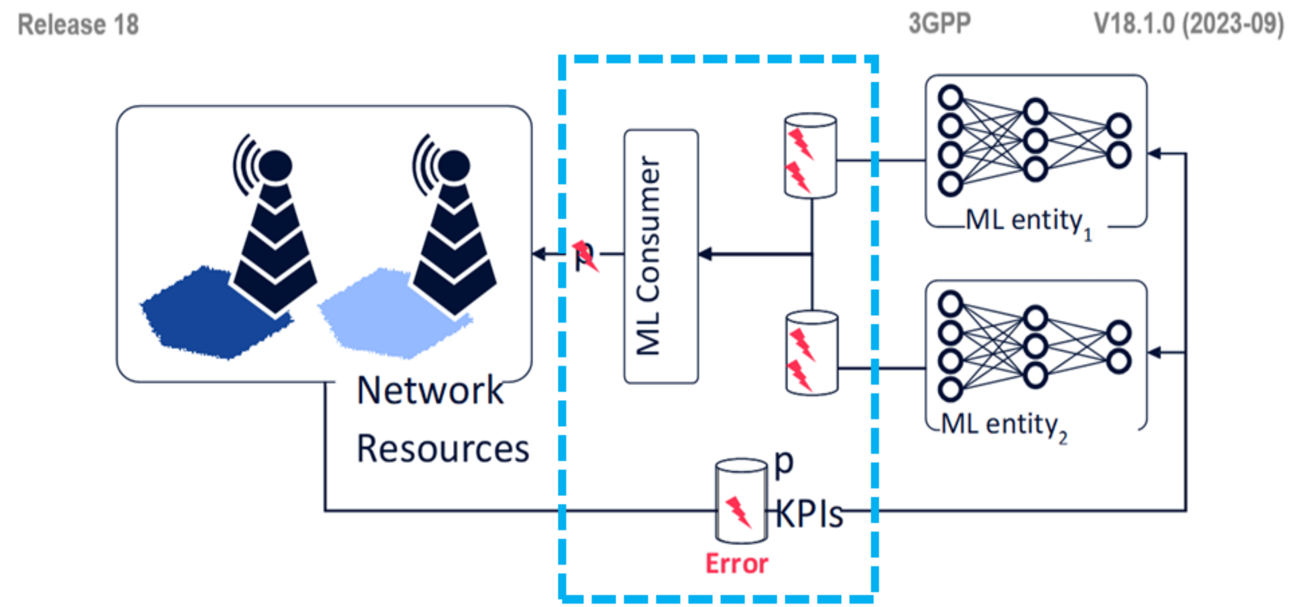


Figure: 5G System AI/ML Management example of Network Resource Propagation of Erroneous Information

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.

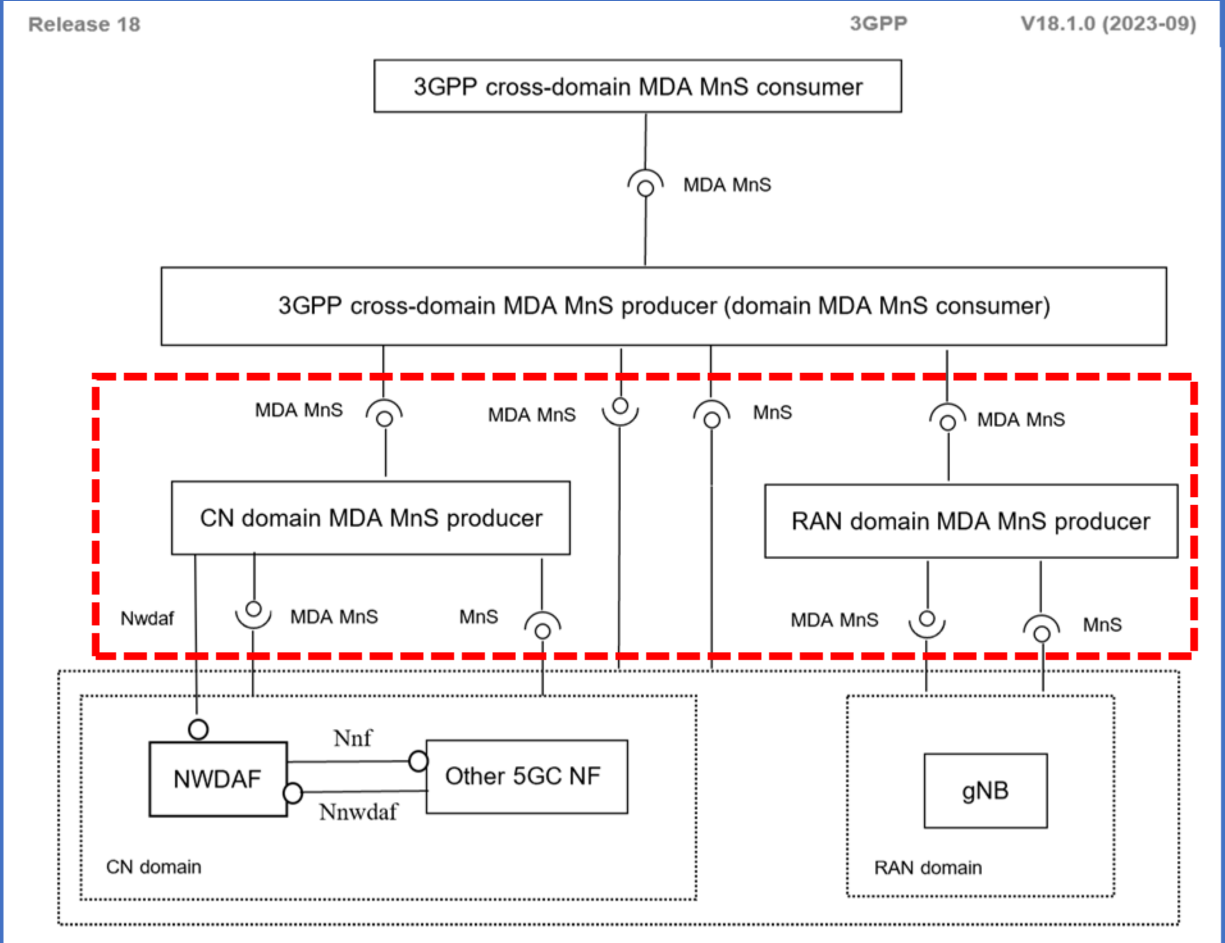
2. 5G System use of AI/ML

Interaction with 5G CN and RAN Domains

The MDA MnS Producer provides Analytics Data for Management Purposes based on Input Data related to different types of NFs or Entities in the Network, e.g. Data reported from gNB and/or specific Core Network Function(s).

Depending on the Use Case (UC) and when needed, the MDA MnS Producer may use the Analytics results produced by NWDAF as Input.

Management Data Analytics Function (MDAF) may act as 3GPP Domain-specific (e.g. RAN or CN) or as 3GPP Cross-Domain MDA MnS Producer. The Figure illustrates the example of co-ordination between NWDAF, gNB and MDA MnS producer(s) for Data Analytics Purpose.



Release 18
3GPP
V18.1.0 (2023-09)
3GPP cross-domain MDA MnS consumer
MDA MnS
3GPP cross-domain MDA MnS producer (domain MDA MnS consumer)
MDA MnS
MDA MnS
MnS
MDA MnS
CN domain MDA MnS producer
RAN domain MDA MnS producer
Nwdaf
MDA MnS
MnS
MDA MnS
MnS
Nnf
NWDAF
Other 5GC NF
Nnwdaf
gNB
CN domain
RAN domain

Figure: 5G System Architecture Management Data Analytics (MDA) example of co-ordination between NWDAF, gNB and MDAS (MDA MnS) Producer

2. 5G System use of AI/ML

Deployment of Multiple MDAs

The 3GPP Cross Domain Management may consume MDA MnS provided by RAN Management as shown in the Figure.

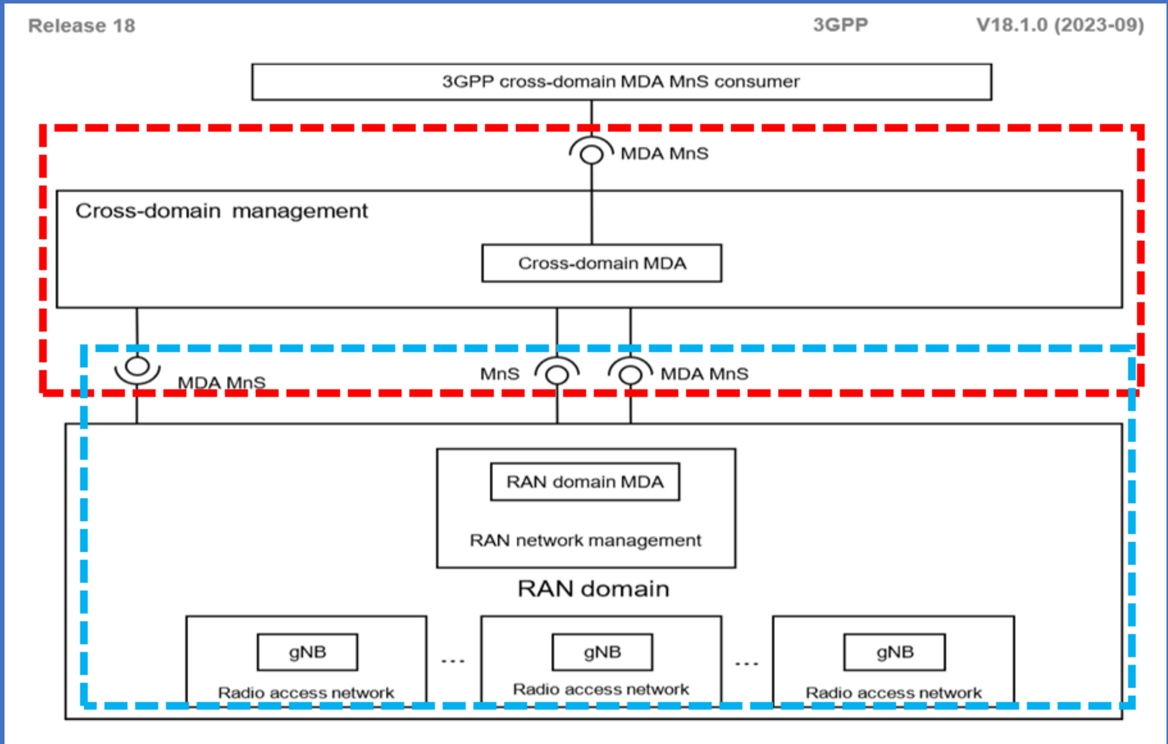
The Management Function (MDAF) playing the Role of *3GPP Cross Domain MDA MnS Producer* interacts with *RAN Domain MDA* per each MDA Use Case/Capability as follows:

- The Cross Domain MDA MnS Producer may consume the RAN Domain MDA MnS.

- The Cross Domain MDA MnS Producer may consume MnS provided by RAN Domains, and produce MDA MnS that may be consumed by 3GPP Cross-Domain MDA MnS Consumer(s).

The Management Function (MDAF) playing the Role of RAN Domain MDA MnS Producer interacts with MnS Producers per each Use Case/Capability as follows:

- The RAN Domain MDA MnS Producer may consume MnS provided by RAN Domain Management, other MDA MnS Producers, Management Data derived by Sub-Network Management Function(s), and Management Data derived by element Management Function(s).



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

Figure: 5G System Architecture Management Data Analytics (MDA) example of coordination Cross-domain MDA and RAN Domain MDA

2. 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).

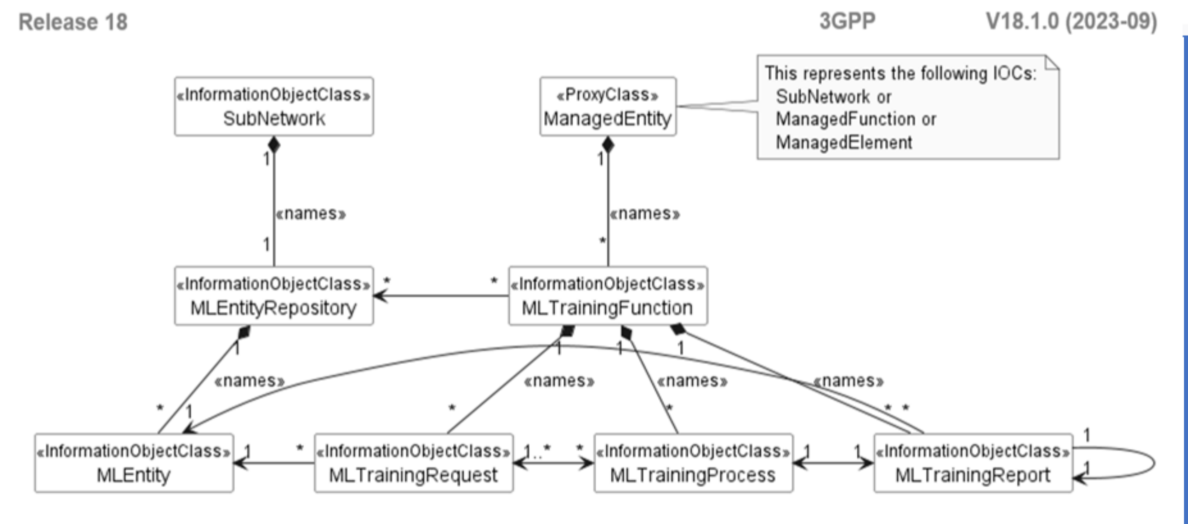

Release 18
3GPP
V18.1.0 (2023-09)
«InformationObjectClass» SubNetwork
«ProxyClass» ManagedEntity
This represents the following IOCs: SubNetwork or ManagedFunction or ManagedElement
«names»
«names»
1
1
1
*
«InformationObjectClass» MLEntityRepository
«InformationObjectClass» MLTrainingFunction
«names»
«names»
«names»
«names»
«InformationObjectClass» MLEntity
«InformationObjectClass» MLTrainingRequest
«InformationObjectClass» MLTrainingProcess
«InformationObjectClass» MLTrainingReport
1
*
1..*
*
1
1
1
1

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

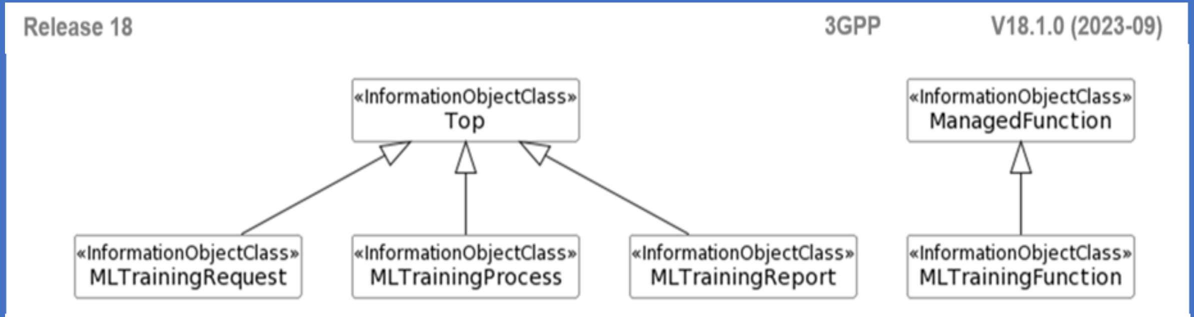
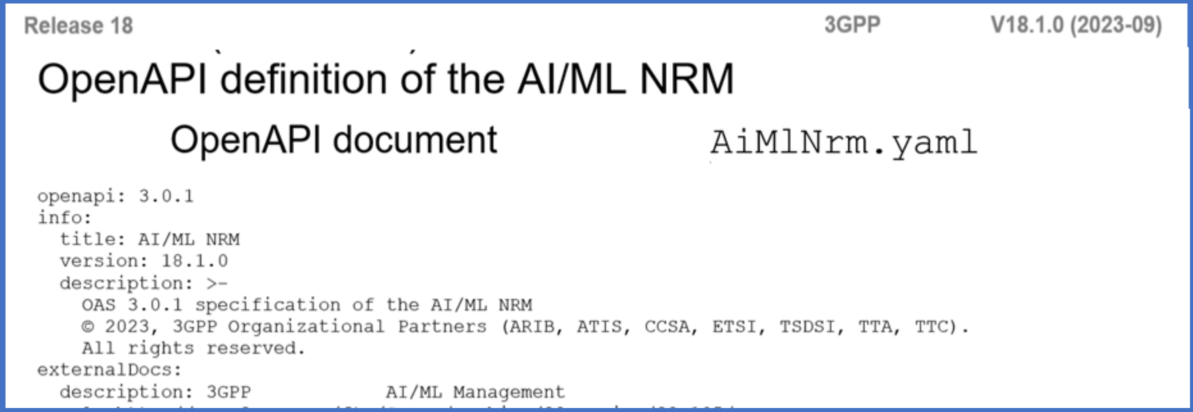

Release 18
3GPP
V18.1.0 (2023-09)
«InformationObjectClass» Top
«InformationObjectClass» ManagedFunction
«InformationObjectClass» MLTrainingRequest
«InformationObjectClass» MLTrainingProcess
«InformationObjectClass» MLTrainingReport
«InformationObjectClass» MLTrainingFunction

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



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
```

2. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - 1

5GS Architecture specification envisages a Set of High-Level Procedures by which Data is collected by a **Network Data Analytics Function** (**NWDAF**) from *UE Application(s)* via an intermediary *Application Function (AF)*.

The Data Collection AF (*DCAF*) may support 5G Architecture Common API Framework (CAPIF) to provide APIs to other Applications (i.e. API Invokers), as defined in 5GS Architecture.

NOTE 1: It is presumed that the User (Resource Owner) has granted "Consent" for its UE Data to be collected, reported and subsequently exposed through interactions with the MNO or the Application Service Provider (ASP), and via any applicable SLA between the MNO and Application Service Provider (ASP).

See on the next slide the Table showing the set User Consent for Data Collection client API Method as specified in 5GS Architecture.

NOTE 2: *The Collection, Reporting and Exposure of Location-based UE Data is expected to comply with Regional Regulatory Requirements and may be further limited by MNO Policy.*

This reference architecture is intended to be instantiated in Domain-specific ways to suit the needs of different features of the 5G System as e.g. the Reference Architecture may be instantiated separately in **different Slices (SST) of a Network.**

Each type of UE Data subject to Collection, Reporting and subsequent Event Exposure in the 5G System is associated with a Logical UE Data domain.

Each such UE Data Domain is associated with a Domain Owner – either the 5G System itself (embodied in a particular deployment by an MNO) or the Application Service Provider (ASP).

Precedence rules on the Exposure (and consequent Collection and Reporting) of UE Data vis-à-vis conflicts between ASP Provisioning Information and System pre-configuration by the MNO or

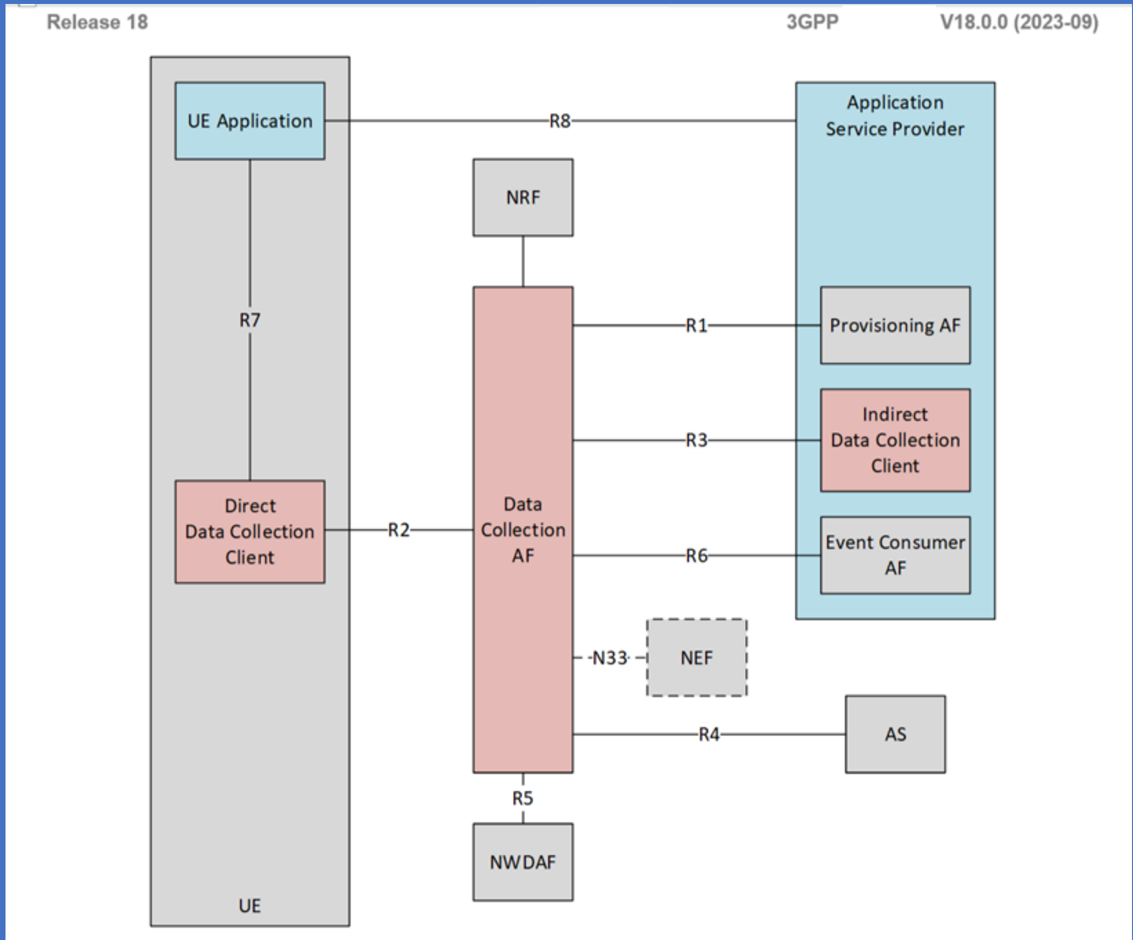

Release 18 3GPP V18.0.0 (2023-09)

Figure: 5G System Reference Architecture for Data Collection and Reporting in 5G SBA Reference Point notation

Note: *The Data Collection AF (**DCAF**) may be deployed outside the trusted domain, in which case the Services it exposes to API Invokers are mediated by the 5G CN NEF node. The Logical Relationships denoted by the Reference Points are unaffected by such deployment choices.*

2. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - 2

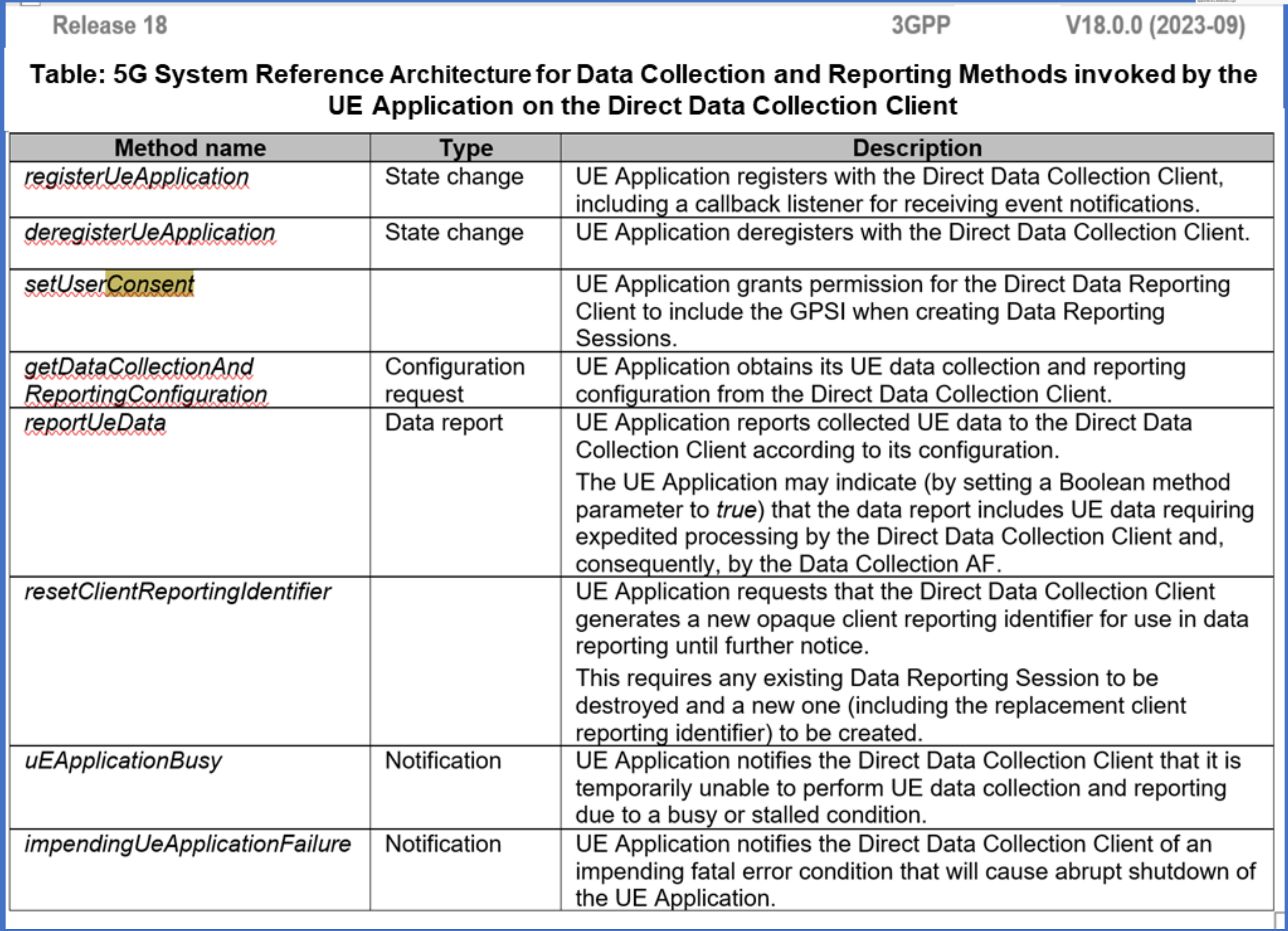
Application registration procedure
Upon activation, the UE Application requests its UE Data Collection and reporting Configuration from the Direct Data Collection Client by invoking the *registerUeApplication* Method at Reference Point *R7*.

The UE Application provides as input parameters its
- External Application Identifier,
- Application Service Provider identifier, and
- Information on its callback listener (for receiving notifications from the Direct Data Collection Client).

The UE Application also indicates its "*consent*" for the *UE Identity (i.e. GPSI*) to be included in *Data Reports sent to the Data Collection AF.*

The *Direct Data Collection Client* establishes a new Data Reporting Session with the *Data Collection AF* using the Procedure specified in the 5GS Reference Architecture.

The *Ndcaf_DataReporting_CreateSession* invocation includes the GPSI of the UE (if consent is given by the UE Application) or otherwise the Direct Data Collection Client shall instead generate an opaque Client reporting Identifier that is Globally Unique and stable (e.g. a UUID) and include this in the invocation of the Service operation.



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. 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. 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.

Procedure for changing Consent to report the UE identifier

The UE Application can change its Consent to reveal the *GPSI of the UE in Data Reports* sent to *the Data Collection AF* during the course of a Data reporting session by invoking the *setUserConsent M*ethod on the *Direct Data Collection Client* at *Reference Point R7.*
The Direct Data Collection Client shall destroy the current Data Reporting Session and create a new one that includes **either** *the GPSI of the UE* ***or*** the *Opaque Client Reporting Identifier*, according to ***whether Consent is granted or withdrawn***.

2. 5G System use of AI/ML

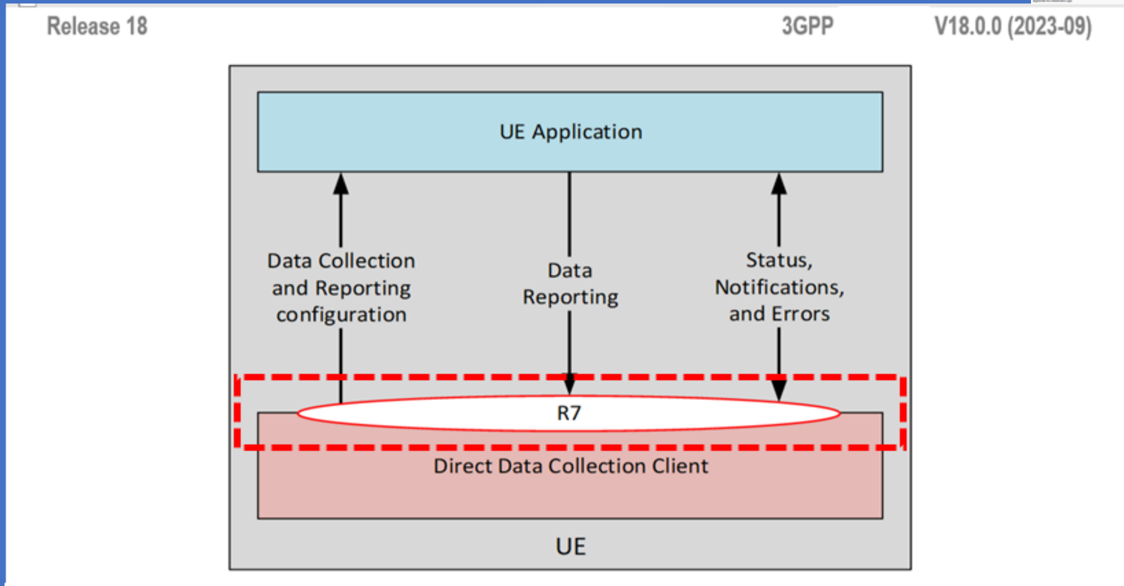
5G System Data Collection and Analytics Reference Architecture - 3

UE Data Collection, Reporting and Notification API

The 5GS Data Collection and Reporting Reference Architecture specifies:
- UE Data Collection, Reporting and Notification API used by internal UE Entities, namely a *UE Application* and the *associated Direct Data Collection Client*, in support of *UE Data Collection* by the *Direct Data Collection Client* for subsequent *reporting to the Data Collection AF*, and related exchange of notifications.

As noted in the Reference Architecture specification, this API is not used when the *Direct Data Collection Client is embedded in the UE Application* (i.e., Collaboration between UE and the DCAF as specified) (see the Figure on "Collaboration" and the text below).

However, this can serve as "guidance" to the Design of the Internal APIs for a UE Application with an embedded Direct Data Collection Client.



Release 18
3GPP
V18.0.0 (2023-09)
UE Application
Data Collection and Reporting configuration
Data Reporting
Status, Notifications, and Errors
R7
Direct Data Collection Client
UE

Figure: 5G System Reference Architecture for Data Collection and Reporting UE Architecture for Data Collection, Reporting and Notification via R7 API

5GS Data Collection & Reporting Architecture Collaboration between UE and DCAF

As specified in this scenario, the *Data Collection Client* is deployed as a sub-Function of the *UE Application*. Therein, *Reference Point R7* is subsumed into the *UE Application*.

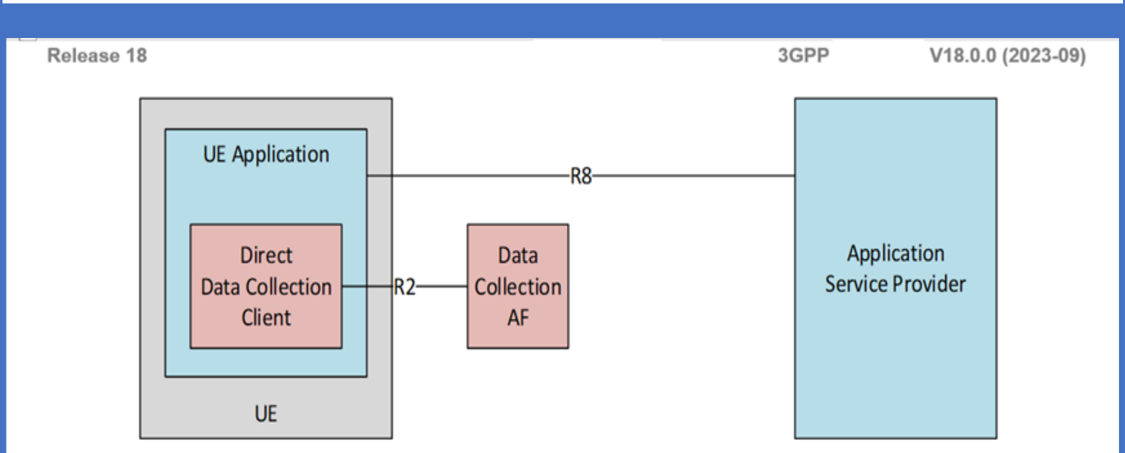


Figure: 5G System Reference Architecture for Data Collection and Reporting Collaboration with UE Data Collection Client deployed as part of the UE Application

The *Direct Data Collection Client* could, e.g., be realized as a SW Library that implements the appropriate Protocol at *Reference Point R2*. In such a realization, the Procedures defined in *Reference Point R7* would likely form *the API of the Data Collection Client Library.*

2. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - 4

The Figure depicts the case where the *Data Collection AF* (*DCAF*) is instead deployed outside the Trusted Domain, along with the Application Service Provider (ASP) and the (external) AS (Application Server).

In this case, *the sub-functions of the Application Service Provider (ASP*) and the (external) AS do not interact with the *Data Collection AF (DCAF*) via the *5G System Service bus*.

The *Ndcaf Service* is therefore not required in such deployments.

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

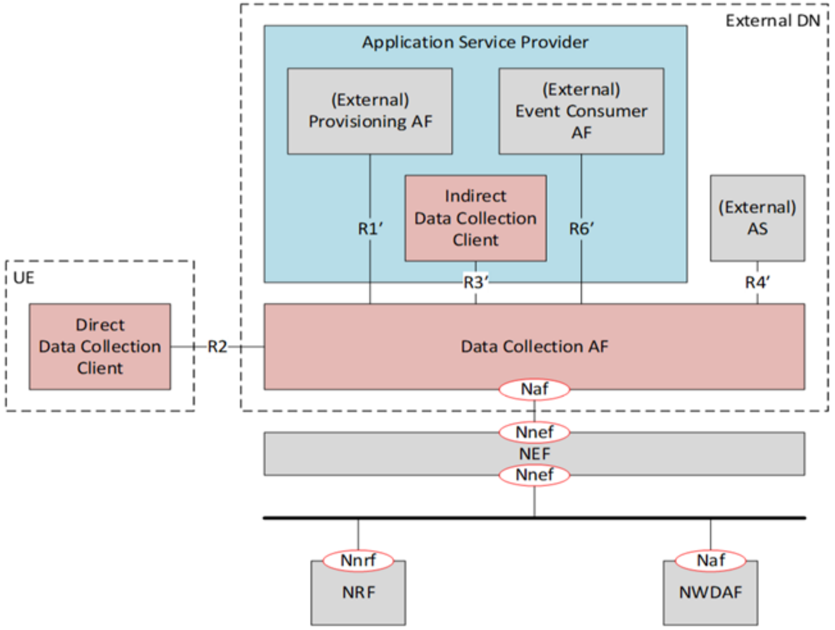

External DN
Application Service Provider
(External) Provisioning AF
(External) Event Consumer AF
Indirect Data Collection Client
R1'
R6'
(External) AS
UE
R3'
R4'
Direct Data Collection Client
R2
Data Collection AF
Naf
Nnef
NEF
Nnef
Nnrf
NRF
Naf
NWDAF

Figure: 5G System Reference Architecture for Data Collection and Reporting in 5G SBA notation when the Data Collection AF is deployed outside the Trusted Domain

2. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - 5

5GS Reference **Architecture Data Collection Domain Model(s)**

The Figure depicts the Static Data Model for the Data Collection and Reporting Domain and is further described in the Figure:

5GS Architecture Service exposure via Common API Framework (CAPIF) for Northbound APIs

When CAPIF is supported in the specified 5G Network configuration, then:
- the Data Collection AF shall *support the CAPIF API Provider Domain* functions as part of a distributed CAPIF deployment, i.e. Ndcaf and Naf via CAPIF 2/2e; and CAPIF 3, CAPIF 4 and CAPIF 5, as specified in 5G Common API Framework Architecture specification;

- the *Data Collection AF* shall support the CAPIF Core Function (CCF) and API provider domain functions as part of a centralized CAPIF deployment, i.e. Ndcaf and Naf via CAPIF 2/2e, as specified in 5G Common API Framework Architecture specification.

The *CAPIF and associated API provider domain functions* are specified in 5G Common API Framework Architecture specification.

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

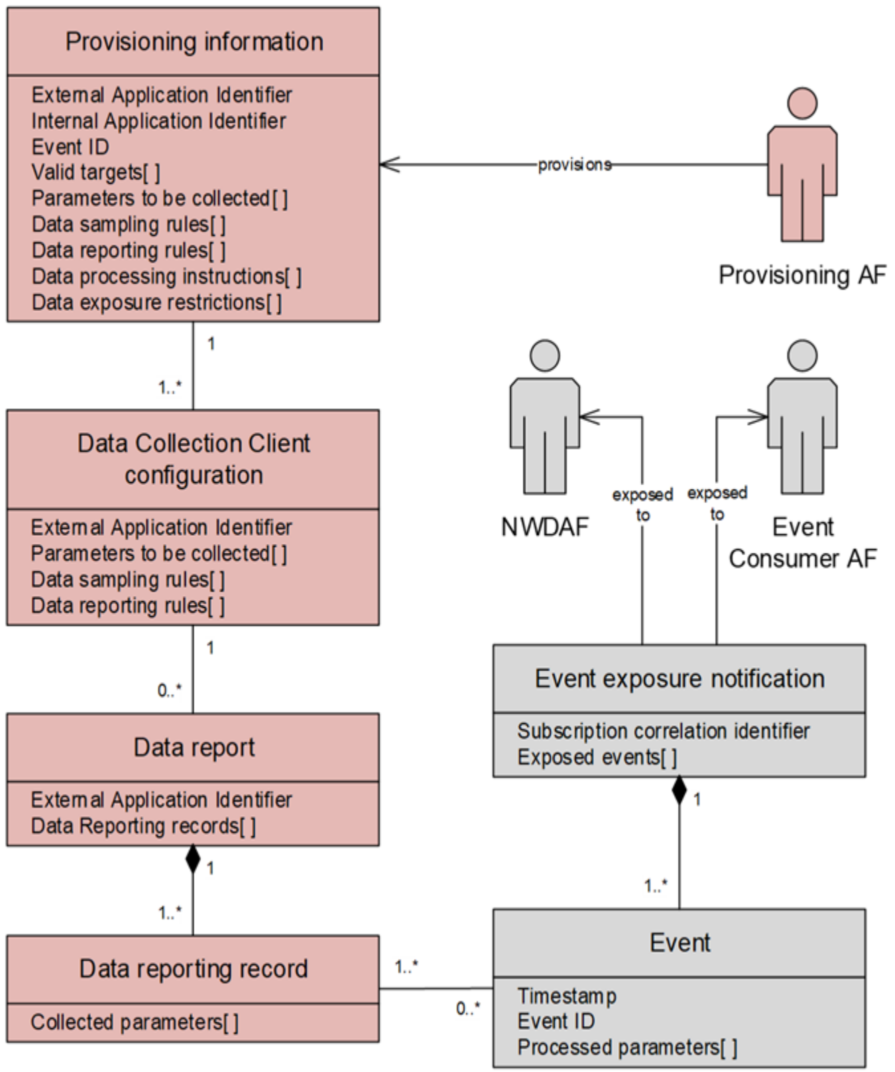

Provisioning information
External Application Identifier
Internal Application Identifier
Event ID
Valid targets[ ]
Parameters to be collected[ ]
Data sampling rules[ ]
Data reporting rules[ ]
Data processing instructions[ ]
Data exposure restrictions[ ]
provisions
Provisioning AF
1
1..*
Data Collection Client configuration
External Application Identifier
Parameters to be collected[ ]
Data sampling rules[ ]
Data reporting rules[ ]
NWDAF
exposed to
exposed to
Event Consumer AF
1
0..*
Event exposure notification
Subscription correlation identifier
Exposed events[ ]
Data report
External Application Identifier
Data Reporting records[ ]
1
1..*
1
1..*
Data reporting record
Collected parameters[ ]
1..*
0..*
Event
Timestamp
Event ID
Processed parameters[ ]

Figure: 5G Data Reporting and Analytics Reference Architecture Static Domain Model

2. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture: *Network Layer* - 6

The *5G System Architecture* allows any ***5GC NF*** to request *Network Analytics Information from* ***NWDAF*** *(Network Data Analytics Function*) containing Analytics Logical Function (**AnLF**). *The NWDAF* belongs to the same *PLMN as the 5GC NF* that consumes the Analytics information.

The *Nnwdaf interface* is defined for *5GC NFs*, to:

- Request *Subscription* to Network Analytics Delivery for a particular Context,

- Cancel Subscription to Network Analytics Delivery and to request a specific report of network analytics for a particular context.

NOTE 1: The 5G System Architecture also allows other "*Consumers*" such as *OAM and CEF (Charging Enablement Function*) to request Network Analytics information from NWDAF.

The 5G System Architecture allows any NF to obtain Analytics from an NWDAF using a *DCCF (Data Collection and Coordination Function)* with associated *Ndccf Services*, as specified.

The *5G System Architecture* allows *NWDAF and DCCF* to request *Historical Analytics from an NWDAF* with associated *Nnwdaf_DataManagement Services* as specified.

The *5G System Architecture* allows ***MFAF*** to fetch *Historical Analytics* from an ***NWDAF*** with associated *Nnwdaf_DataManagement Service* as specified.

As depicted in the Figure, the *Ndccf interface* is defined for *any NF to support Subscription Request(s) to Network Analytics,* to cancel subscription for Network Analytics and to request a Specific Report of Network Analytics.

If the Analytics is not already being collected, the *DCCF* requests the Analytics from the *NWDAF* using *Nnwdaf Services*. The *DCCF* may collect the Analytics and deliver it to the *NF*, or the *DCCF* may rely on a Messaging Framework to collect Analytics and deliver it to the NF.

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

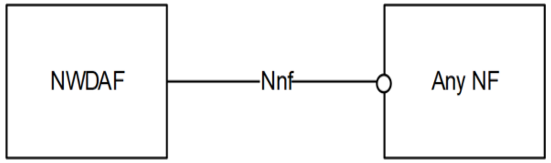


Figure: 5G System Data Analytics Collection and Reporting Architecture from any 5G Core Network Function (NF)

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

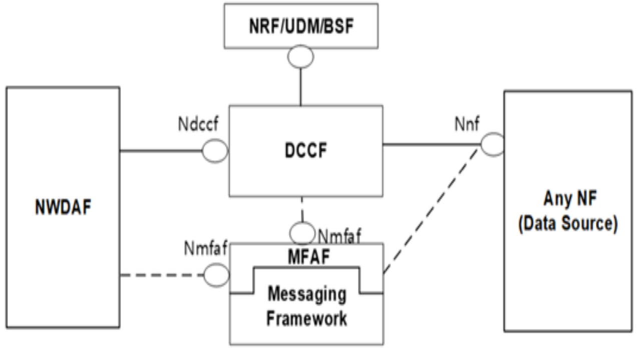


Figure: 5G System Data Analytics Collection and Reporting Architecture using Data Collection Co-ordination

2. 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).*

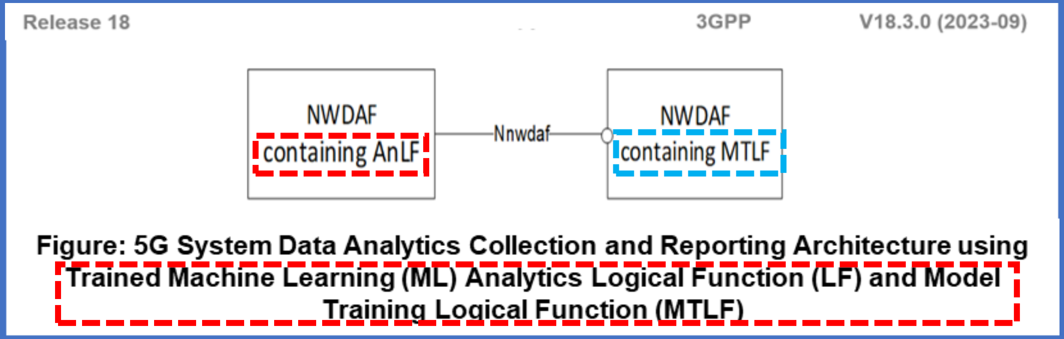

Release 18 3GPP V18.3.0 (2023-09)

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

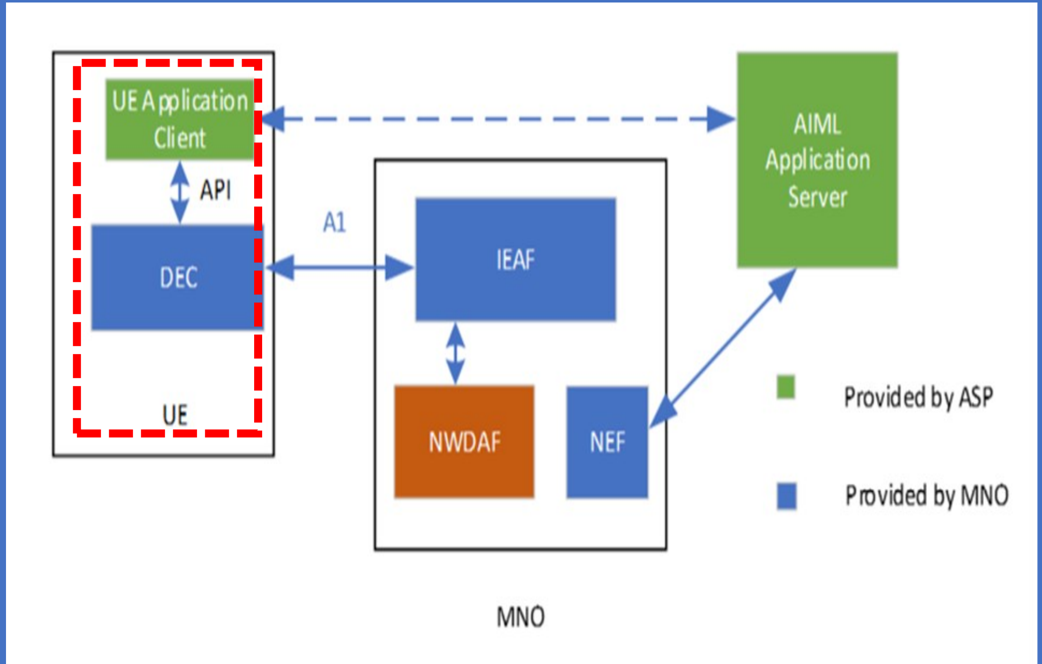


Figure: 5G Data Analytics, Collection and Reporting Architecture evolved IEAF Data Information Collection Function

2. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - ***use of AI/ML*** - 8

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:

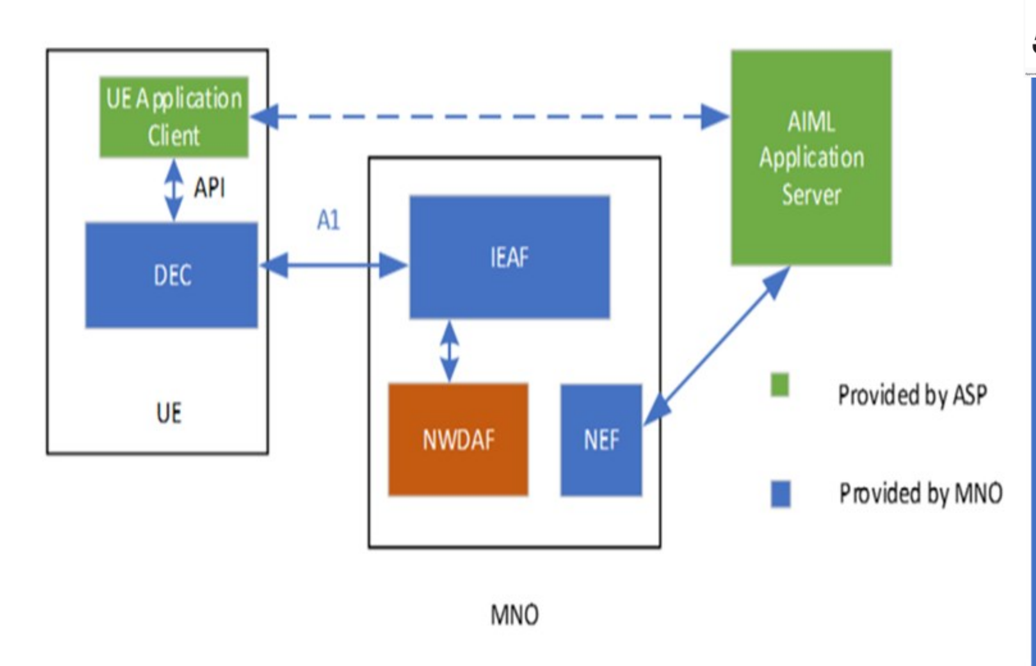


Figure: 5G Data Analytics, Collection and Reporting Architecture evolved IEAF Data Information Collection Function

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.

2. 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.*

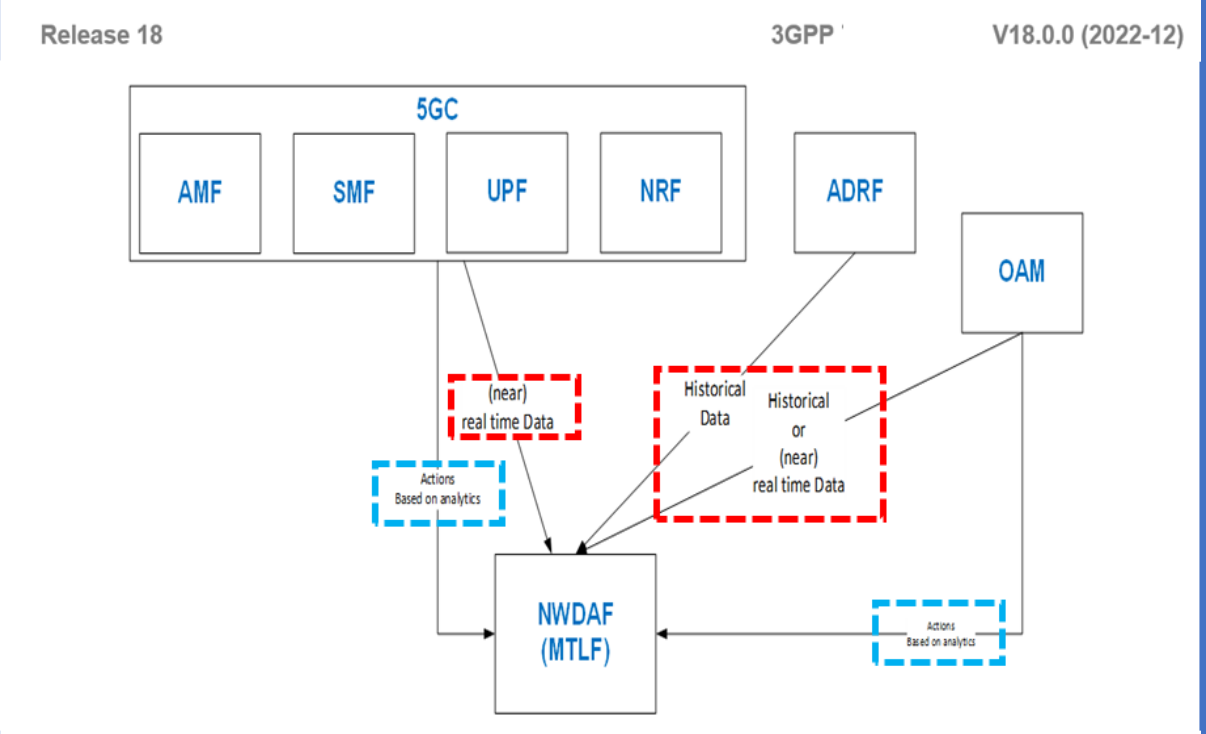


Figure: 5G System Data Analytics Collection and Reporting Architecture Model drift detected at Network Data Analytics Function (NWDAF) Model Training Logical Function (MTLF)

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".*

2. 5G System use of AI/ML

5G System Data Collection and Analytics Reference Architecture - 10

Roaming Capability Architecture

Based on Operator's *Policy* and Local Regulations (e.g. *Privacy*), *Data* or *Analytics* may be *exchanged between PLMNs* (i.e. *HPLMN* and *VPLMN*).

In a *PLMN*, an *NWDAF* is used as exchange point to exchange Analytics and to collect *Input Data for Analytics* with other *PLMNs*.

The NWDAF with Roaming exchange Capability is called *Roaming Exchange NWDAF (RE-NWDAF).*

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

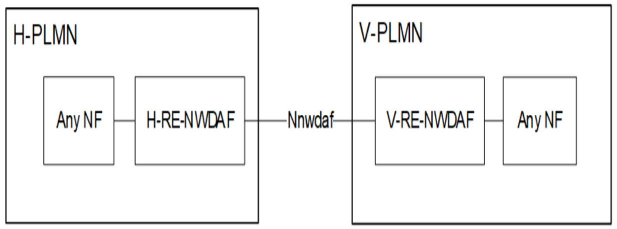


Figure: 5G System Data Analytics Collection and Reporting Roaming Architecture to exchange Input Data or Data Analytics between V-PLMN and H-PLMN

Using the Architecture shown in the Figure:

- For *Outbound Roaming Users*, the *NF "Consumer*" in the *HPLMN* can retrieve *Analytics* from the *VPLMN* via the *H-RE-NWDAF in HPLMN* and *V-RE-NWDAF in VPLMN.*

NOTE 1: *The Analytics from the VPLMN may be generated by the V-RE-NWDAF in the VPLMN or by other NWDAFs in the VPLMN.*

- *For Outbound Roaming Users*, the *H-RE-NWDAF in HPLMN* can collect *Data* from the *VPLMN via V-RE-NWDAF in VPLMN*.

- For *Inbound Roaming Users*, *the NF "Consumer*" in the *VPLMN* can retrieve *Analytics* from the *HPLMN* via *V-RE-NWDAF in VPLMN and H-RE-NWDAF in HPLMN.*

NOTE 2: The Analytics from the HPLMN may be generated by H-RE-NWDAF in the HPLMN or other NWDAFs in the HPLMN. For Inbound Roaming Users, the V-RE-NWDAF can collect data from the HPLMN via the H-RE-NWDAF.

NOTE 3: *Both Local Breakout (LBO) and Home Routed (HR) Roaming Architectures support the Data or Analytics exchanging between PLMNs.*

NOTE 4: Interactions between RE-NWDAFs of different PLMNs may be via ***SEPPs,*** which are not depicted in the Architecture for the sake of clarity.

2. 5G System use of AI/ML

5G System ***Application Data Analytics Enablement Service*** **(ADAES) Architecture** (for VAL - *Application Layer*) - 1

5G System Architecture Application Data Analytics Enablement ***(ADAE) internal Architecture***

In ADAE Framework, **A-DCCF** and **A-ADRF** can be defined as Functionalities within the *internal ADAE Architecture* and can offer the following Functionalities:

- ***Application Layer*** - *Data Collection and Coordination Function (**A-DCCF**)*

A-DCCF coordinates the Collection and Distribution of Data requested by the "*Consumer" (ADAE Server*).

Data Collection Coordination (**DCC**) is supported by a ***A-DCCF.***
ADAE Server can send requests for Data to the ***A-DCCF*** rather than directly to the *Data Sources*.
A-DCCF may also perform Data Processing/Abstraction and Data Preparation based on the VAL Server Requirements.

- *Application Layer - Analytics and Data Repository Function (**A-ADRF**)* stores Historical Data and/or Analytics, i.e., Data and/or Analytics related to past time period that has been obtained by the *"Consumer* (*e.g. ADAE Server).*

After the "*Consumer"* obtains Data and/or Analytics, *"Consumer*" may store *Historical Data* and/or *Analytics in an A-ADRF*.

Whether *the "Consumer"* directly contacts *the A-ADRF* or goes via the *A-DCCF* is based on configuration.

The Figure illustrates *the Generic Functional Model for ADAE* when re-using the ***3GPP Network Data Analytics (NWDAF) Model.***

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

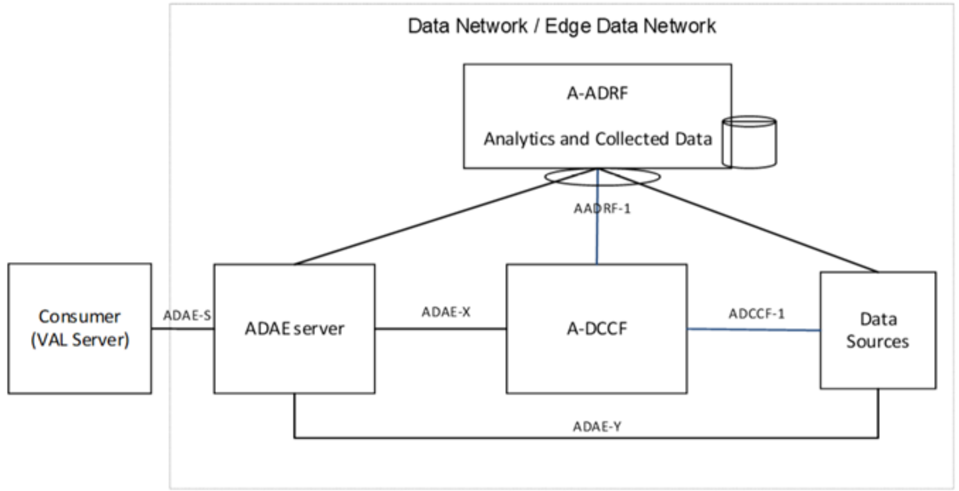


Figure: 5G System Data Analytics Collection and Reporting Architecture Application Data Analytics Enablement Internal Functional Architecture

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

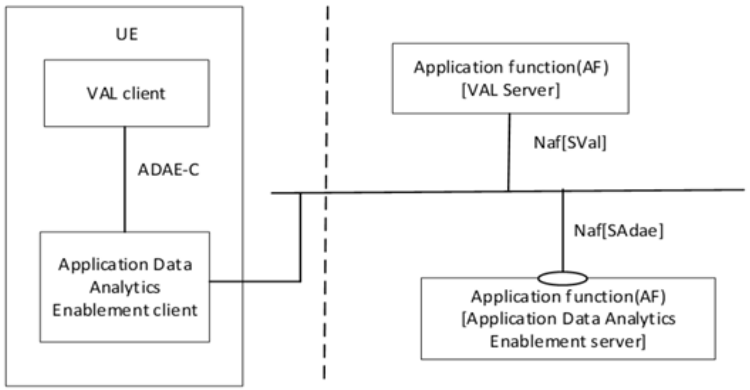

UE
VAL client
ADAE-C
Application Data Analytics Enablement client
Application function(AF) [VAL Server]
Naf[SVal]
Naf[SAdae]
Application function(AF) [Application Data Analytics Enablement server]

Figure: 5G System Architecture for Application Data Analytics Enablement in 5G Service-based Interface (SBI) Representation

2. 5G System use of AI/ML

5G System *Application Data Analytics Enablement Service* (ADAES) Architecture (for VAL - *Application*

5G System Architecture Application Data Analytics Enablement (ADAE) Deployment Scenarios

There could be three (3) ADAE Deployment Options:

1. *ADAES can be deployed at a Centralized Cloud Platform*, and collects **Data from multiple EDNs**

2. *ADAES can be deployed at the Edge Platform* (3GPP EDGEAPP)

3. *Coordinated ADAES deployment*, where multiple ADAE Services are deployed in Edge or Central Clouds.

Such deployment allows *for Local-Global Analytics for System wide optimization*

ADAE Layer APIs

The following ADAE Capabilities are offered as APIs:

- ADAE Server APIs;

- A-ADRF APIs;

The Service Enablement Architecture Layer and Network Slice capability Enablement Service APIs are specified and support:

- Group Management Server APIs;

- Location Management Server APIs;

- Configuration Management Server APIs;

- Identity Management Server APIs;

- Key Management Server APIs; and

- Network Slice Capability Enablement APIs.

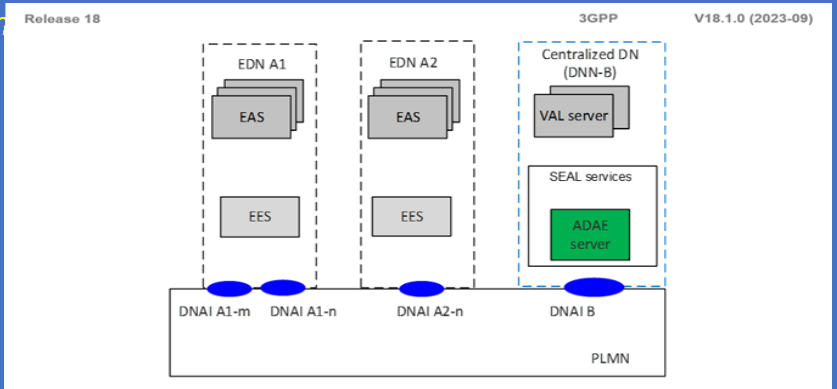


Figure: 5G System Architecture for Application Data Analytics Enablement Cloud deployment ADAE option

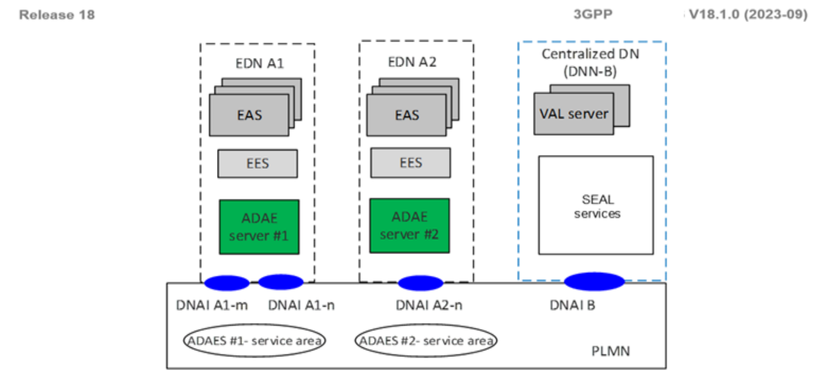


Figure: 5G System Architecture for Application Data Analytics Enablement 5G EDGEAPP Architecture deployed ADAE option

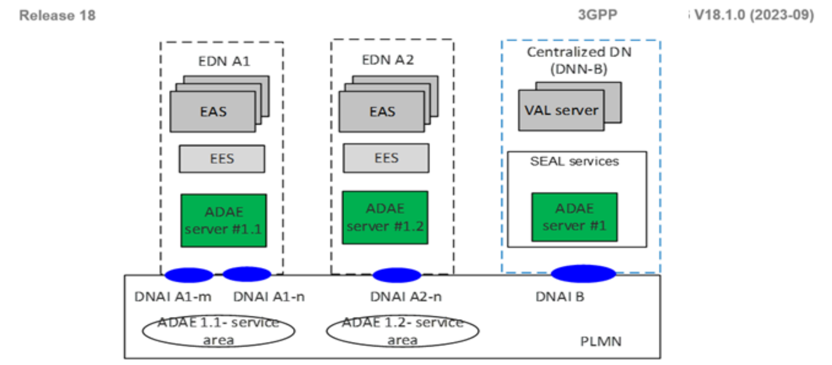


Figure: 5G System Architecture for Application Data Analytics Enablement Co-ordinated deployed ADAE option

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.*

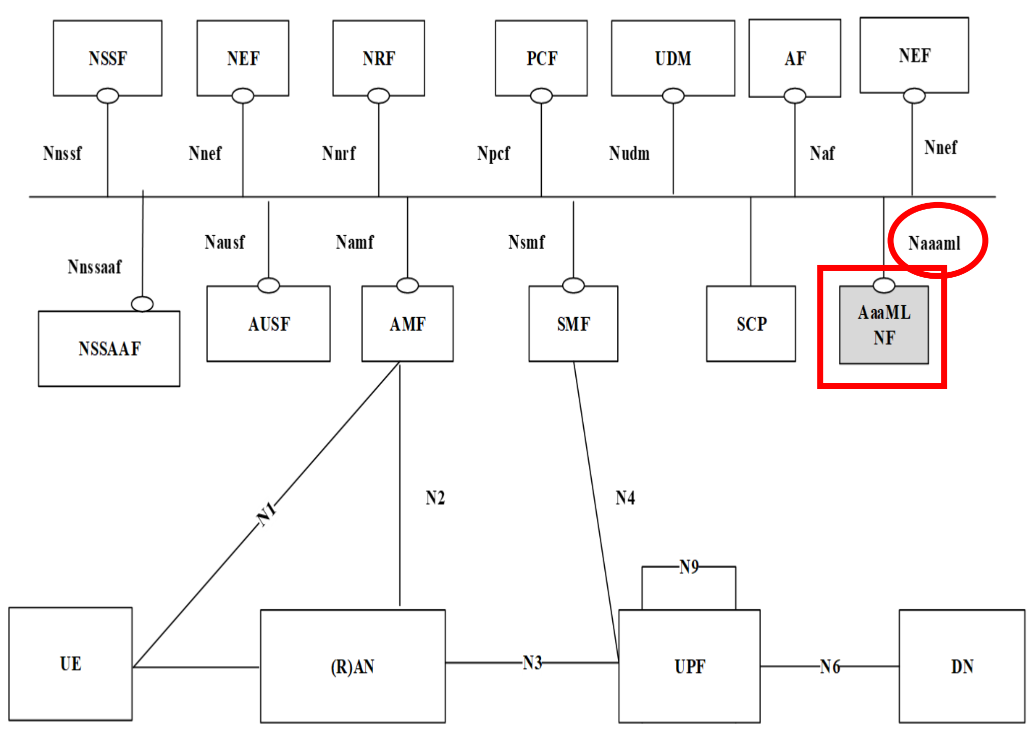
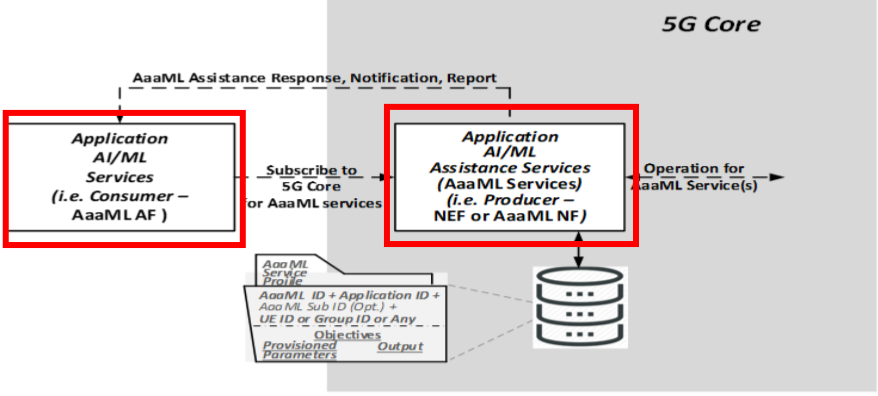


Figure: 5G System Service Architecture with AaaML NF



5G Core
AaaML Assistance Response, Notification, Report
Application AI/ML Services (i.e. Consumer – AaaML AF)
Subscribe to 5G Core for AaaML services
Application AI/ML Assistance Services (AaaML Services) (i.e. Producer – NEF or AaaML NF)
Operation for AaaML Service(s)
AaaML Service Profile
AaaML ID + Application ID + AaaML Sub ID (Opt.) + UE ID or Group ID or Any
Objectives
Provisioned Parameters
Output

Figure: 5G Application AI/ML Service Assistance Framework

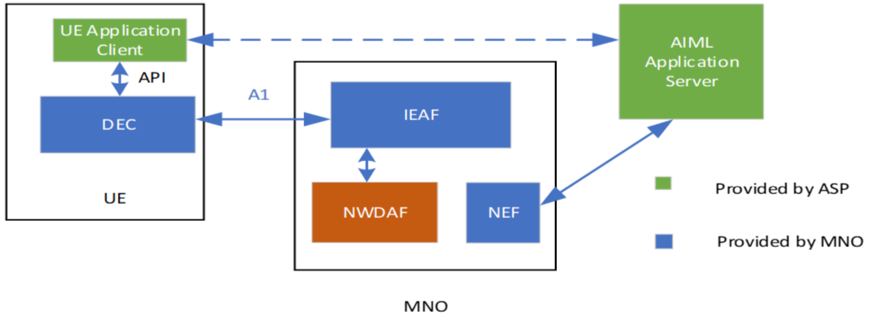


Figure: 5G IEAF (Data Information AF)

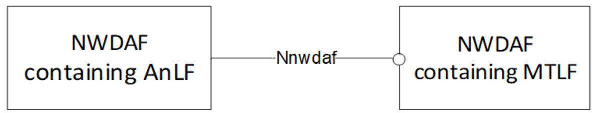


Figure: 5G NWDAF containing AnLF (Analytics Logical Function) for request and subscribe to ML Model Provisioning Architecture

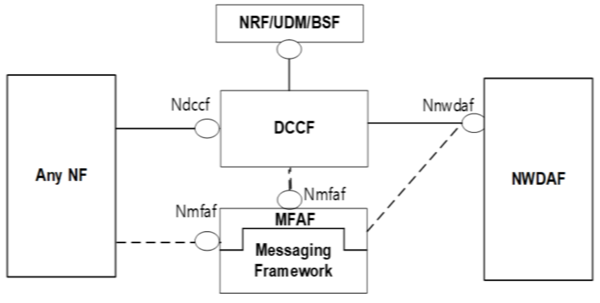


Figure: 5G Network Data Analytics Exposure Architecture using DCCF

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.

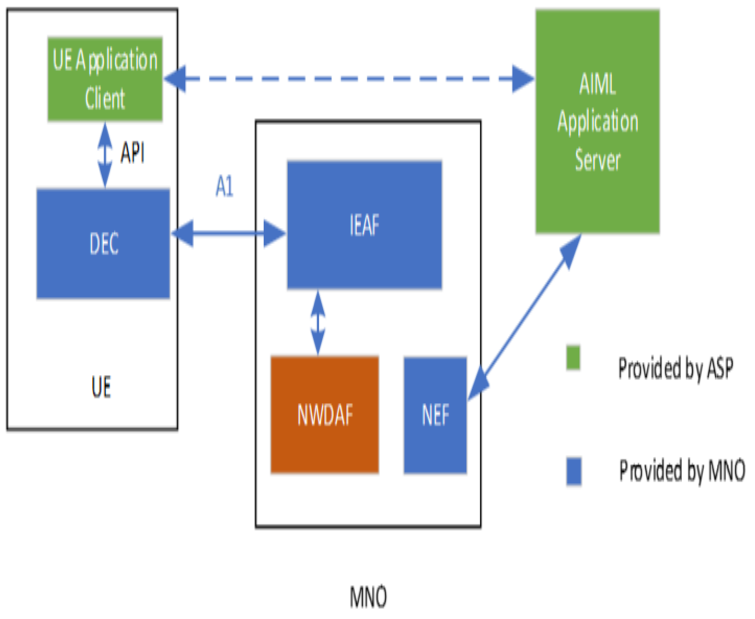


Figure: 5G IEAF Data Information Collection

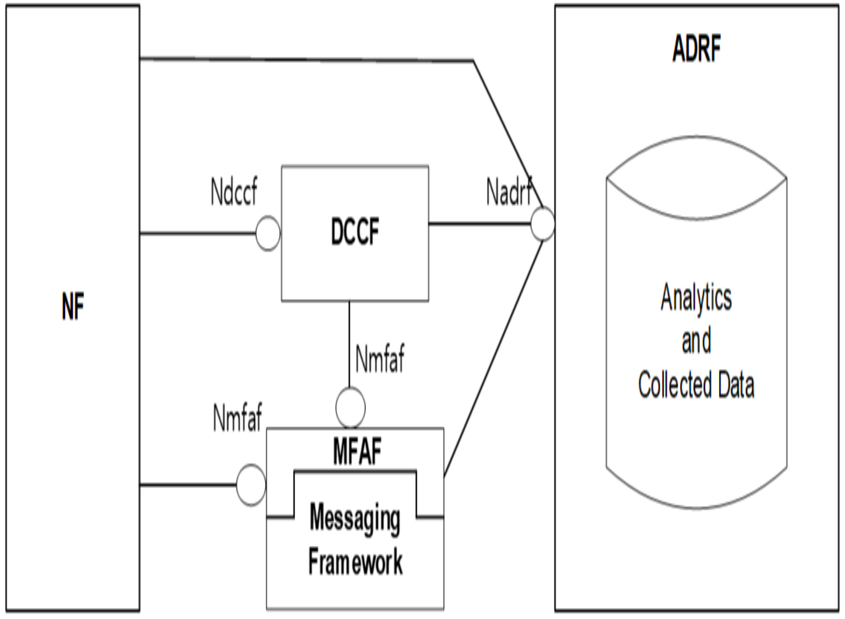
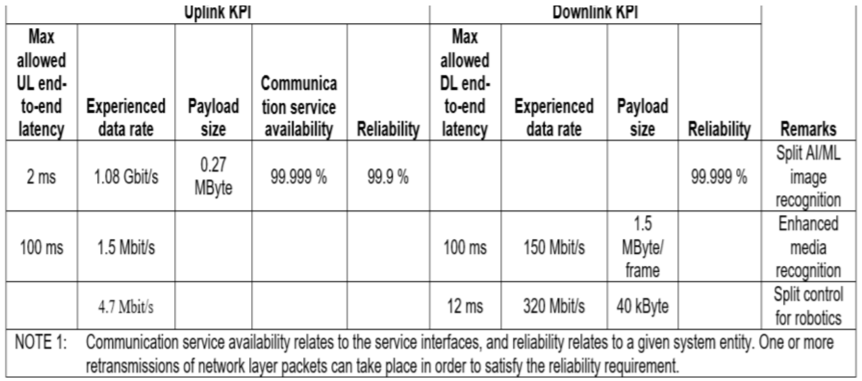


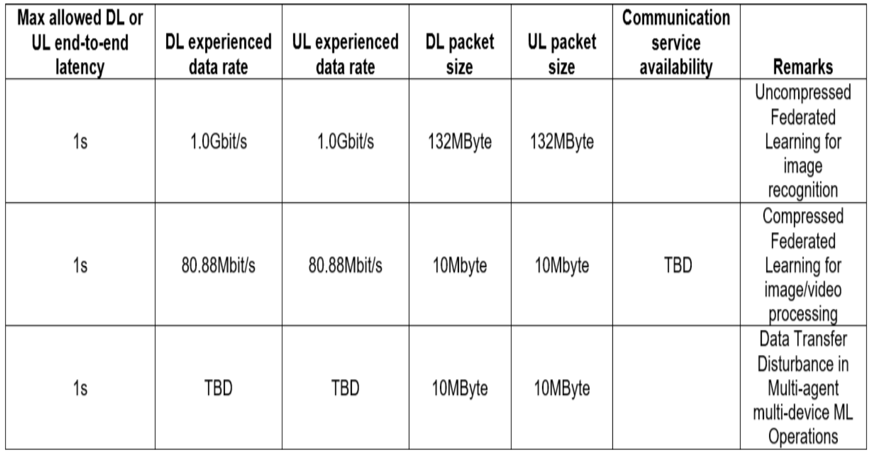
Figure: 5G Data Storage for Analytics and Collected Data

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	Communica tion 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

2. 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)

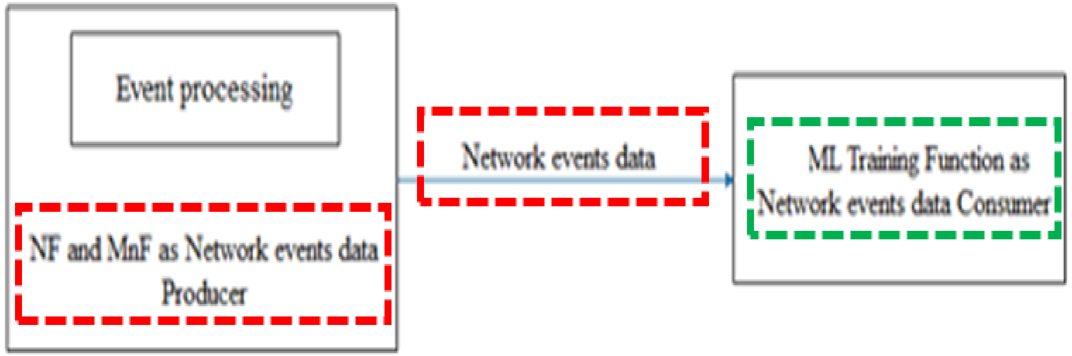

Event processing
NF and MnF as Network events data Producer
Network events data
ML Training Function as Network events data Consumer

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

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

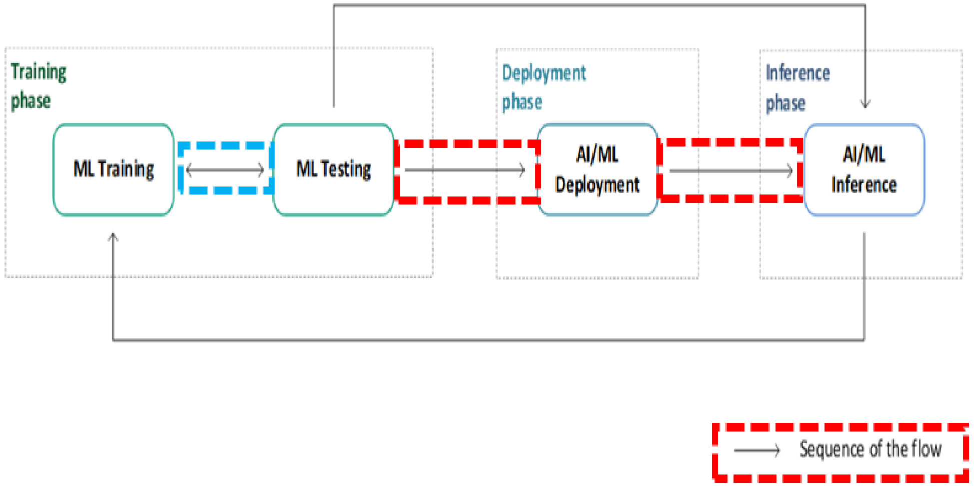


Figure: 5GS (including 5GC, NG-RAN and Management System), and the generic AI/ML Operational Workflow of the Operational Steps in the Lifecycle of an ML Model or Entity

2. 5G System use of AI/ML

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*.

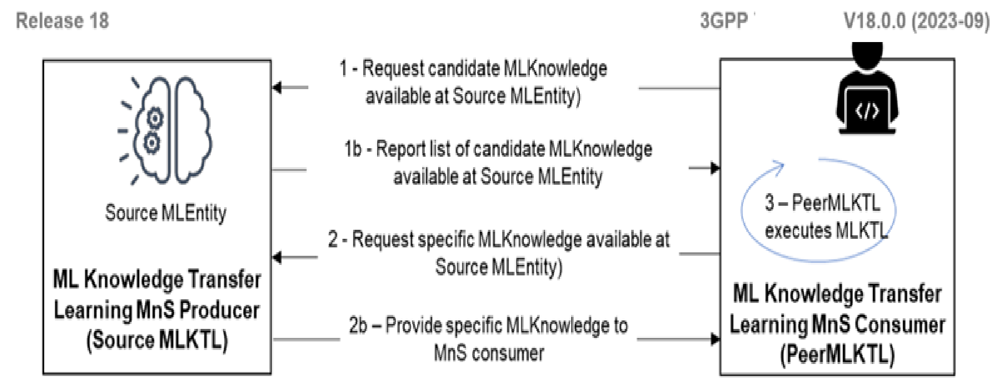

Release 18
3GPP
V18.0.0 (2023-09)
Source MLEntity
ML Knowledge Transfer Learning MnS Producer (Source MLKTL)
1 - Request candidate MLKnowledge available at Source MLEntity)
1b - Report list of candidate MLKnowledge available at Source MLEntity
2 - Request specific MLKnowledge available at Source MLEntity)
2b – Provide specific MLKnowledge to MnS consumer
3 – PeerMLKTL executes MLKTL
ML Knowledge Transfer Learning MnS Consumer (PeerMLKTL)

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

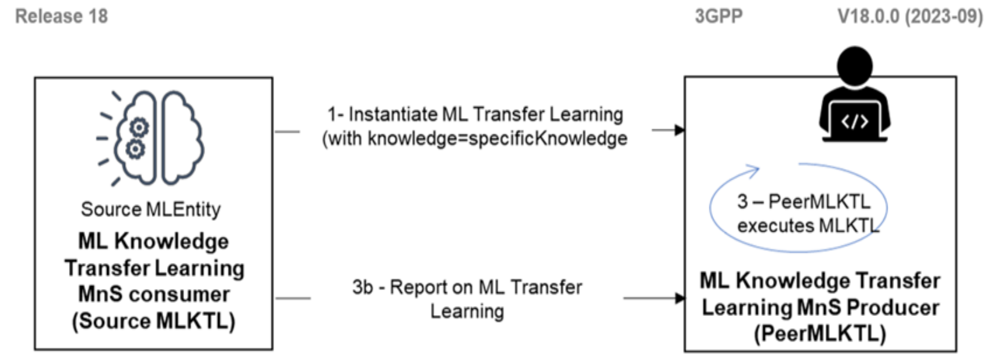

Release 18
3GPP
V18.0.0 (2023-09)
Source MLEntity
ML Knowledge Transfer Learning MnS consumer (Source MLKTL)
1- Instantiate ML Transfer Learning (with knowledge=specificKnowledge
3b - Report on ML Transfer Learning
3 – PeerMLKTL executes MLKTL
ML Knowledge Transfer Learning MnS Producer (PeerMLKTL)

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

2. 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*.

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

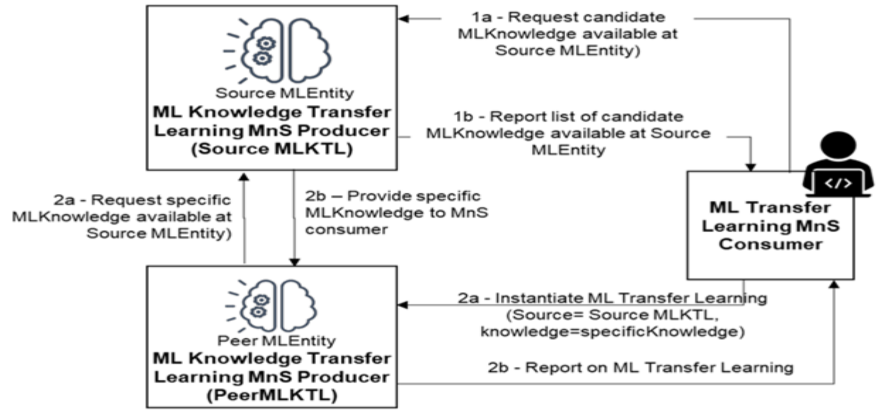

Source MLEntity
ML Knowledge Transfer Learning MnS Producer (Source MLKTL)
1a - Request candidate MLKnowledge available at Source MLEntity)
1b - Report list of candidate MLKnowledge available at Source MLEntity
ML Transfer Learning MnS Consumer
</>
2a - Request specific MLKnowledge available at Source MLEntity)
2b – Provide specific MLKnowledge to MnS consumer
Peer MLEntity
ML Knowledge Transfer Learning MnS Producer (PeerMLKTL)
2a - Instantiate ML Transfer Learning (Source= Source MLKTL, knowledge=specificKnowledge)
2b - Report on ML Transfer Learning

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

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

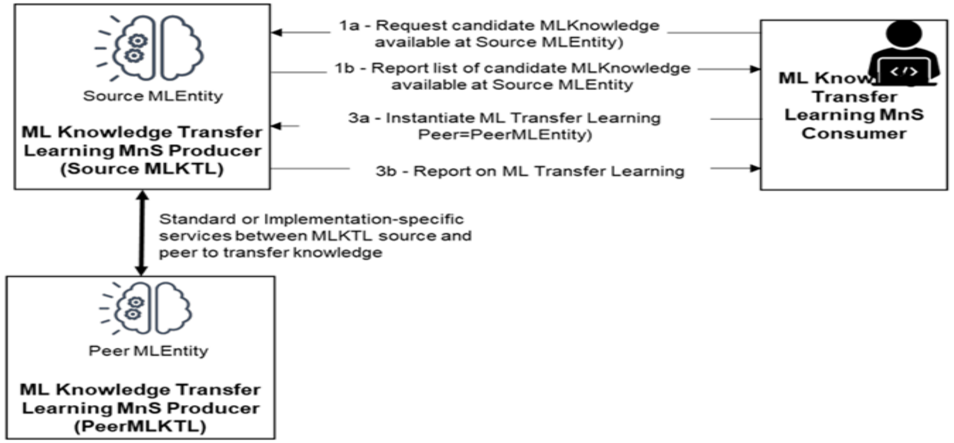

Source MLEntity
ML Knowledge Transfer Learning MnS Producer (Source MLKTL)
1a - Request candidate MLKnowledge available at Source MLEntity)
1b - Report list of candidate MLKnowledge available at Source MLEntity
3a - Instantiate ML Transfer Learning Peer=PeerMLEntity)
3b - Report on ML Transfer Learning
ML Know Transfer Learning MnS Consumer
</>
Standard or Implementation-specific services between MLKTL source and peer to transfer knowledge
Peer MLEntity
ML Knowledge Transfer Learning MnS Producer (PeerMLKTL)

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

5GS Management Service (*MnS*) use of 5GS Network Layer support for NF Service "Producer" - NF Service "Consumer" Interaction - 1

5GS Management Service (*MnS*) *"Producers", "Consumers" and "Exposure*"

The **Management Services** (*MnSs*) for a Mobile Network with or without Network Slicing may be produced by any Entity.

For example, it can be *Network Functions (NFs),* or Network Management Functions.

The ***Entity**** may provide ("*produce")* such Management Services as, for example, the
- Performance Management Services,
- Configuration Management Services and
- Fault Supervision Services

The **Management Services** (*MnSs*) can be "*consumed" by another Entity,* which may in turn "*produce" (expose) the Service to other Entities.*

The Figure shows an example of the *Management Service X*, which is initially "*produced" by the "Entity A*", which is an *NF*, then "*consumed" by another "Entity B"* which is a Network Management Function (*NMF*). Then "*Entity B"* in turn *exposes* it (the *same "Management Service X*" to the *"Entity C".*

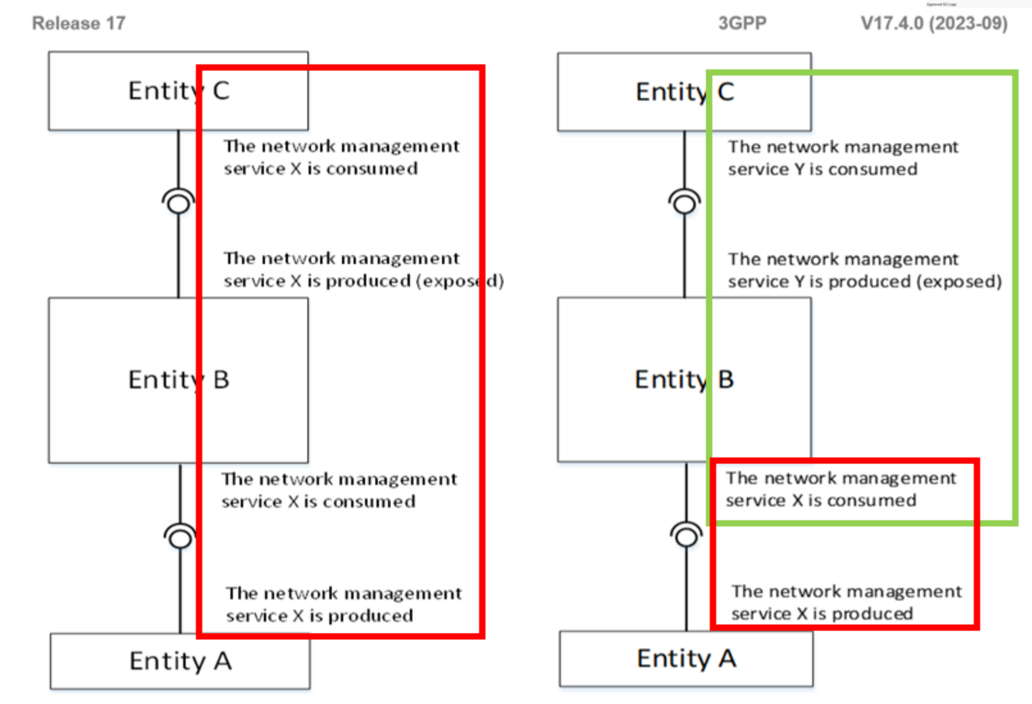

Release 17
3GPP
V17.4.0 (2023-09)
Entity C
The network management service X is consumed
The network management service X is produced (exposed)
Entity B
The network management service X is consumed
The network management service X is produced
Entity A
Entity C
The network management service Y is consumed
The network management service Y is produced (exposed)
Entity B
The network management service X is consumed
The network management service X is produced
Entity A

Figure: 5G System Management Architecture Framework Reference Model for Management Service (MnS) Producers, Consumers and Exposure

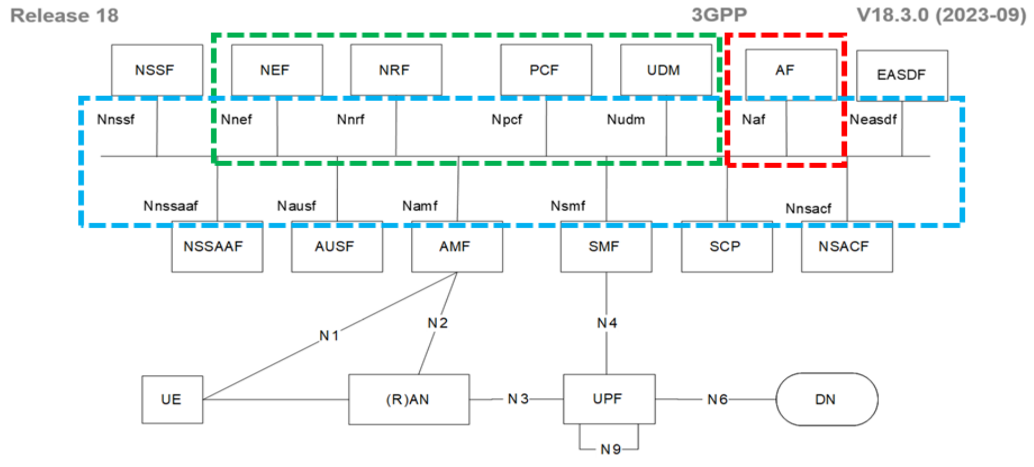


Figure: 5G System Non-Roaming Architecture

**Entity*"* *as being defined within the updated definition of "Context" used in 3GPP 5G System Architecture and ETSI*

5GS Network Layer support for NF Service "Producer" - NF Service "Consumer" Interaction - 2

In 5GS, a NF Service is one (1) *"Type of Capability"* exposed by an ***NF*** (*NF Service "Producer")* to other authorized **NF** (*NF Service "Consumer"*) through a *Service-based Interface (SBI*).

A Network Function (NF) may expose one (1) or more NF Services. The following Criteria specify NF Services:

- *NF Services* are derived from the *System Procedures* that describe *End-to-End (E2E) Functionality*, where applicable (see 5GS Architecture Procedure specification, Annex B drafting rules).

Services may also be defined based on *information flows from other 3GPP specifications.*

- 5G System Procedures can be described by a sequence of NF Service Invocations.

- NF Services may communicate "directly" between NF Service "Consumers" and NF Service "Producers", or "indirectly" via an SCP.

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

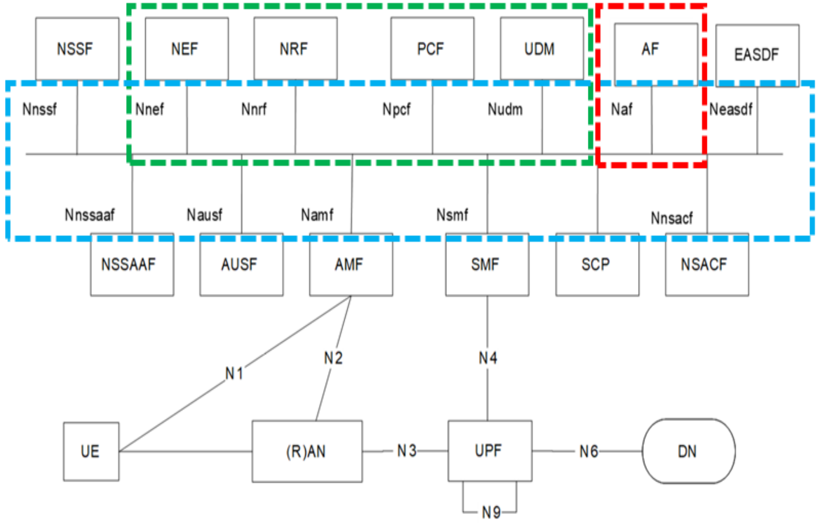


Figure: 5G System Non-Roaming Architecture

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

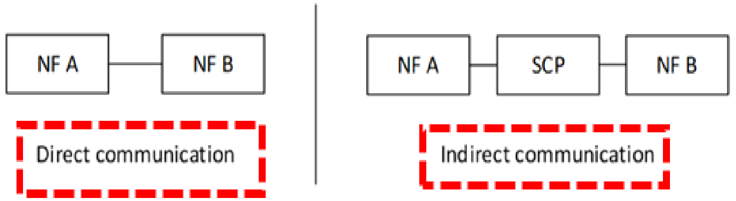


Figure: 5G System Architecture NF-to-NF Service Inter Communication

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

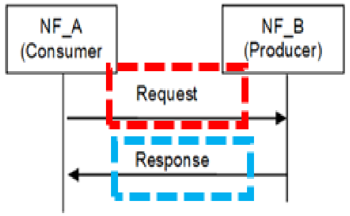


Figure: 5G System Architecture "Request-Response" NF Service Communication

5GS Network Layer support for: NF Service "Producer" - NF Service "Consumer" Interaction - 3

The *E2E interaction* between (2) Network Functions, *"Consumer" and "Producer*", within this *NF Service Framework* follows *two (2) Mechanisms*, irrespective of whether *"Direct Communication" or "Indirect Communication"* is used:

- *"Request-Response":* **A Control Plane (CP) NF_B** (*NF Service "Producer*") is requested by another **Control Plane (CP) NF_A** (*NF Service "Consumer*") to provide a certain *NF Service*, which either A) Performs an Action or B) Provides Information or C) Both.
NF_B provides an NF Service based on the request by NF_A.
In order to fulfil the request, NF_B may in turn "consume" NF Services from other NFs.

In "*Request-Response" Mechanism*, Communication is one to one between two NFs (Consumer and "Producer") and a one-time response from the producer to a request from the "Consumer" is expected within a certain timeframe.

The *NF Service "Producer*" may also *add a Binding Indication* in the Response, which may be used by the *NF Service "Consumer*" to select suitable *NF Service "Producer" instance(s*) for subsequent Requests.

For *indirect communication*, the *NF Service "Consumer*" copies the *Binding Indication* into the *Routing Binding indication*, that is included in subsequent requests, to be used by the *SCP* to discover a *suitable NF Service "Producer" instance(s).*

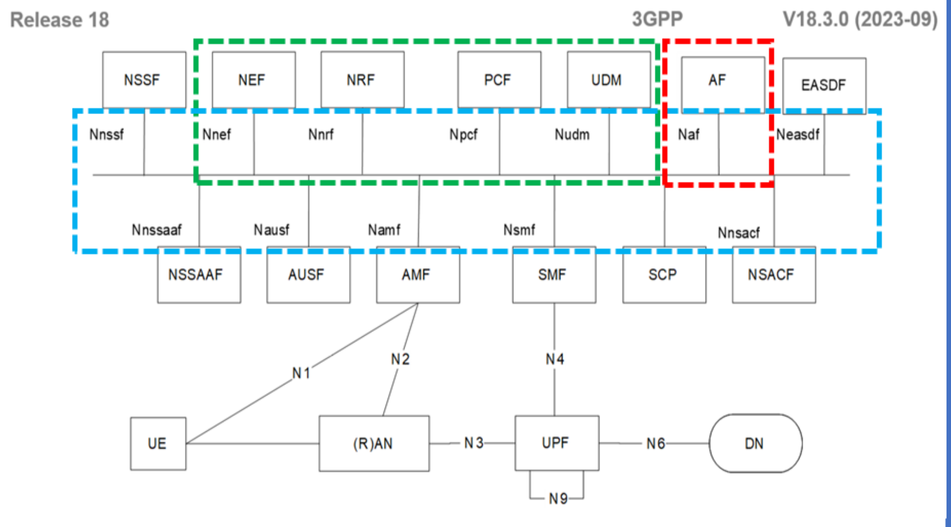


Figure: 5G System Non-Roaming Architecture

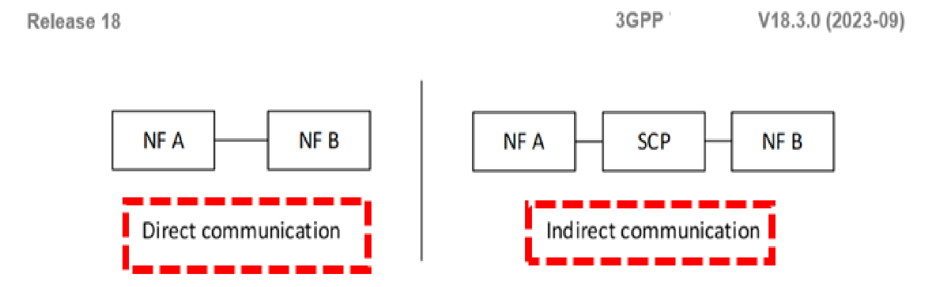


Figure: 5G System Architecture NF-to-NF Service Inter Communication

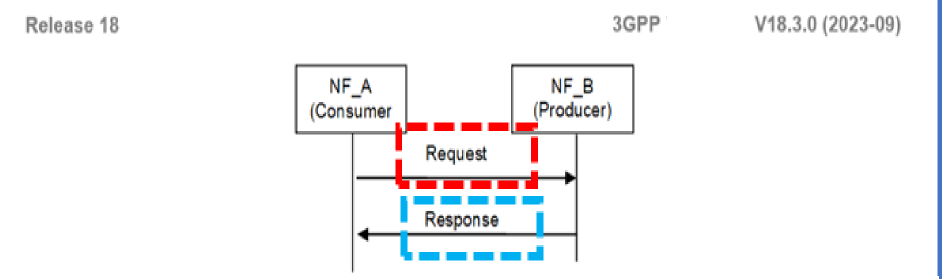


Figure: 5G System Architecture "Request-Response" NF Service Communication

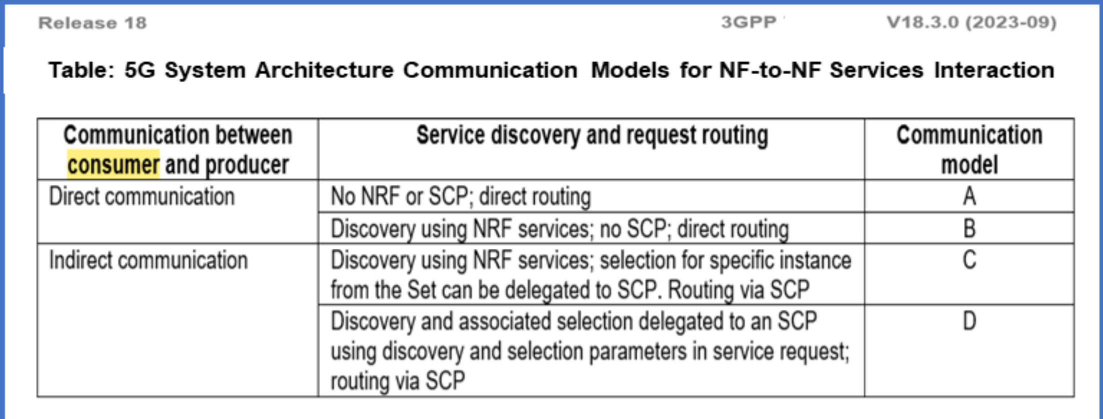
5GS Network Layer support for NF Service "Producer" - NF Service "Consumer" Interaction - 4

Model A - *Direct Communication without NRF interaction*: Neither NRF nor SCP are used. "Consumers" are configured with "Producers' "NF Profiles" and directly communicate with a "Producer" of their choice.

Model B - *Direct Communication with NRF interaction*: "Consumers" do discovery by querying the NRF. Based on the discovery result, the *"Consumer"* does the selection. The *"Consumer"* sends the request to the selected *"Producer".*

Model C - *Indirect Communication* without delegated discovery: *"Consumers"* do discovery by querying the NRF. Based on discovery result, the *"Consumer"* does the selection of an *NF Set* or a specific *NF instance of NF set*. The *"Consumer"* sends the request to the SCP containing the address of the selected *Service "Producer"* pointing to a NF Service Instance or a set of NF service instances. In the latter case, the SCP selects an *NF Service instance*. If possible, the SCP interacts with NRF to get selection parameters such as Location, Capacity, etc. The SCP routes the request to the selected *NF Service "Producer" instance*.

Model D - *Indirect Communication* with delegated discovery: "*Consumers"* do not do any discovery or selection. The *"Consumer*" adds any necessary discovery and selection parameters required to find a suitable "Producer" to the Service Request. The SCP uses the request address and the discovery and selection parameters in the request message to route the request to a suitable *"Producer" Instance*. The SCP can perform discovery with an NRF and obtain a discovery result.



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

Table: 5G System Architecture Communication Models for NF-to-NF Services Interaction

Communication between consumer and producer	Service discovery and request routing	Communication model
Direct communication	No NRF or SCP; direct routing	A
	Discovery using NRF services; no SCP; direct routing	B
Indirect communication	Discovery using NRF services; selection for specific instance from the Set can be delegated to SCP. Routing via SCP	C
	Discovery and associated selection delegated to an SCP using discovery and selection parameters in service request; routing via SCP	D

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

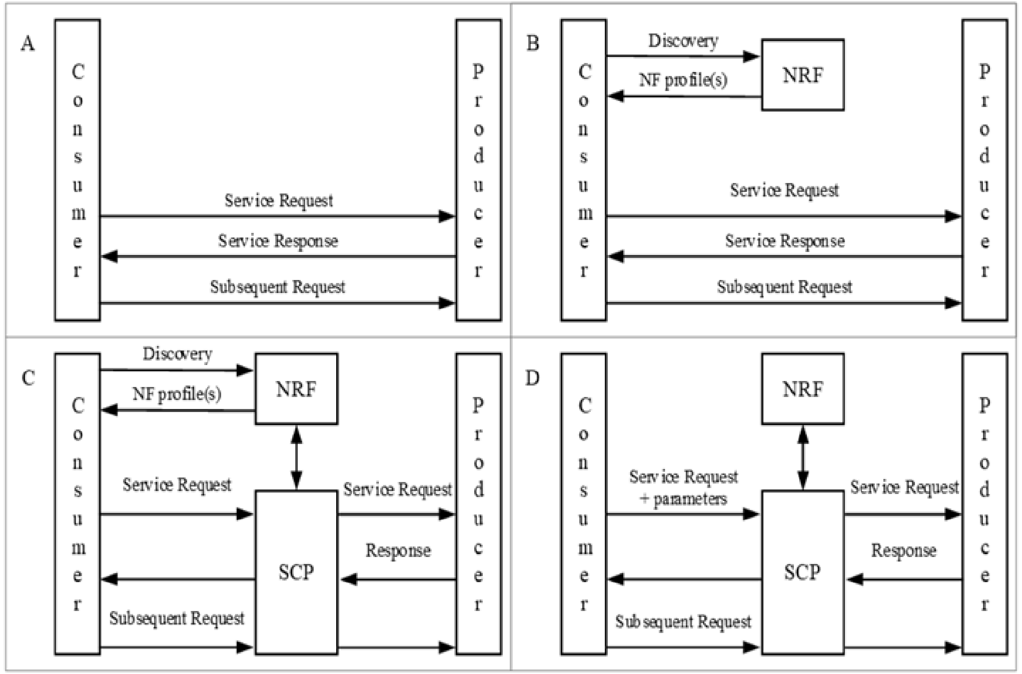


Figure: 5G System Architecture Communication Models for NF-to-NF Services Interaction

Annex 1: Mobile Networks to evolve from:

a Design that offers "Best-effort Services

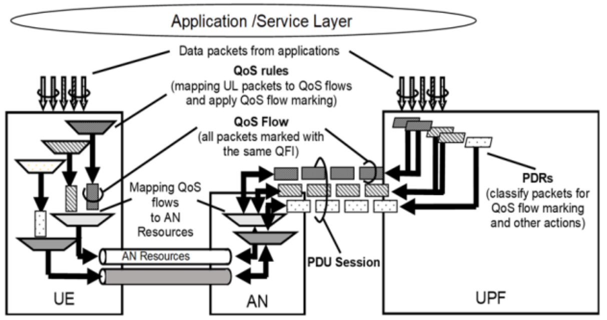
to

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.

Release 18 3GPP V18.2.0 (2023-06)



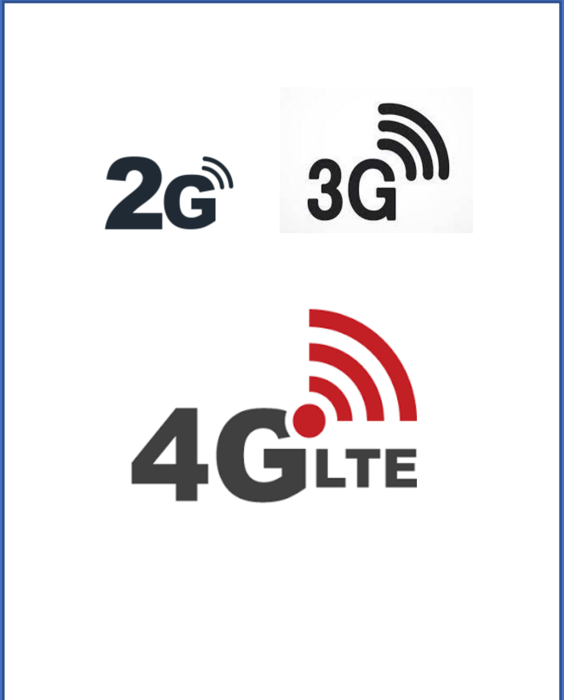
Application /Service Layer
Data packets from applications
QoS rules
(mapping UL packets to QoS flows and apply QoS flow marking)
QoS Flow
(all packets marked with the same QFI)
Mapping QoS flows to AN Resources
AN Resources
PDU Session
PDRs
(classify packets for QoS flow marking and other actions)
UE
AN
UPF

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

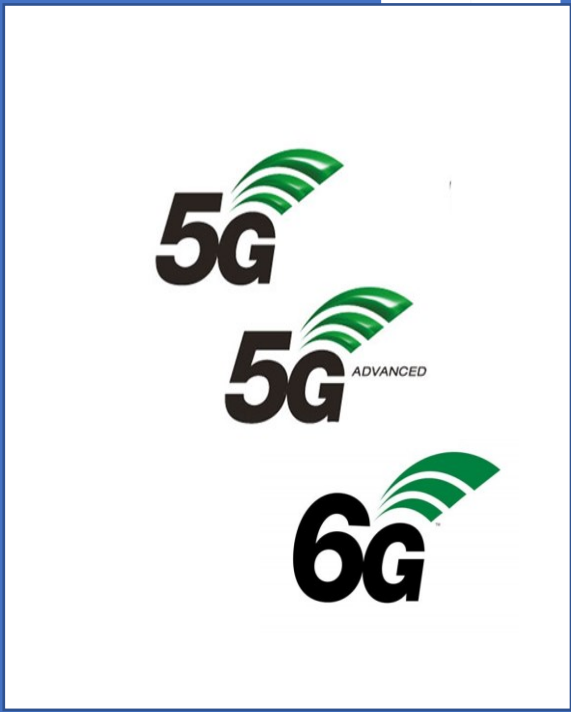
Measurement Assistance Information shall include the addressing information of *a* ***Performance Measurement Function*** **(PMF) in the UPF, the UE can send PMF protocol messages** incl.:

- 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**.



=>



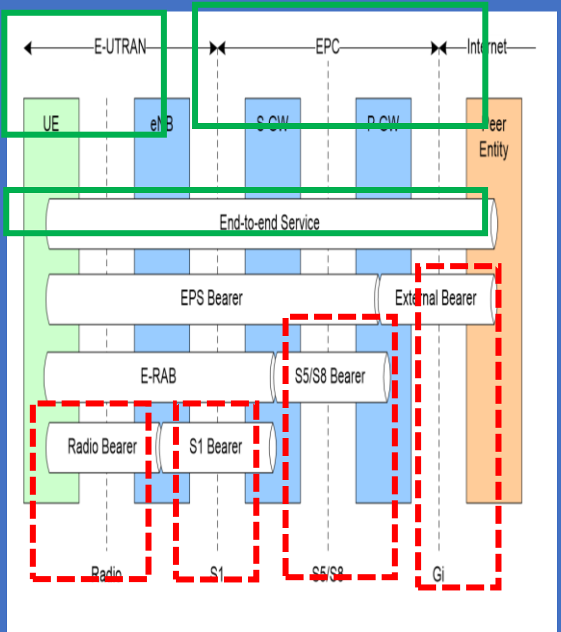


Figure : EPS Bearer Service Architecture

=>

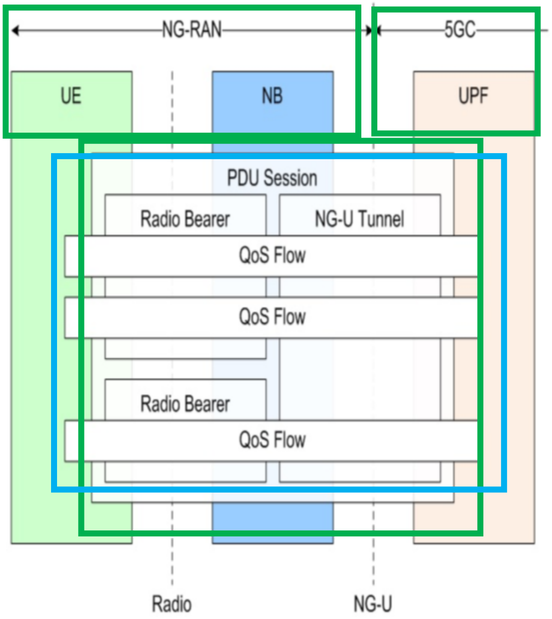


Figure: 5G CN NG-RAN Bearer Services QoS Architecture

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).

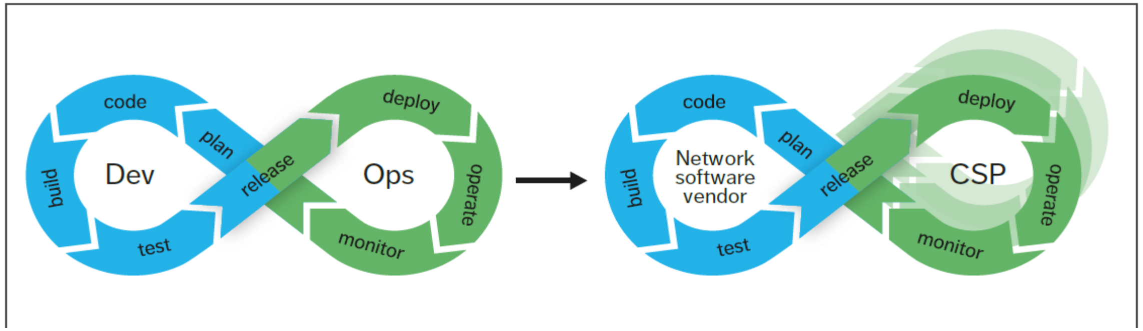


Figure 1: The DevOps and (Telecom) aaP Business Models

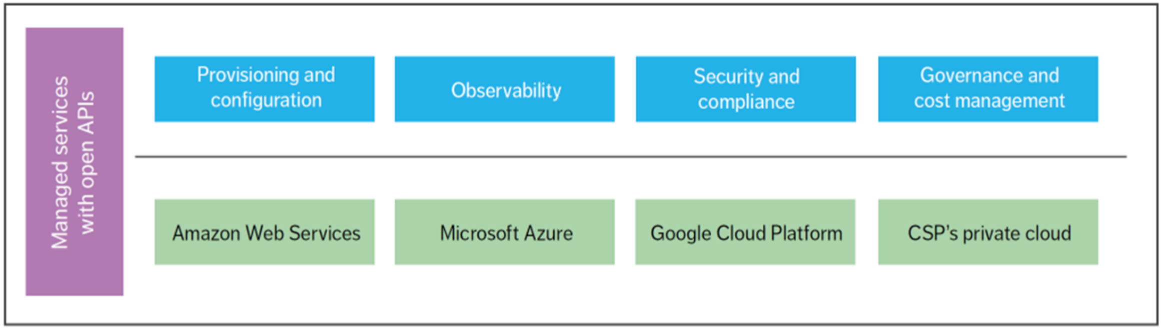
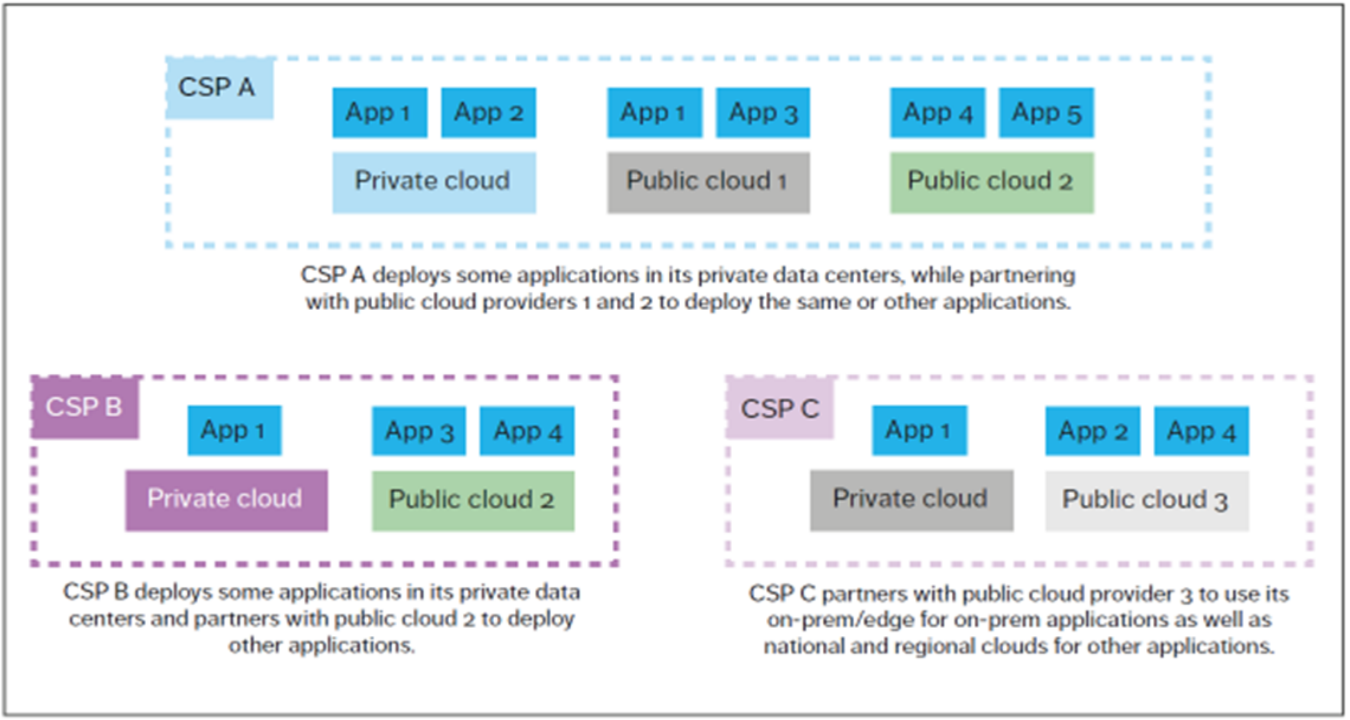


Figure 3: Key Enablers for a Multi-Cloud Native Application



CSP A deploys some applications in its private data centers, while partnering with public cloud providers 1 and 2 to deploy the same or other applications.

CSP B deploys some applications in its private data centers and partners with public cloud 2 to deploy other applications.

CSP C partners with public cloud provider 3 to use its on-prem/edge for on-prem applications as well as national and regional clouds for other applications.

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

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.

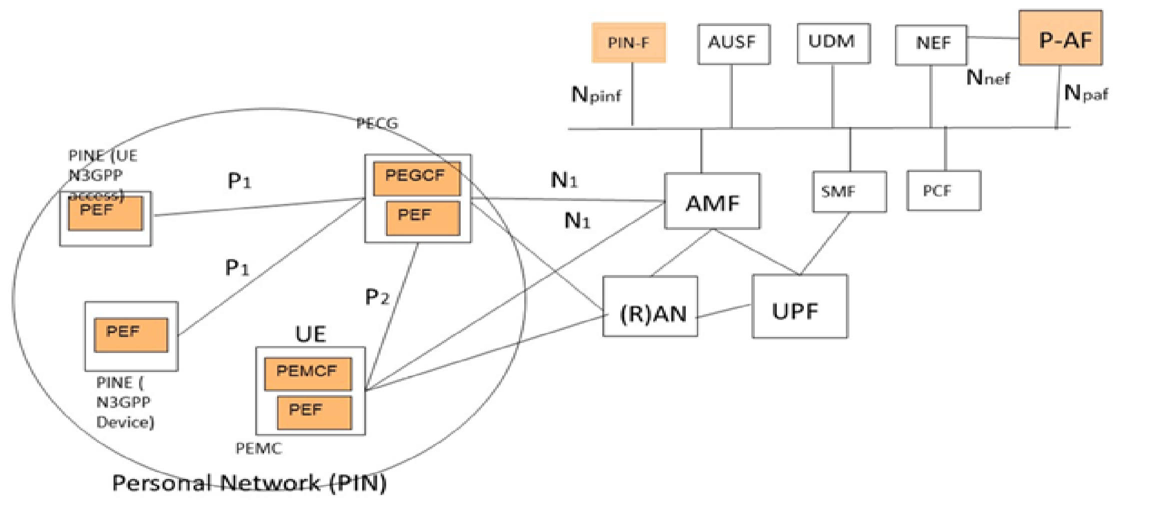
PINE-to-Network routing: the traffic is routed by a PEGC between PINE and 5GS, the PINE direct connects with the PEGC via non-3GPP access separately.

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

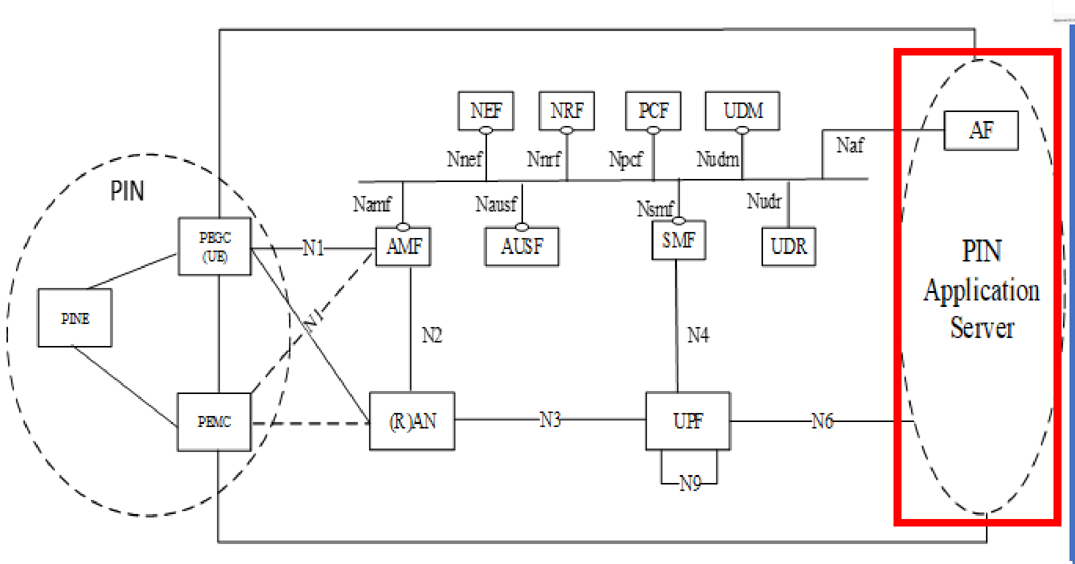


Figure: 5GS PIN Personal IoT Network Reference Architecture

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.

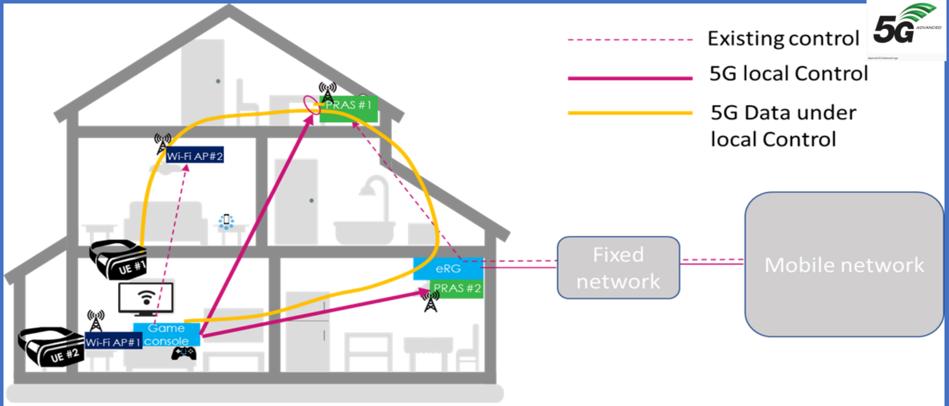


Figure: 5G Local Control of Premise Radio Access Stations (PRASs) for UE to access CPN Device

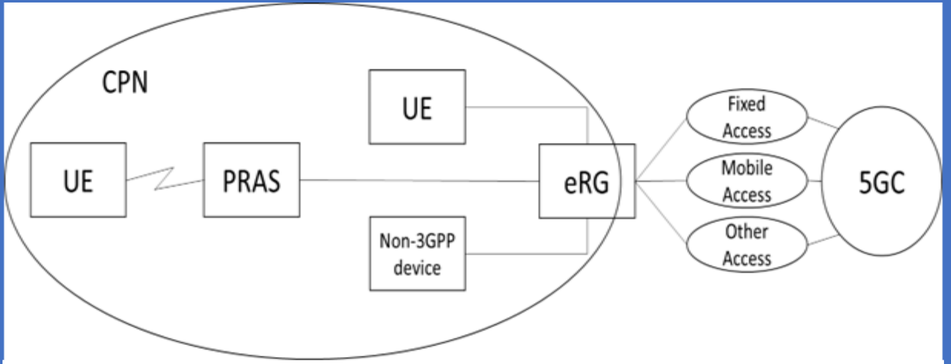
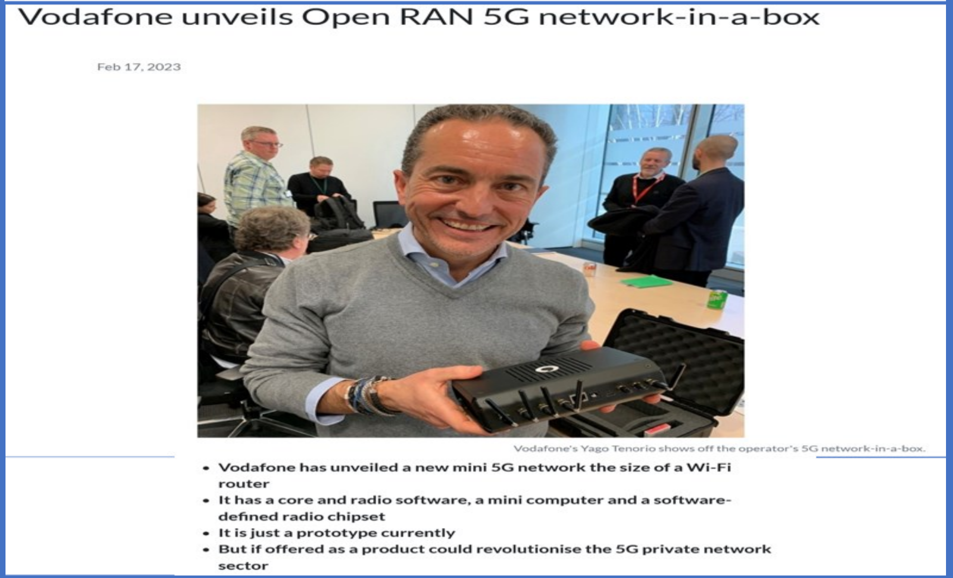


Figure: Customer Premises Network (CPN) connected to 5GC



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

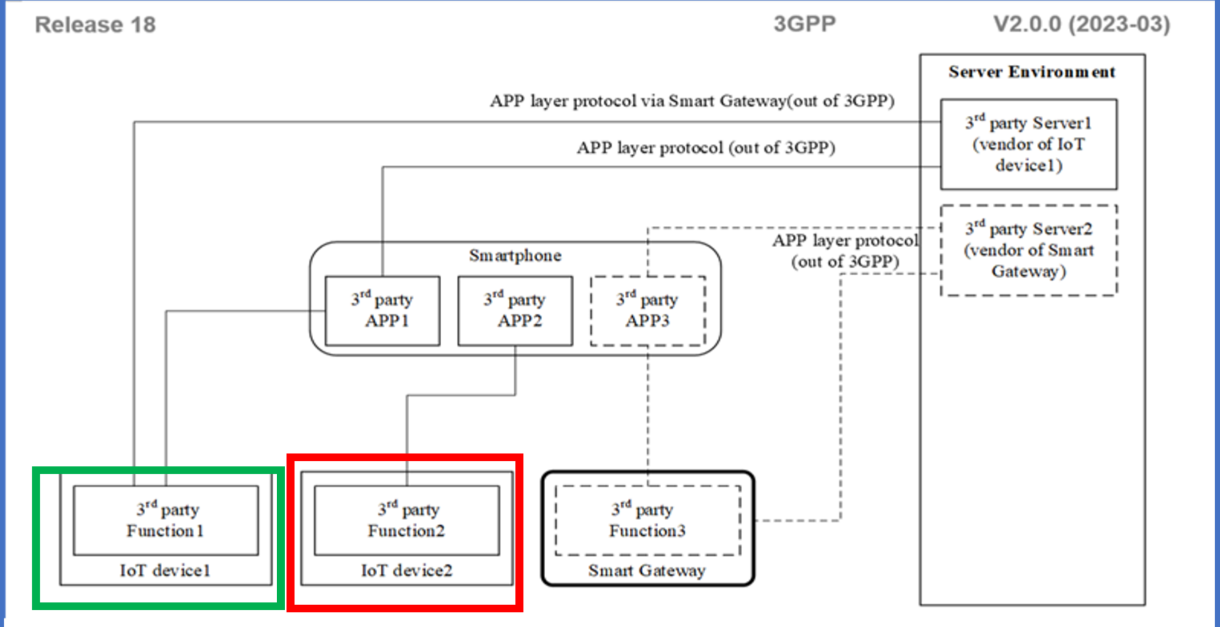
Feb 17, 2023

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

Annex 3 - 3GPP 5G Advanced Release specification for PINs (Personal IoT Network(s)) - 4

A Current Smart Home IoT Deployment Example



Release 18
3GPP
V2.0.0 (2023-03)
Server Environment
APP layer protocol via Smart Gateway(out of 3GPP)
3rd party Server1 (vendor of IoT device1)
APP layer protocol (out of 3GPP)
3rd party Server2 (vendor of Smart Gateway)
APP layer protocol (out of 3GPP)
Smartphone
3rd party APP1
3rd party APP2
3rd party APP3
3rd party Function1
IoT device1
3rd party Function2
IoT device2
3rd party Function3
Smart Gateway

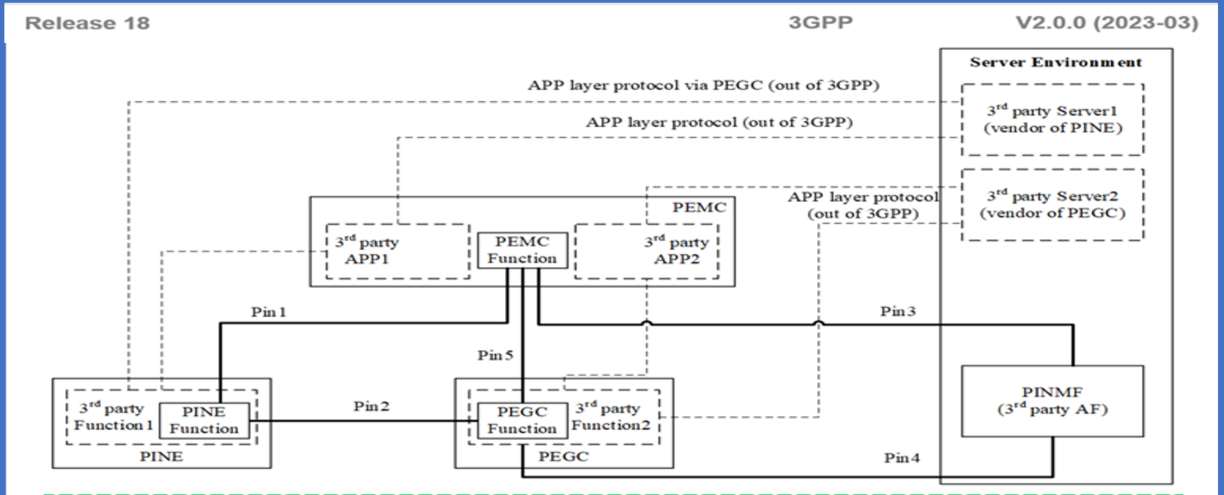
Figure: Example of Current IoT Smart Home Deployment

The IoT Device1 is initially discovered by a Smartphone using the 3rd Party APP1 installed in the Smartphone, and then the Smartphone is able to connect with the IoT Device1 assisted by the 3rd Party APP1.

The 3rd Party APP1 is developed by the Vendor of the IoT Device1. The IoT Device1 is able to visit the 3rd Party Server1 over Internet via the Smart Gateway, and the 3rd Party APP1 also can visit the 3rd Party Server1 over Internet, so that the Smartphone is able to control the IoT Device1 via internet assisted by the 3rd Party Server1.

The IoT Device2 is manufactured by a different Vendor from that of the IoT Device1, and is not able to be controlled by a Smartphone via Internet.

A Deployment Example of the PIN that the PINMF can be a NF, Trusted AF, or 3rd Party AF.



NOTE 1: The PEMC Function instance can be a function in a 3rd party APP, a standalone APP, or middleware,

Figure: 5G PIN Deployment Functions Example

For the case of NF/trust AF, one Operator only has one (1) PINMF, the PEMC can use pre-configured information for PIN Service Operations, e.g. FQDN of the Operator's PINMF.

For the Case of 3rd Party AF, there may be multiple PINMFs, which one is used determines by the User, and the Serving PINMF should register itself for the User to handle the PIN Service Operations.

If both PINMF as NF/trust AF and PINMF as 3rd party AF are deployed, which one is used is determined by PEMC implementation.
In the deployment example, the 3rd party APP and 3rd party Function can assist the initial discovery and initial direct connection setup between PINE/PEGC and PEMC without user input information.

An example of the use case with the deployment example is as following:

Remarks & Questions?