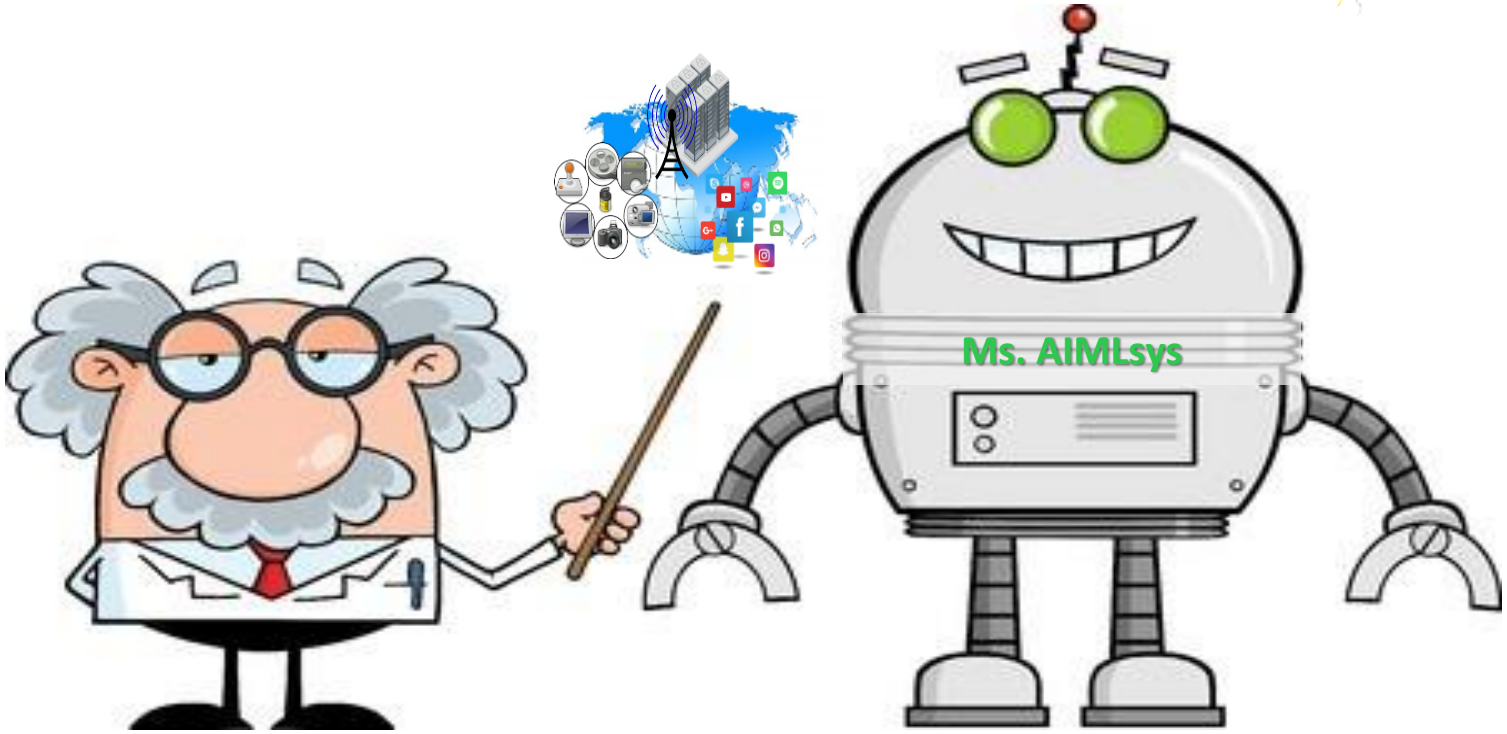
3GPP TSG-SA2 Meeting #157
Berlin, Germany May 22 – 26 2023

S2-2306424

Rel-19 - 5G System AI/ML Study

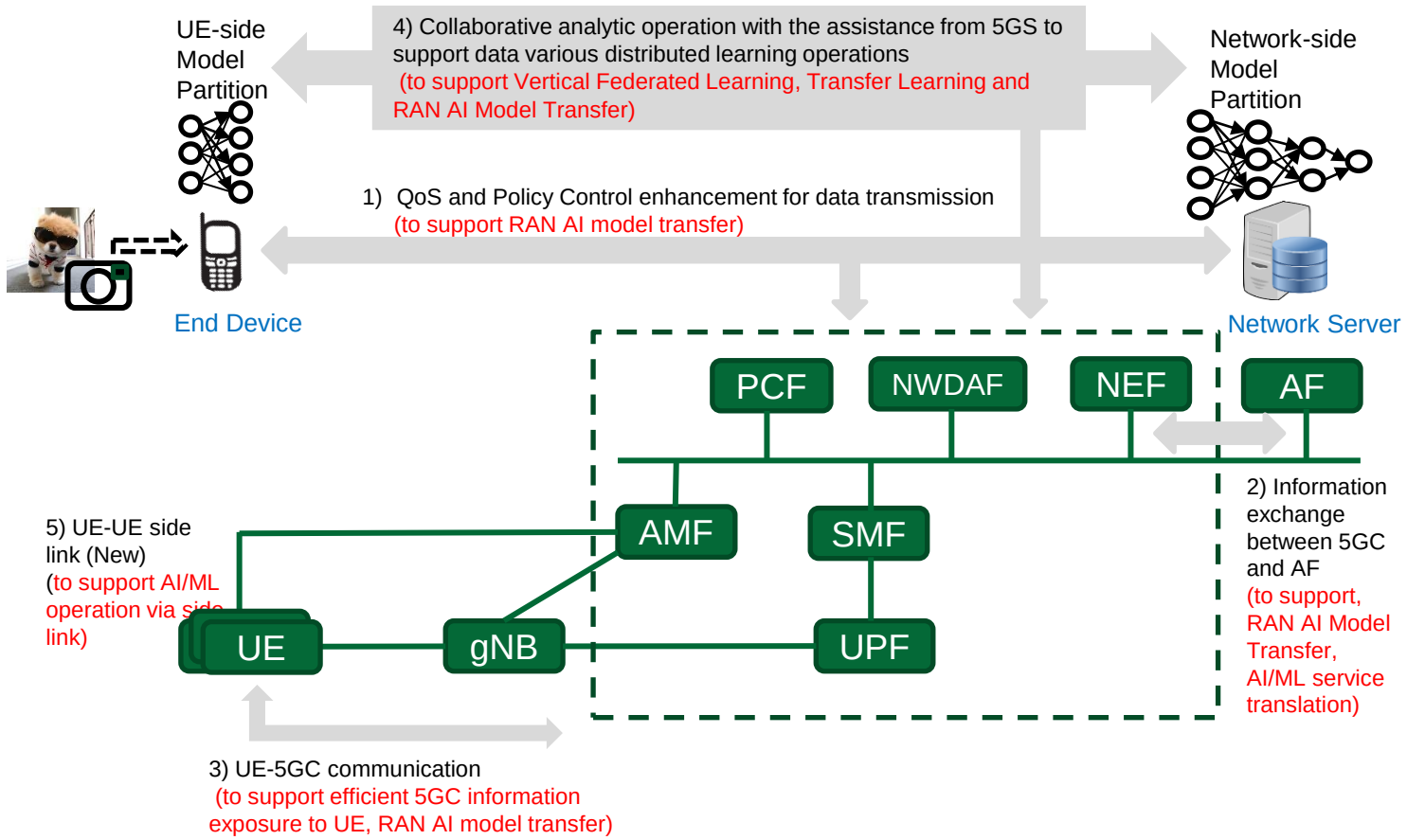
To Enable AI/ML Support Across Multiple Domains for Be ond 5G



AI/ML Advances the 5G Advanced System to be Even More Advanced!

Overall Considerations for Rel-19 5G System AI/ML Study –
Consistent Network Solutions to support AI/ML Across Multiple Domains

Considering the established solution framework for the AI/ML support in 5G Core and for service applications in Rel-18, it is beneficial to have a common 5GC Stage-2 study in Rel-19 to ensure consistent 5G network solutions to support end-to-end AI/Ml operations across multiple domains.



R19 Considerations to extend the AI/ML Functionalities Across Multiple 5GS domains

1. QoS and Policy Control enhancement for data transmission

   Support the RAN AI model transfer over air interface with the consideration of the R18 AIMLsys functionalities (e.g., the QoS and policy applied for AIML model transfer)

2. Information exchange between 5GC and AF

   Support the Vertical-FL, RAN AI model transfer and AIML service translation with the consideration of the R18 AIMLsys functionalities (e.g., external parameter provisioning, member selection) and solution in AIMLsys TR (e.g., NEF/AF based translation solution)

3. UE – 5GC communications

   Support the RAN AI Model Transfer (e.g. AI/ML Positioning) as well as the R18 leftover issues (5GC information exposure to UE) by leveraging the previous proposals (e.g. DCAF/IEAF or NAS based solution) captured in R18 AIMLsys TR

4. Collaborative analytic operation with the assistance from 5GS to support various distributed learning operations

   Support the Vertical FL, RAN AI model transfer and Transfer learning with the consideration of the R18 AIMLsys functionalities (e.g., Member UE selection mechanism, Consolidated QoS support)

5. UE-UE side link (New)

   Support the AI/ML operation via side link according to the requirements in SA1 AIML-Ph2 study

Outline – Multi-domain AI/ML Support Considerations

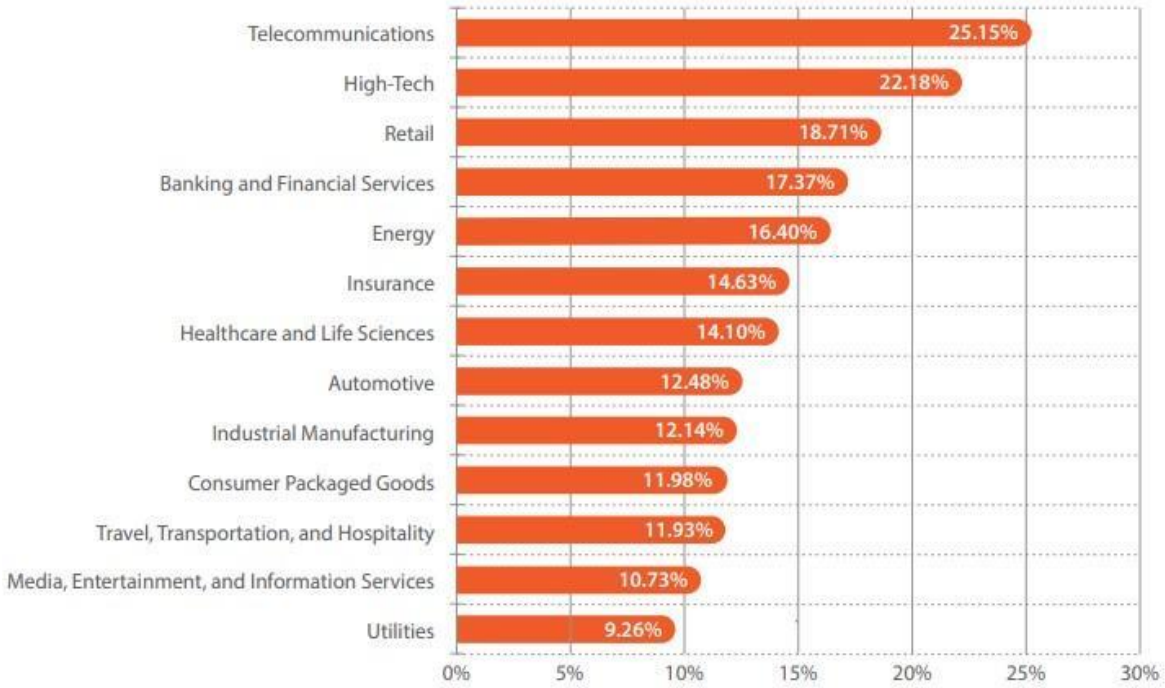
- Motivations of AI/ML Support Across Multiple Domains for 5G Advanced System
- Proposals for Multi-domain AI/ML (MDAIML) support at a glance
- Proposed Contents for MDAIML
  1. 5GS assisted Cross-domain AI training and inference (Vertical FL)
  2. 5GC Assistance for RAN AI/ML Model Transfer/Delivery
  3. 5GC Assistance for RAN AI-based Positioning
  4. 5GS Assistance for Transfer Learning
  5. Leveraging Sidelink for AIML service for Model Splitting
  6. Leveraging Sidelink for AIML service for Distributed learning and inference
  7. Efficient 5GC Information Exposure to UE for AI/ML operation
  8. AIML-Translation between application request and 5GC APIs

Note: "Multiple Domains" include UE, RAN, OAM, Core Network and Applications

- Conclusions

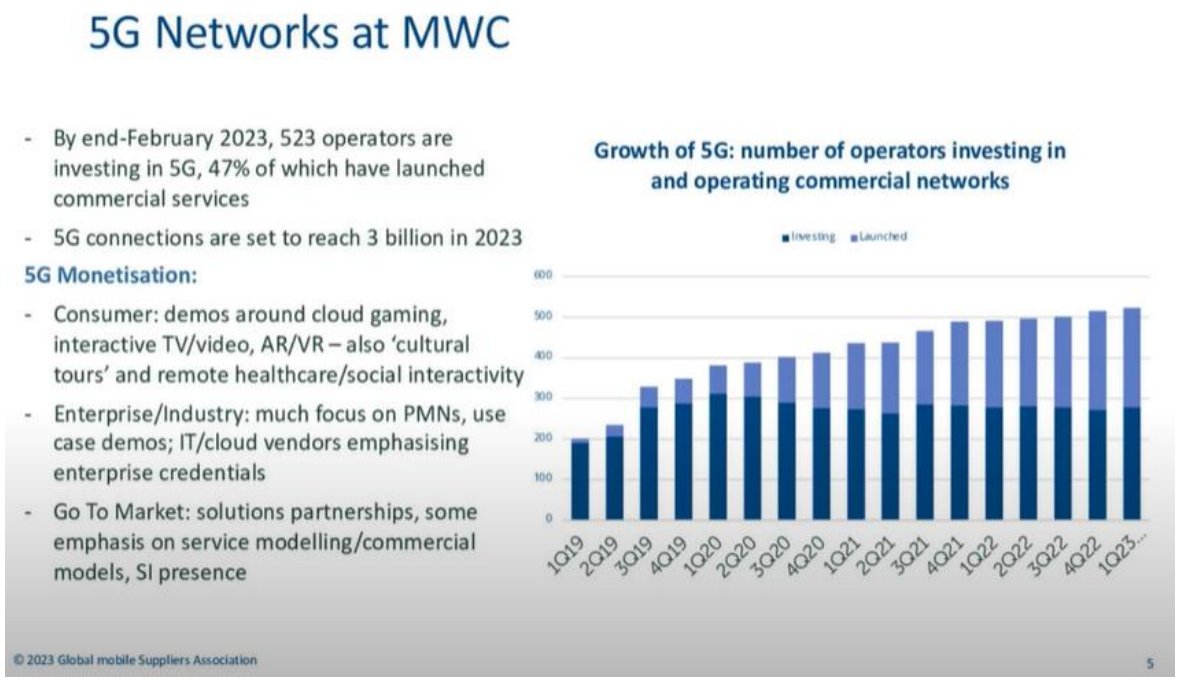
Motivations of End-to-end AI/ML Support for 5G Advanced System – Technology & Market are READY

- Proven trend of Artificial Intelligence (AI) proliferates the growth of the global industries business. In the meantime, the increasing advanced 5G system with high performance and capacity smartphones present the golden opportunity for MNO to leverage its 5G assets to offer new breed of mobile services for its Vertical customers to assist their AI/ML operations.



The biggest revenue gains due to AI implementation

Credit to: Ramaswamy, S. "Getting Smarter by the Sector: How 13 Global Industries Use Artificial Intelligence," TCS Global Trend Study



As of March 2023, over 60% USA smartphones are 5G devices.

Global Industries are READY for a bigger Boost from More Advanced 5G System

Motivations of Multi-domain AI/ML Support for 5G Advanced System

— Valuable Data Sets from Different Domains to Boost Cross-domain Training & Inference Precision

Value Propositions

Service QoE Assurance

Adaptive to customized Vertical service requirements with more accurate network capabilities/resources projections to ensure service delivery and performance

Mobile System Enhancement

More accurate and rapid engineer to adapt to changing network conditions and to improve network resources and energy performance efficiency

Streamlined Service Enablement

Ability to expose more precise 5G system information to authorized 3rd party and UEs to enable business opportunities for valued added services to enrich MNOs' service portfolio offerings.

Enabler

System Performance Paybacks

- QoE/SLA Prediction
- QoE/SLA Assurance
- Energy and Power Saving
- Load Balancing
- Multi-connectivity/Carrier Aggregation
- Mobility Optimization
- Mobility Pattern Prediction
- Interference Mitigation/Management
- Ultra Reliability & Low Latency
- Coverage & Capacity Extensions
- Service Discovery & Positioning
- User Distribution Prediction
- Service Translation/Profiling
- Radio & Network Performance Prediction
- Radio & Network Performance Assurance

Enabler

Multi-domain AI/ML Support

Converged 5G AI/ML Architecture

UE Domain | RAN Domain | Core Domain | OAM | Application Domain

Data Collection:-

- ✓ RAN {L1/L2 Measurement, MDT, Signaling}, Core {performance, mobility, QoS etc.}, UE {performance, comp resource etc.},

Data Analytics:

- ✓ Radio Finger-print and Traffic characteristic recognition (type, rate, delay, etc.)

AI/ML:

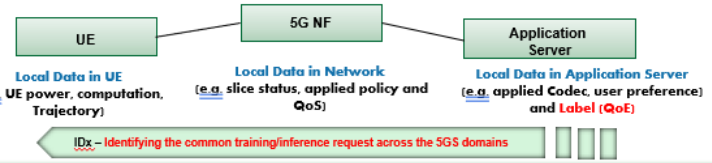
- ✓ Horizontal/Vertical FL with Model/Data distribution, Transfer/splitted Learning, Centralized/Distributed Inference, Transfer Learning,

Cross Domains Coordination:

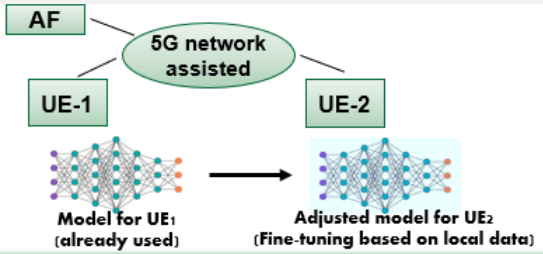
- ✓ Feature selection, Label determination, Member UE selection, Intelligent Data Transport Scheduling, Policy/QoS management, Model Identification/registration, Cross domain AI/ML session coordination, Performance monitoring

Proposals for MDAIML At A Glance

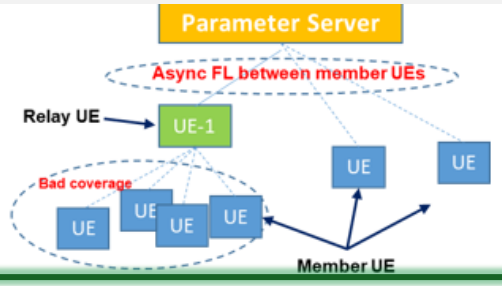
5GS assistance Cross-domain AI Training & Inference



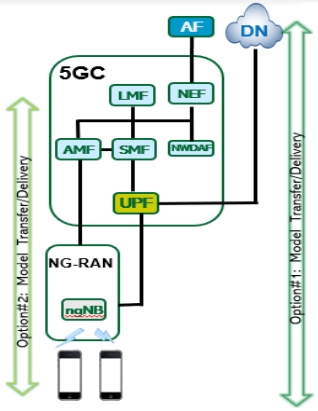
5GS Assistance for Model Distribution to support Transfer Learning



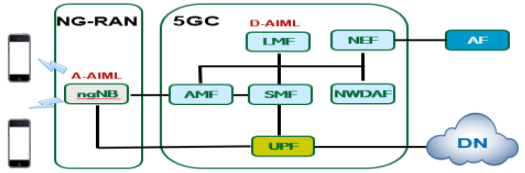
Leveraging Sidelink for AIML service for Horizontal Federated Learning



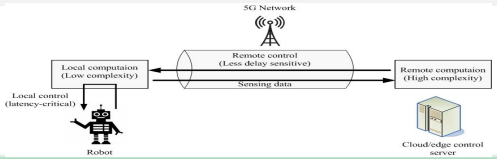
5GC Assistance for RAN AI/ML Model Transfer/Delivery



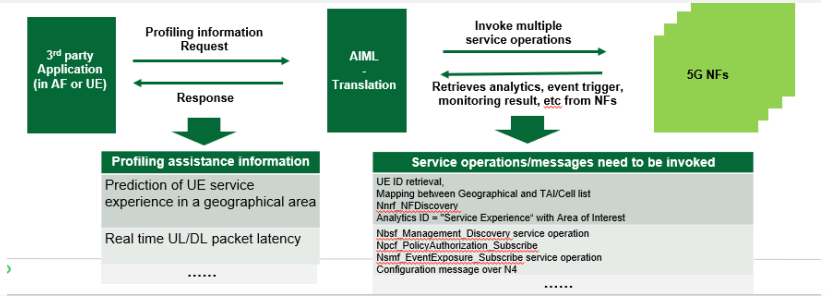
5GC Assistance for RAN AI-based Positioning



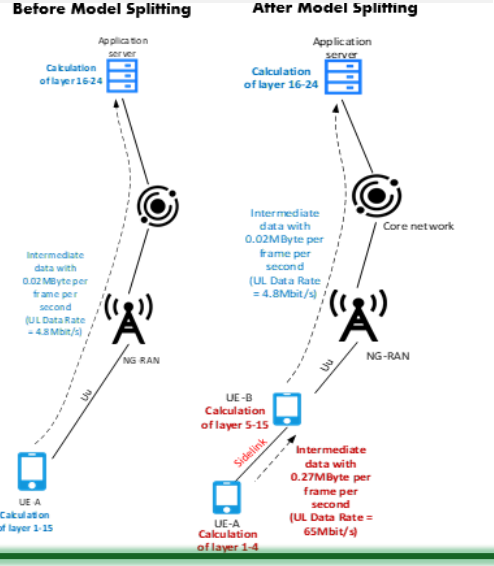
Efficient 5GC Information Exposure to UE for AI/ML operation



AIML-Translation between application request and 5GC APIs



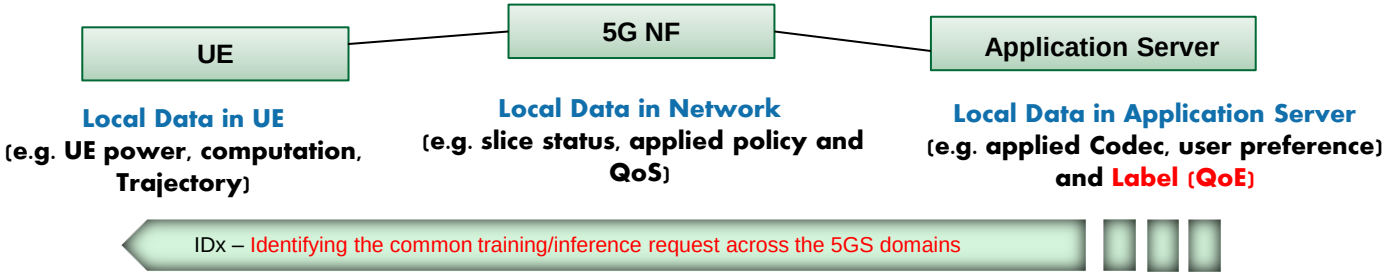
Leveraging Sidelink for AIML service for Model Splitting

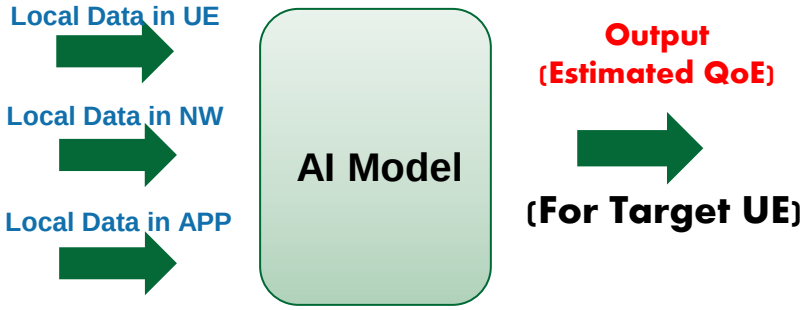


1. 5GS assisted Cross-domain AI training and inference (Vertical FL)

Background:

- Application layer User experience (QoE) is affected by the factors in different domains, mainly related to network information (e.g. slice status, applied Policy), UE information (e.g. UE power, Trajectory) and Application layer information (e.g. applied codec, QoE metric).
- Due to security, privacy and/or commercial considerations, the local data in each domain cannot be shared with the AF to perform the AI data analytics to determine the ongoing QoE or to predict the QoE. Instead, by applying the VL technique with the collaboration with the UE and the 5GC NF(s) over the 5G system, the information of each domain (i.e. UE, 5GC NF(s)) corresponding to specific UE can be collected by the AF to apply the AL data analytics.
- As the result, by applying the cross domain VL, the QoE corresponding to the target UE can be determined/predicted at the AF.





Gap Analysis:

- ☑ AF initiates the cross-domain V-FL operation identified by a specified Label with the target UE and selected 5GC NFs to collaborate on the AI model training/inference operation
- ☑ With the assistance of 5GC, AF selects the target 5GC NF(s) via Service Registration and/or Discovery and/or Provisioning to participate in the V-FL operation for AI model training/inference operation of which the AI training support includes the feature alignment across the 5GC and UE domains.
- ☑ 5GC assistance for the service and policy configuration required by the execution of V-FL operation within 5GS
- ☑ Determine the proper transport mechanism to transfer the AI model training/inference data across the domains

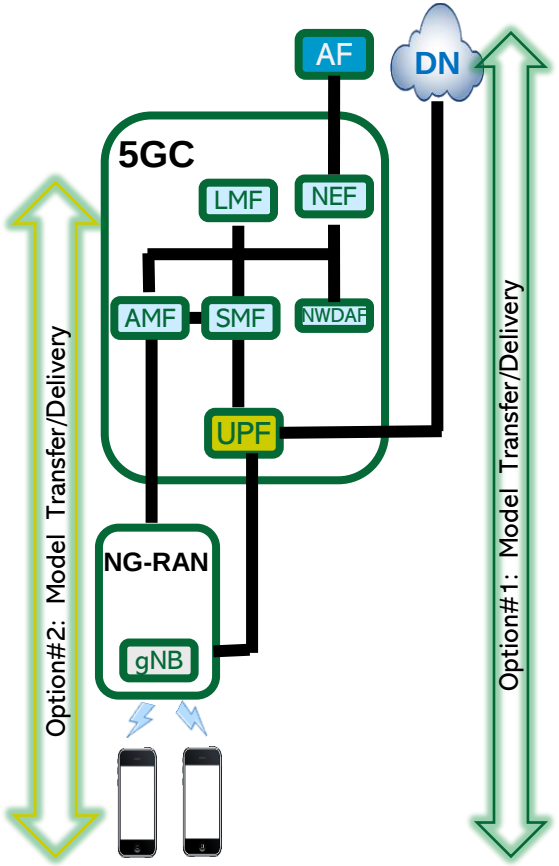
2. 5GC Assistance for RAN AI/ML Model Transfer/Delivery

Background:

- RAN has defined 6 possible options to support RAN AI/ML Model Training/Transfer/Delivery
- Two architecture paths can be taken for Model Transfer/Delivery:
  - Option-1 : (Case-y) OTT approach (similar to functionalities of 5GS assistance for cross-domain AI Training/Inference for VFL discussed previously)
    - OTT server can transfer the AI/ML model to the UE via application layer. Then the 5GC needs to support such transfer e.g., determine the suitable QoS, ensure this model transfer will not impact other service, etc. by reusing the existing mechanism that has been defined in R18 AIMLsys as much as possible.
    - OTT server can transfer the model to the CN, then CN can transfer/deliver AI/ML model(s) to UE via NAS signalling or UP data. Since the model need to be aware in the 5GC to let the 5GC manage the model (e.g., monitor the model performance and do the model registration). OTT can first provision some model info to the CN and then transfer the model to the CN. Or when the UE register it model to the CN, CN can use the provisioned information to do the checking.
  - Option-2: (Case-z) 3GPP approach (i.e., 5GC delivers to the UE)
    - 5GC coordination with RAN to support RAN AI/ML Model Transfer & Delivery

The following Cases further detail the different options for model delivery/transfer to UE, training location, and model delivery/transfer format combinations for UE-side models and UE-part of two-sided models:

Case	Model delivery/transfer	Model storage location	Training location
y	model delivery (if needed) over-the-top	Outside 3gpp Network	UE-side / NW-side / neutral site
z1	model transfer in proprietary format	3GPP Network	UE-side / neutral site
z2	model transfer in proprietary format	3GPP Network	NW-side
z3	model transfer in open format	3GPP Network	UE-side / neutral site
z4	model transfer in open format of a known model structure at UE	3GPP Network	NW-side
z5	model transfer in open format of an unknown model structure at UE	3GPP Network	NW-side



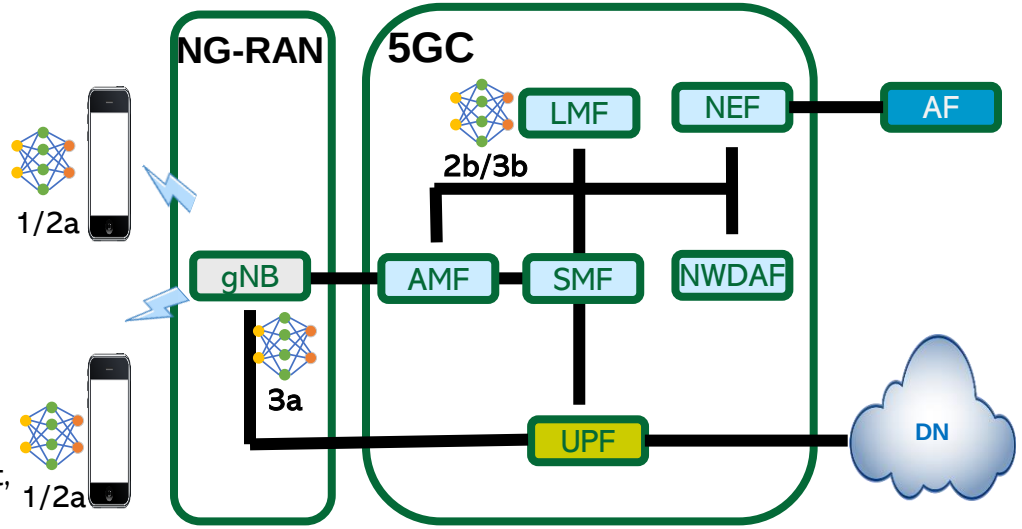
Gap Analysis:

- ✅ **Transport mechanism and QoS/Policy control to support RAN AI/ML Model delivery to UE**
- ✅ **Model Identification/registration between the UE and 5GC for the given AI/ML Model transfer/delivery**
- ✅ **Model information provisioning between CN and OTT**
- ✅ **Data collection and performance monitoring**

3. 5GC Assistance for RAN AI-based Positioning

Background:

- RAN has defined 5 cases to support RAN AI-based Positioning.
  - Case 1: UE-based positioning with UE-side model, direct AI/ML or AI/ML assisted positioning
  - Case 2a: UE-assisted/LMF-based positioning with UE-side model, AI/ML assisted positioning
  - Case 2b: UE-assisted/LMF-based positioning with LMF-side model, direct AI/ML positioning
  - Case 3a: NG-RAN node assisted positioning with gNB-side model, AI/ML assisted positioning
  - Case 3b: NG-RAN node assisted positioning with LMF-side model, direct AI/ML positioning
- Two approaches are considered:
  - Direct AI/ML positioning (D-AIML):
    - AI/ML model output: UE location
    - e.g. fingerprinting based on channel observation as the input of AI/ML model
  - AI/ML assisted positioning (A-AIML):
    - AI/ML model output: new measurement and/or enhancement of existing measurement
    - e.g. assistance information - LOS/NLOS identification, timing and/or angle of measurement, likelihood of measurement



Gap Analysis:

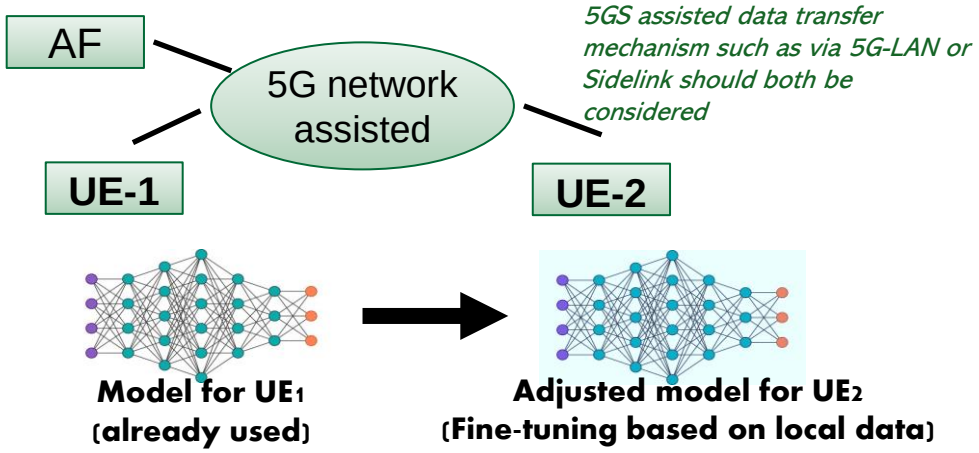
☑ Impact to 5GC/LMF to support AI training/inference of the positioning model w.r.t. D-AIML or A-AIML approach
e.g., UE selection support for the positioning model training/inference

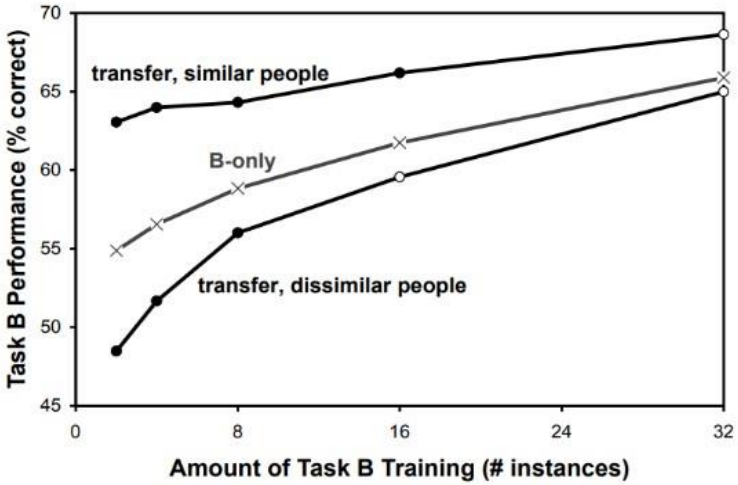
☑ Data collection for AI-based positioning model training/inference over the air
- data collection for offline training/inference/monitoring
- other AI/ML-related signaling (Model control, management)

4. 5GS Assistance for Transfer Learning

Background:

- Usually, AI model is acquired from a pretrained knowledge instead of starting from scratch. Transfer learning is a key technique for efficient/optimal model training.
- Fine-tuning, as the widely used transfer learning technique, helps to train model for UEs which share similar user characteristics/factors such as location, QoS, trajectory, applied policy, etc.
- Transfer Learning is an effective technique to support model training efficiently.





(Reference: To Transfer or Not To Transfer, Massachusetts Institute of Technology, MIT, Michael T. Rosenstein, et al.)

Gap Analysis:

- ☑ In R18, Member UE selection assistance functionality for federated learning has been defined. This functionality can be extended to support Transfer Learning when it is requested by the 3rd party AF.
- ☑ Policy & QoS control and service provisioning for the traffic for Transfer Learning needs to be studied
- ☑ Study the proper transport mechanism over the sidelink to transfer the AI model between UEs

5. Leveraging Sidelink for AIML service for Model Splitting

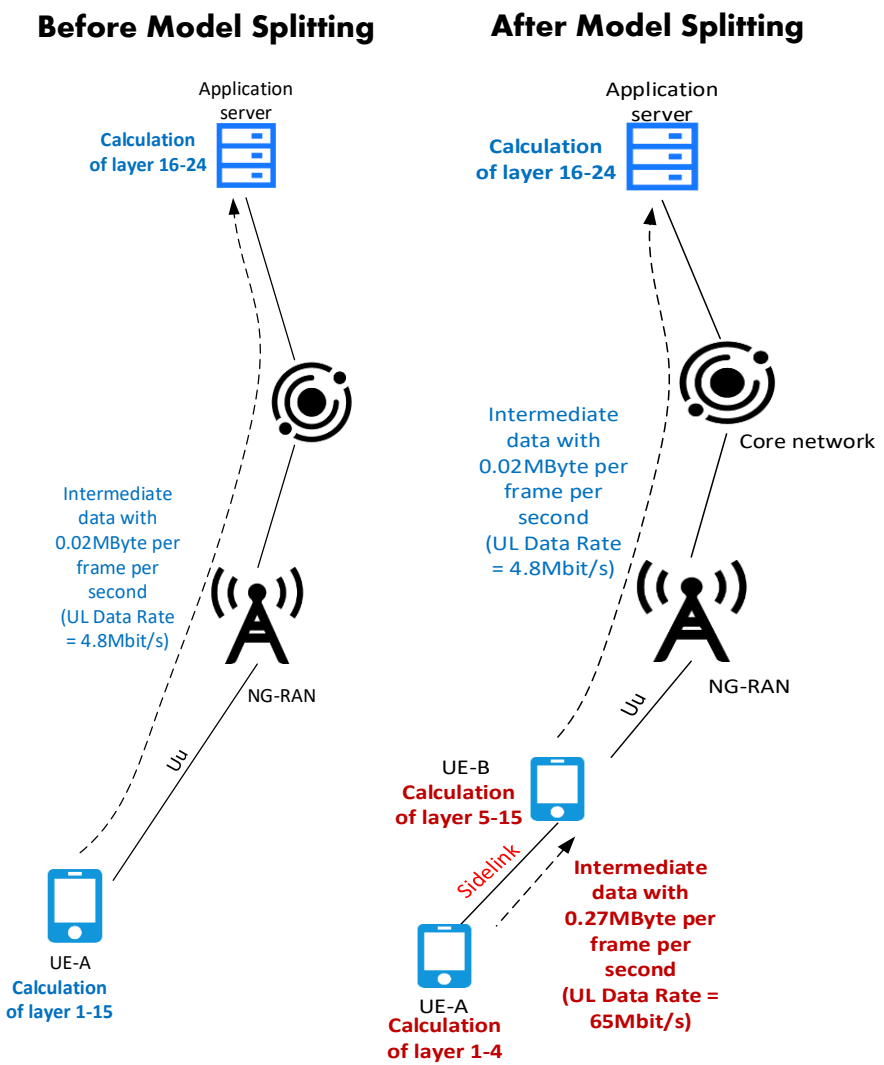
Background:

- **Task offloading based on model splitting**

  Offloading AI inference task from one UE to another UE in proximity. By doing so, the UE who has interesting data set can be used while the computation work task can be offloaded to others, and Uu interface capacity impact can be reduced.

Gap analysis:

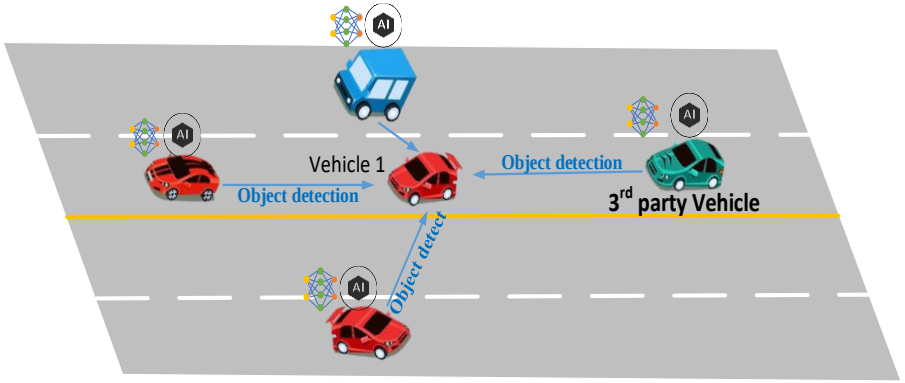
- ✅ Mechanism to select the Relay-UE and Remote-UE to assist the federated learning needs to be extended to enable Model Splitting when it is requested by the 3rd party AF.
- ✅ Determine the proper transport mechanism over the side link, such as Policy & QoS control and service provisioning for the traffic between the Uu and Sidelink to support the Transfer Learning need to be studied



6. Leveraging Sidelink for AIML service for Distributed learning and inference

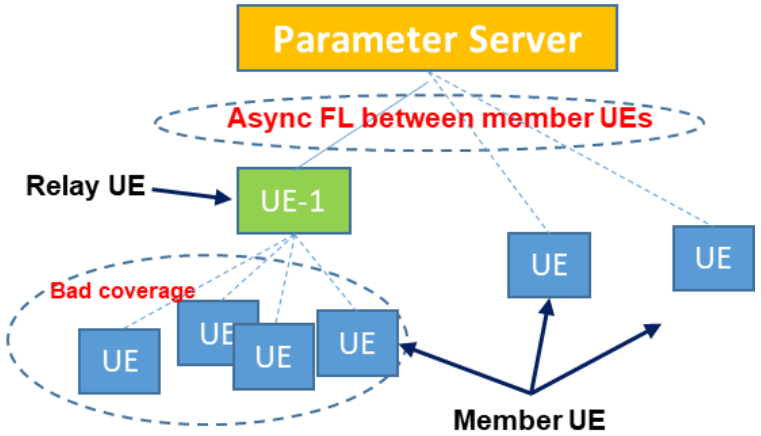
Background:

- **Sidelink based Federated Horizontal Learning & Inference**
  - Leveraging sidelink to provide backup connectivity with the AF for the UE(s) who are out of coverage or have bad QoS over Uu interface after being selected to participate in the FL operation;
  - Monitoring the aggregated QoS threshold at the Relay-UE connecting to the Remote-UE(s) which are participating in the FL and/or Inference operation;
  - Applying the Member selection for target UE(s) participating in the FL/Inference operation.



Gap analysis:

- ☑ Studying any extension to existing PC5 discovery of the Relay UE. 5GS assistance (e.g. leverage member UE selection) to find proper UEs for the joint inference task
- ☑ Studying any Policy & QoS control, service provisioning extensions and service continuity for supporting FL/Inference operation over sidelink
- ☑ Supporting the aggregated QoS threshold monitoring to ensure no overload to the Relay UE when it is connecting the Remote-UE(s) which are participating in the FL/Inference operation.



7. Efficient 5GC Information Exposure to UE for AI/ML operation

Background:

- **QoS information exposure efficiently to UE are necessary in order to achieve an ideal inference performance**
  - For many of ongoing user specific applications (including AI based), its policy enforcement is carried out by the UE
  - 5GAA have identified the scenarios and requirements to ensure high accuracy and rapid notification to UE in order to adopt the changing condition in-time
  - SA1 AMMT has identified the significant role play by QoS for 3rd party's AIML operation
  - UE does not always hold an active connection to application server (case-1: OS level connection; case-2: application-level connection)
  - 3rd party's work for adaptation to 5GC need to be simplified or avoided

Gap Analysis:

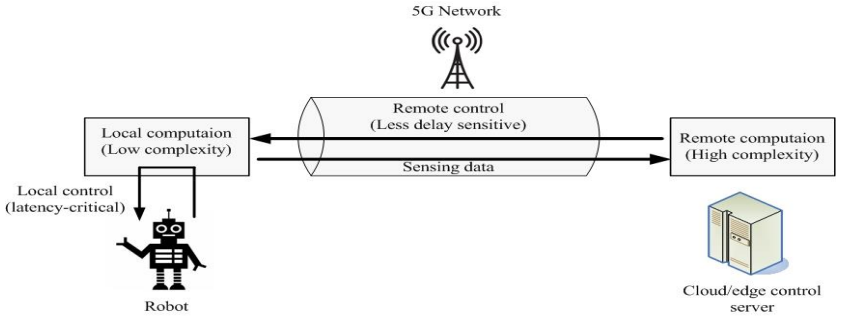
✅ **Efficient/Effective Communication path between 5GC and UE**

Common connection for in-time message delivery especially when UE is in IDLE mode

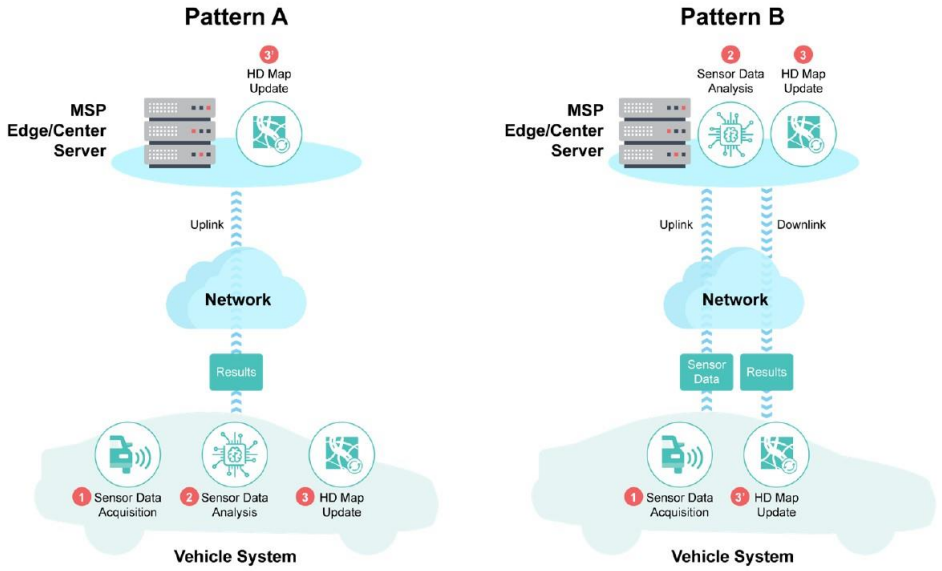
✅ **Real time pre-alerting with High information accuracy**

Reporting the pre-alerting (e.g. QoS Change) within the Allowed Anticipated Notification time with high accuracy, e.g. Event to be happened in 10 seconds.

Use case example-1: AR/VR glasses based motion-to-photon



Use case example-2: Intelligent driving analytics for dynamic HD map

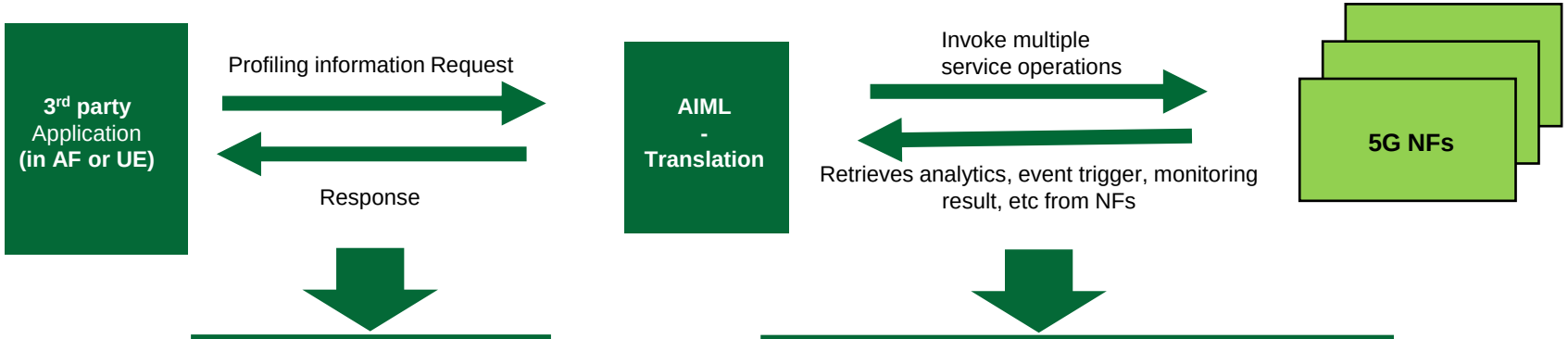


8. AIML-Translation between application request and 5GC APIs

Background:

- More and more newly defined 5GC APIs (Service operations) makes the adaptation work complicated for a 3rd party application (both UE and AF)
- Mapping and coordination between application requests (Profiling information similar to setting up the QoS profile) and 5GC service operation will be much efficient and effective to expedite the service deployment

Gap Analysis:



Profiling assistance information
Prediction of UE service experience in a geographical area
Real time UL/DL packet latency
……

Service operations/messages need to be invoked
UE ID retrieval, Mapping between Geographical and TAI/Cell list Nnrf_NFDiscovery Analytics ID = "Service Experience“ with Area of Interest
Nbsf_Management_Discovery service operation Npcf_PolicyAuthorization_Subscribe Nsmf_EventExposure_Subscribe service operation Configuration message over N4
……

Conclusions – Common Study for multi-domain AI/ML support for 5G Advanced

Rel-18

AI/ML framework support for different domains have been established.

Rel-19

Common 5GC Stage-2 study prevents overlapped objectives and solutions to be defined and to ensure consistent AI/ML architecture across multiple domains.

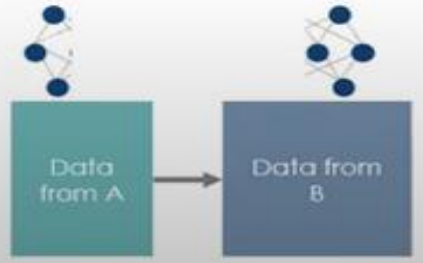
With the consideration of tight Rel-19 timeline, common study for AI/ML support across multiple domains is the right approach to conserve effort and to achieve maximum results.

Let's get it done!

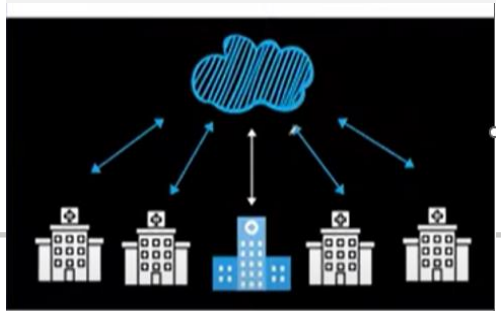
Backup Slides

Concepts - Different Distributed/Federated Learning/Inference Techniques

	Horizontal Federated Learning	Vertical Federated Learning	Transfer Learning	Splitting Learning
Description	Data providers have similar data structures (e.g. text data on mobile phones). Federated is used to increase the size of the data.	Data providers have different data modalities but share overlapping clients (e.g. Banking, Background Check)	A classifier trained on one dataset is used to predict a label on a second dataset.	The execution of an already-trained AI/ML model is divided into two parts: A fraction of the model is executed at the UE, while the remaining part runs at a network endpoint, which then communicates the result to the UE.
Data separation	Same features Different samples	Same samples, Different features	Different features Different samples	



Example: MRI images in different hospitals



Example: Financial Background Check

Loan	Repayment
$6,120	$6,120
$11,560	$10,000
$89,478	$0

Credit Check company

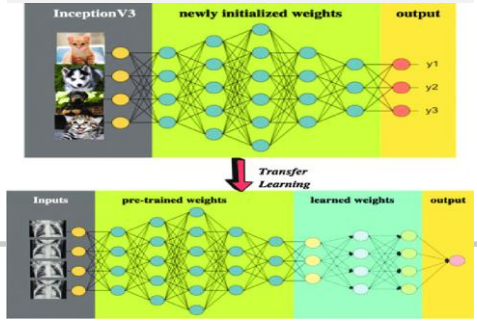
Monthly deposit	Account balance
$10,100	$78,456
$0	$11,120
$0	$561

Bank

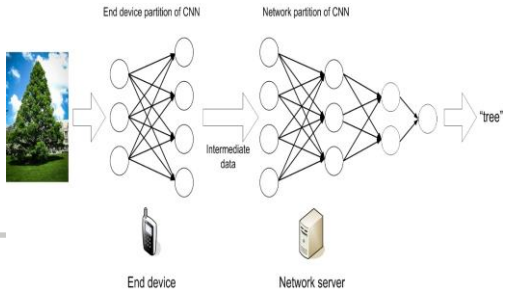
Online Shopping
$14,256
$4,842
$189

E-Commerce company

Example: Imaging Recognition



Example:



рахмат
danke
謝謝
ngiyabonga
teşekkür ederim
Баярлалаа
спасибо
faafetai lava
mersi
barka
welalin
tack
spas
kia ora
vinaka
спасибі
blagodaram
dank je
misaotra
matondo
paldies
grazzi
mahalo
tapadh leat
thank you
gracias
хвала
asante
manana
obrigada
murakoze
tenki
kiitos
dankie
dhanyavad
nanni
nandri
bayarlalaa
gràcie
hvala
maururu
köszönöm
aciù
dankon
akun
djiere dieuf
tau
mochchakkeram
enkosi
bedankt
dziękuję
дякую
mamnun
chnorakaloutioun
gratias ago
gràcies
sulpáy
go raibh maith agat
chokrane
sobodi
dèkuji
sagolun
sukriya
kop khun krap
taiku
obrigado
mèsi
didi madloba
najis tuke
ありがとう
tanemirt
grazie
arigatô
takk
dakujem
trugarez
terima kasih
rahmet
diolch
dhanyavadagalu
shukriya
merci
мерси
kam sah hamnida
rahmat
তোমাকে ধন্যবাদ
감사합니다
xièxie
ευχαριστώ
merci