Motivation for AI enabled 5G System

Xiaomi

© 3GPP 2021

AI enabled 5G System

Motivations (RAN1)

- Study on Artificial Intelligence (AI)/Machine Learning (ML) for NR air interface (RP-221348) has started to study the AI framework to enable performance improvement by using AI models in UE/NG-RAN in order to support following scenarios, e.g.
  - AI based CSI,
  - AI based beam management
  - AI based positioning
- RAN1 has identified 6 cases in the table,
  - Model distribution
  - Model storage location
    - Out of 3GPP network
    - 3GPP network (e.g., NWDAF, OAM, etc.）
  - Training location
    - UE side/Neutral side/ NW-side (e.g., NWDAF, OAM, etc.）

Case	Model distribution (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

- **Gaps:** currently SA2 doesn’t support the RAN use cases above, 5GS architecture needs to be enhanced to support:
  - Lifecycle management (LCM) for AI models
  - AI model storage, AI model training, AI model distribution (e.g., to UE/NG-RAN/5GC NF(s))
  - Performance monitoring, Data collection

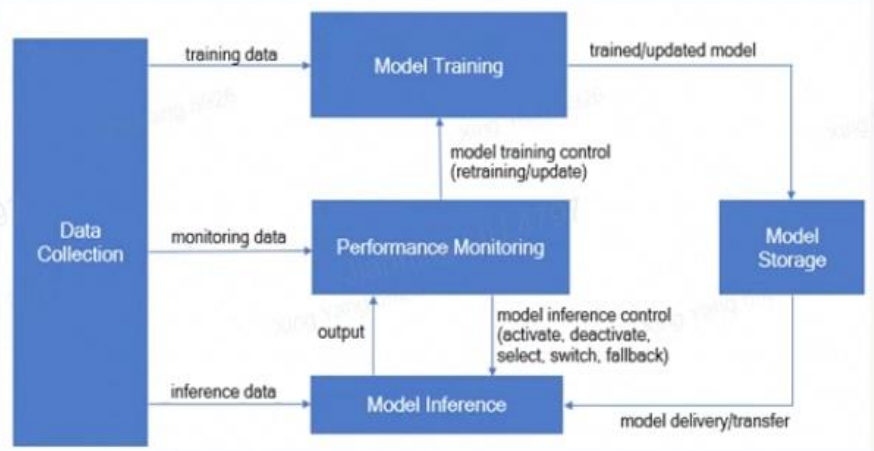
© 3GPP 2021

AI enabled 5G System

Motivations (RAN2)

- For the AI Model distribution, RAN2 has studied different solutions in the table, from gNB to UE, CN/LMF to UE, Server (e.g., OAM, OTT) to UE, including Control Plane and User Plane solutions.
- RAN2 has concluded that the general AI/ML framework consist of, (i) Data Collection, (ii) Model Training, (iii) Model Management, (iv) Model Inference, and (v) Model Storage.

Solution 1a: gNB can transfer/deliver AI/ML model(s) to UE via RRC signalling.
Solution 2a: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via NAS signalling.
Solution 3a: LMF can transfer/deliver AI/ML model(s) to UE via LPP signalling.
Solution 1b: gNB can transfer/deliver AI/ML model(s) to UE via UP data.
tSolution 2b: CN (except LMF) can transfer/deliver AI/ML model(s) to UE via UP data.
Solution 3b: LMF can transfer/deliver AI/ML model(s) to UE via UP data.
Solution 4: Server (e.g. OAM, OTT) can transfer/delivery AI/ML model(s) to UE (e.g. transparent to 3GPP).



- **Gaps:** based on the general AI/ML framework, current 5GS architecture cannot support the AI/ML framework. SA2 needs to study how to distribute these entities (i.e. Data Collection, Model Training, Model management, Model Inference and Model storage) to 5GS architecture.

© 3GPP 2021

AI enabled 5G System

Motivations (RAN3)

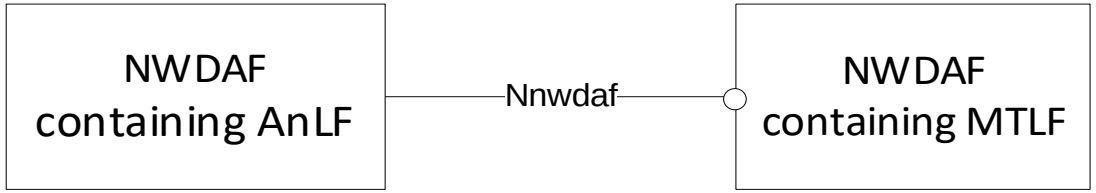
- Artificial Intelligence (AI)/Machine Learning (ML) for NG-RAN (RP-220635) has started to specify data collection enhancements and signaling support within existing NG-RAN interfaces and architecture for the following scenarios, e.g.
  - AI/ML-based Network Energy Saving
  - AI/ML-based Load Balancing
  - AI/ML-based Mobility Optimization
- For the deployments of RAN intelligence, following scenarios may be supported in R18:
  - AI/ML Model Training is located in the OAM and AI/ML Model Inference is located in the gNB.
  - AI/ML Model Training and AI/ML Model Inference are both located in the gNB.
- **Gaps:**
  - Existing RAN3 mechanism doesn't support multi-vendor interoperability and collaboration with 5GC AI/ML enabled functions, which limits the scenarios and performance of AI/ML for NG-RAN.
  - New uses cases such as AI/ML based QoE optimization, AI/ML based network slicing management may need collaboration with 5GC in data collection, model training and model transfer and management.

© 3GPP 2021

AI enabled 5G System

Motivations (SA2)

- Enablers for Network Automation for 5G (FS_eNA_Ph2/3), which provides/consumes the AI models within NWDAF, and sharing the AI models between NWDAF(s).
  - Analytics logical function (AnLF): performs inference, derives analytics information and exposes analytics.
  - Model Training logical function (MTLF): trains Machine Learning (ML) models and exposes new training services.



- Observation: currently the AI models is not supported to be shared by other 5GC NFs. It is benefit for NWDAF to share the AI models with other NFs, e.g., NG-RAN, 5GC NF(s) or UE.

© 3GPP 2021

AI enabled 5G System

Potential architecture enhancement (option 1)

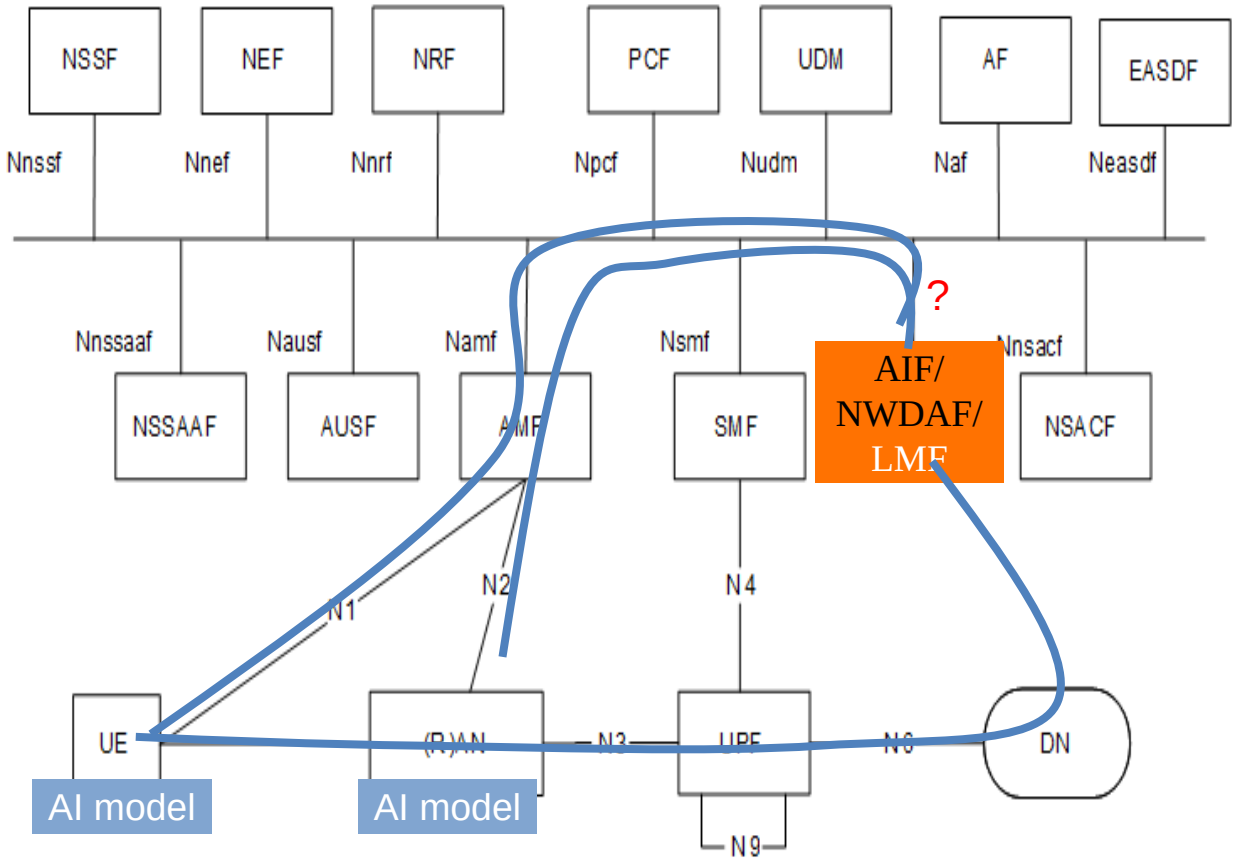
- **Option1** (case Z1-Z5)
  - AI model storage location: Network

    e.g., NWDAF/LMF or new NF
  - AI model training location: UE, and (R)AN
  - AI model inference: UE, RAN
  - AI model distribution:

    - From NWDAF/new NF/LMF to (R)AN

    - From NWDAF/new NF/LMF to UE

      - Control plane
      - User plane



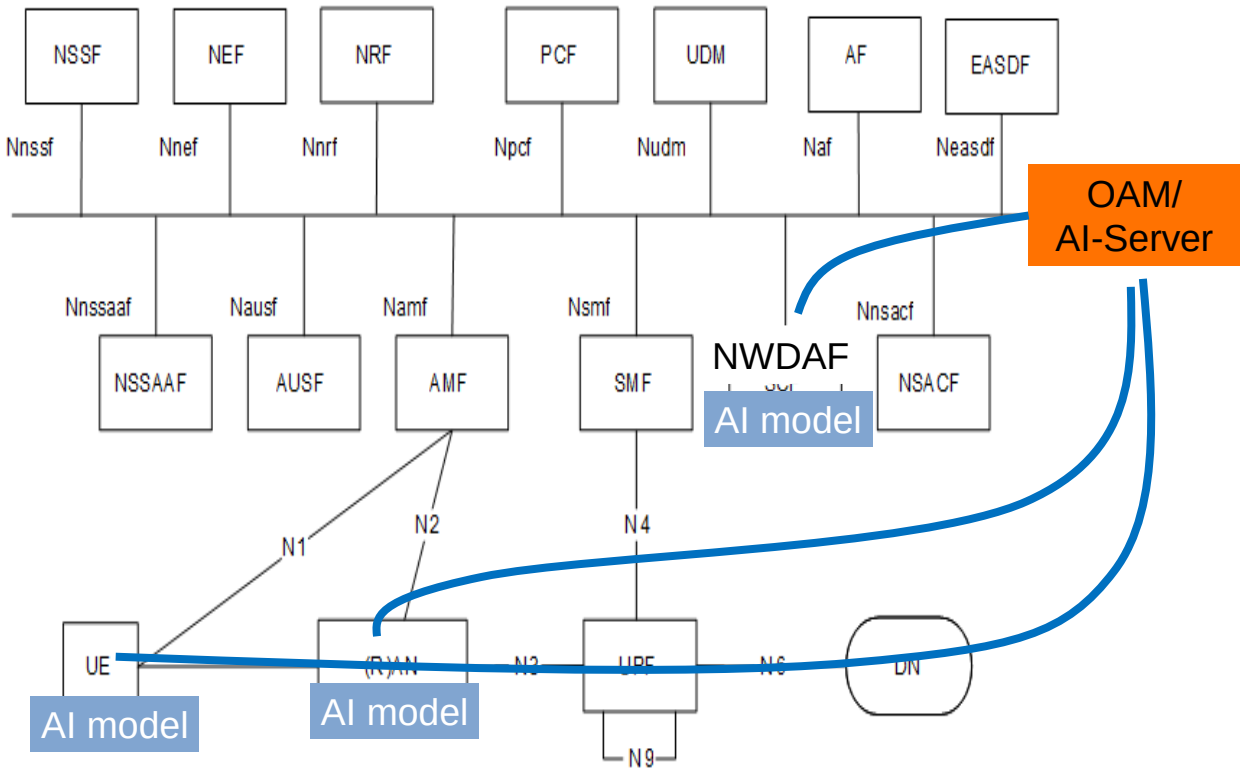
© 3GPP 2021

AI enabled 5G System

Potential architecture enhancement (option 2)

- **Option2 (case Y)**
  - AI model storage location: outside 3GPP

    e.g., OAM, or AI-Server
  - AI model training location: UE, NWDAF, OAM
  - AI model inference: UE, RAN
  - AI model distribution:
    - From OAM to (R)AN
    - From OAM to NWDAF
    - From AI-Server to UE via user plane



© 3GPP 2021

AI enabled 5G System

Potential SA2 Objectives

This SID is aiming to introduce AI technologies to 5G System for the enhancement of radio resources management, positioning/location, etc.

- **The potential Objective includes:**
  - Identify the architecture impacts based on the use cases provided from RAN (i.e. Y, Z1-Z5)
  - Study on how to distribute the entities of AI framework in RAN into 5G architecture, including Data Collection, Model training, performance monitoring, Model interference, Model storage, etc. How to apply the AI enabled 5G architecture to support the scenarios identified by RAN [e.g., AI based CSI/beam management/position]
  - The management/operations of AI models by the NG-RAN, 5GC NF or UE, including,
    - AI models registrations/deregistration
    - AI models creation/updates/removal
    - AI storage/AI models training/inference
    - AI models distribution to the required NG-RAN, 5GC NF or UE, e.g. over user plane or control plane
    - Data collection from UE/NG-RAN/5GC NF(s)/application server for AI models training.
  - QoS mechanism enhancement to support, e.g., the delivery of AI models.
  - Security

© 3GPP 2021

Thank You!

© 3GPP 2021