ML/AI integration in 5G and Beyond 5G

Dejan Vukobratovic
Professor
Faculty of Technical Sciences
University of Novi Sad

ASMS/SPSC 2025
27.02.2025, Sitges, Spain

Dejan Vukobratovic

Professor

Head of Communications and Signal Processing Group

Homepage: https://sites.google.com/view/vukobratovic

Google Scholar:

https://scholar.google.com/citations?user=MugUYHgAAAAJ

The Institute of Artificial Intelligence R&D of Serbia

Affiliated Senior Researcher

Webpage: https://ivi.ac.rs/en/

Founding director of ICONIC centre

ICONIC: Centre for intelligent communications, networking and information processing

https://iconic.ftn.uns.ac.rs/

University of Novi Sad

http://www.uns.ac.rs/

Faculty of Technical Sciences

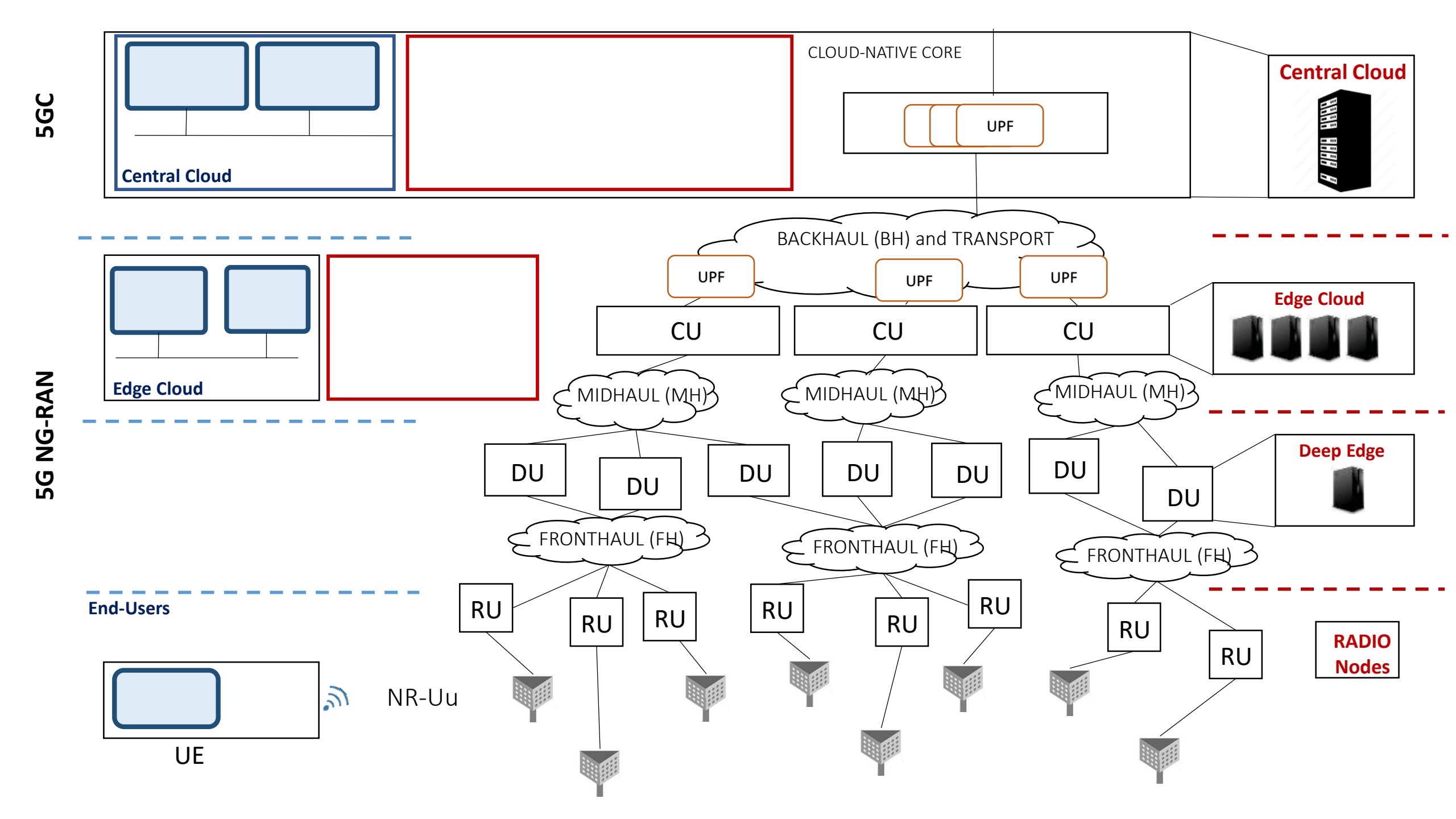
http://ftn.uns.ac.rs/

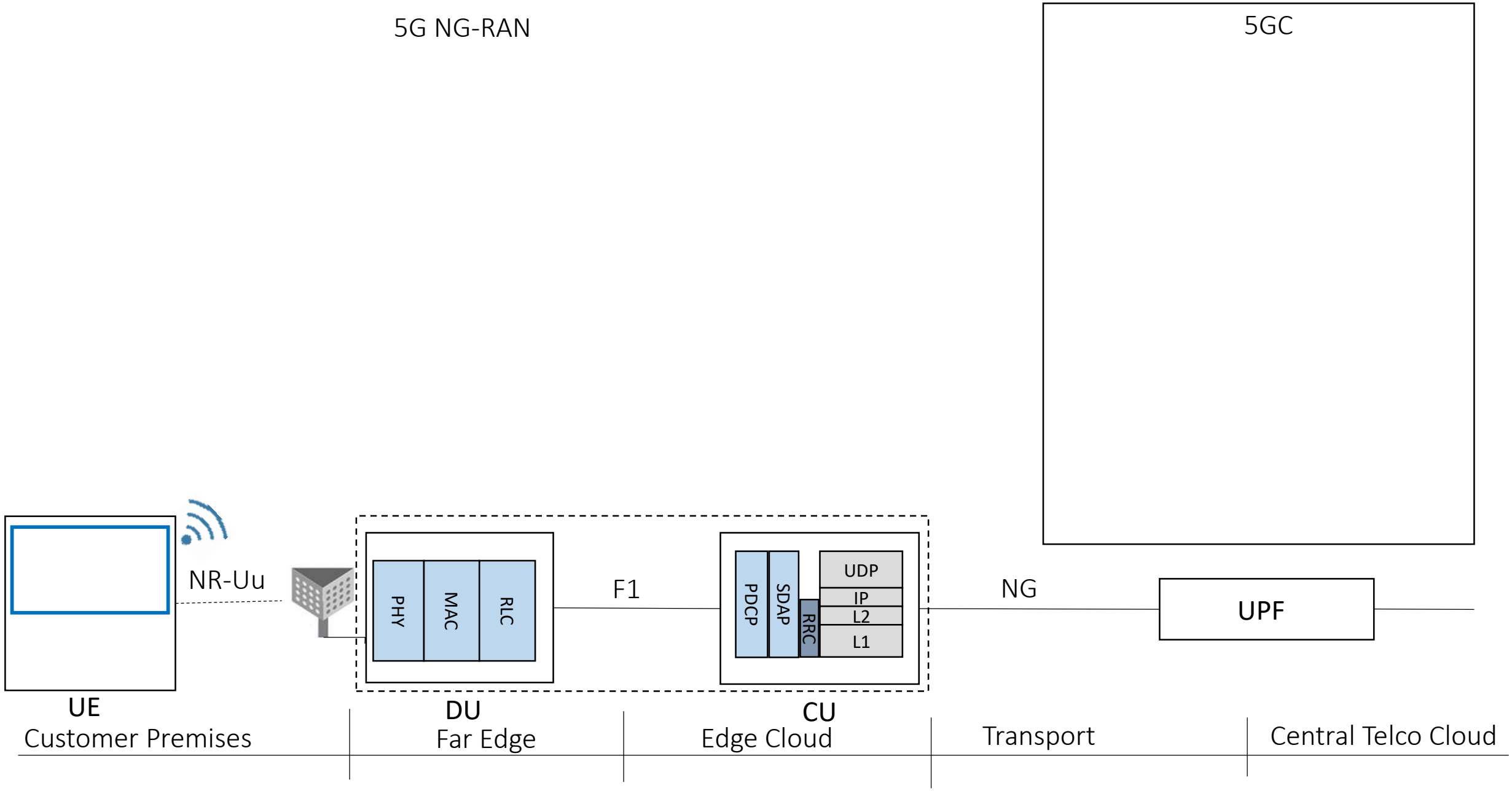
Department of Power, Electronic and Communications Engineering

http://deet.ftn.uns.ac.rs/

Communications and Signal Processing

Electronics


5GC
Central Cloud
CLOUD-NATIVE CORE
UPF
Central Cloud
5G NG-RAN
Edge Cloud
BACKHAUL (BH) and TRANSPORT
UPF
UPF
UPF
CU
CU
CU
Edge Cloud
MIDHAUL (MH)
MIDHAUL (MH)
MIDHAUL (MH)
DU
DU
DU
DU
DU
DU
DU
Deep Edge
FRONTHAUL (FH)
FRONTHAUL (FH)
FRONTHAUL (FH)
RU
RU
RU
RU
RU
RU
RU
RU
RADIO Nodes
End-Users
NR-Uu
UE


5G NG-RAN
5GC
NR-Uu
PHY
MAC
RLC
F1
PDCP
SDAP
RRC
UDP
IP
L2
L1
NG
UPF
UE
Customer Premises
DU
Far Edge
CU
Edge Cloud
Transport
Central Telco Cloud

Outline of the talk

- AI/ML in the Core Network
- AI/ML in the RAN
- AI/ML for the PHY
- AI/ML at the Application Layer
- AI/ML-driven Network Management and Orchestration

Outline of the talk

- **AI/ML in the Core Network**
- AI/ML in the RAN
- AI/ML for the PHY
- AI/ML at the Application Layer
- AI/ML-driven Network Management and Orchestration

5G Core (5GC) Network

Cloud-Native Service-Based Architecture (SBA)

- Collection of virtual Network Functions (NFs)
- Sofware-based implementation in virtualized environment
- NFs offers one or more services via their APIs to other NFs/consumers

Support for Network Slicing

- Support for services with different requirements
- Creating multiple virtual networks on a shared physical infrastructure
- 5G slice provides complete network functionality (RAN, CN, transport)

3GPP **TS 23.501**, “System architecture for the 5G System (5GS),” V18.0.0, December 2022.

AI/ML in the 5G Core Network

ML/AI Integration in 5GC

- Initial Steps started in Rel. 15 in 5GC
  - **NWDAF** – Network Data Analytics Function
- **Goal:** Enable automated data collection and data analytics provisioning

Evolution of NWDAF

- Initial function (Rel. 15) to provide network slice analytics (load level)
- Provide data collection and analytics exposure to other NFs (Rel. 16)
- UE apps data collection, ML/AI model training/deployment (Rel. 17/18)

3GPP **TS 29.520**, “5G System; Network Data Analytics Services; Stage 3,” V18.6.0, June 2024.

Network Data Analytics Function (NWDAF)

Main Functions

- Data Collection Interface for Network Nodes
- Predefined Data Analytics Functions
- Data and Analytics Exposure Interface for Authorised Consumers (via NEF)

NWDAF Services (Rel-15/16)

- **Analytics Subscription/Analytics Information**, e.g., Load and Mobility Prediction, Predictive QoE, Slice SLA Assurance
- **ML Model Information/Provision** provides information and model request and retrieval
- **Distributed Implementation** – Central NWDAF (e.g., AI model repository) and Edge NWDAF (low-latency use cases)

Chouman, A., Manias, D.M. and Shami, A., „Towards supporting intelligence in 5G/6G core networks: NWDAF implementation and initial analysis," *IWCMC 2022,* pp. 324-329, 2022.

NWDAF Architecture Refinements (Rel-17/18)

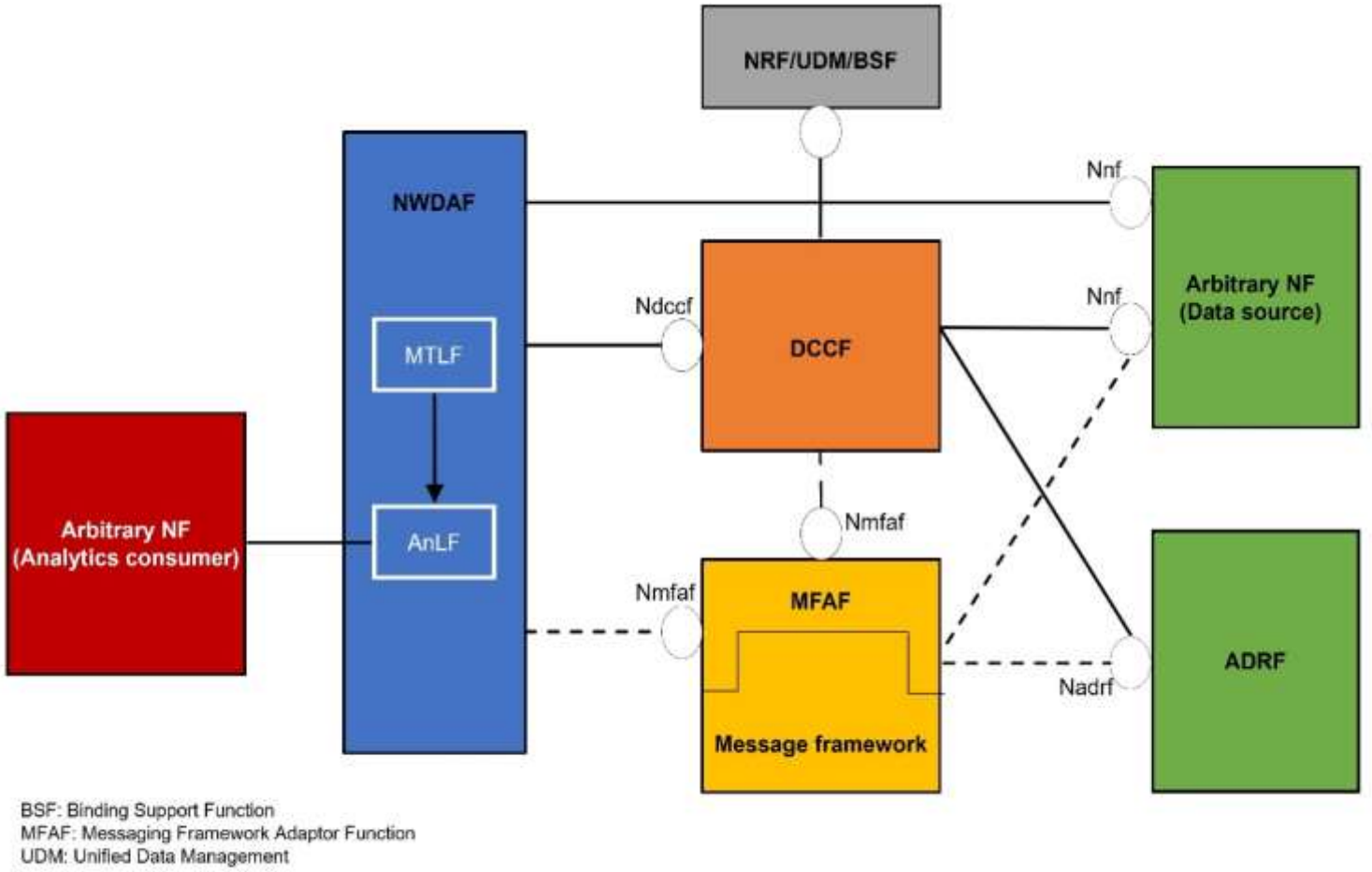
MTLF – Model Training Logical Function (LF)

AnLF – Analytics LF

DCCF – Data Collection Coordination Function

ADRF – Analytics Data Repository Function (data lake)

MFAF – Messaging Framework Adaptor Function

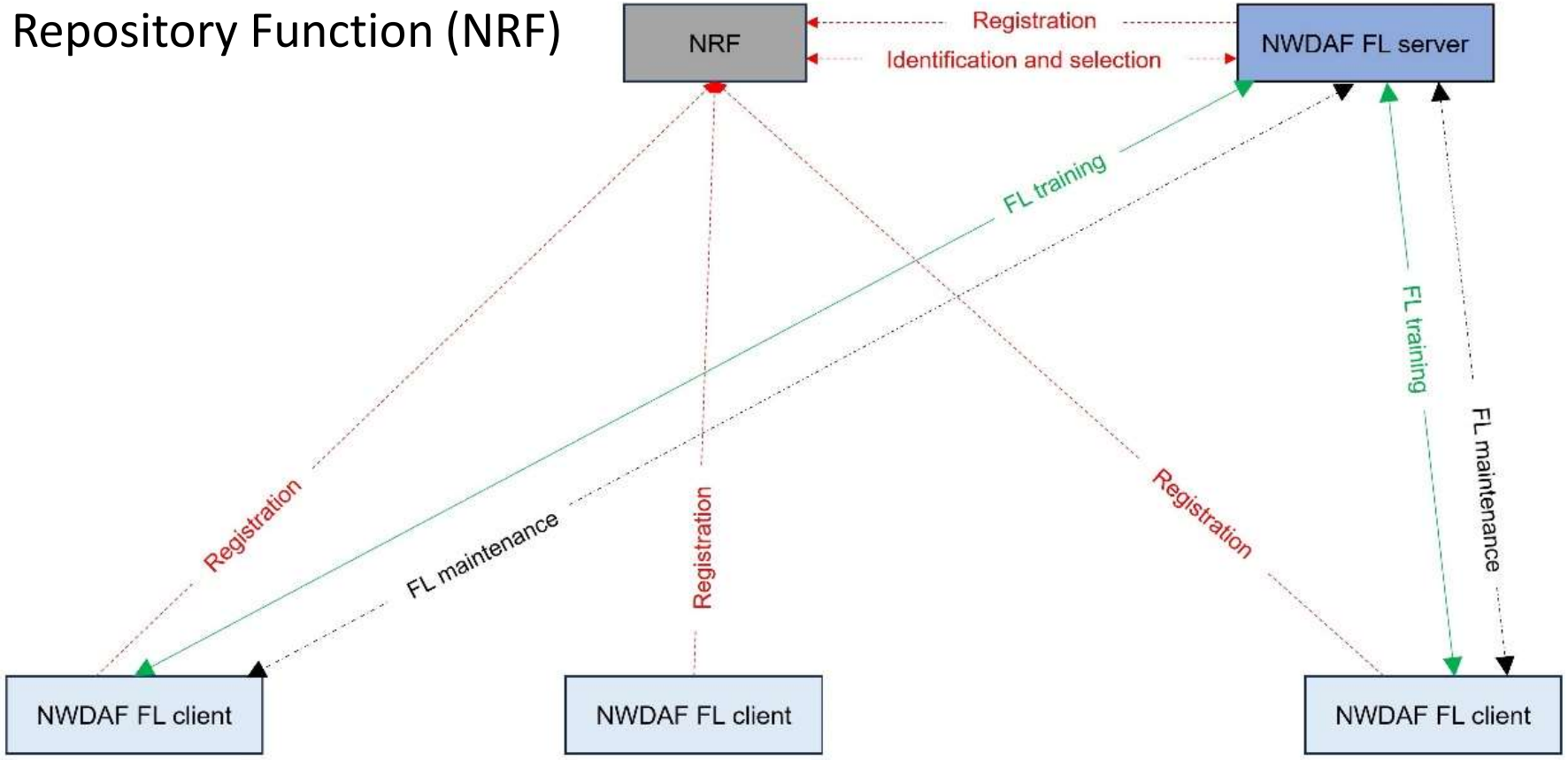


Architecture for network data analytics defined in Rel-17

Figure: https://www.docomo.ne.jp/english/corporate/technology/rd/technical_journal/bn/vol26_3/003.html
3GPP TR23.700-80 Study on 5G System Support for AI/ML-based Services
3GPP TR23.700-81 Study of Enablers for Network Automation for the 5G System (5GS)

NWDAF Federated Learning (Rel-18)

1 – Registration to Network Repository Function (NRF)
2 – FL training iteration
3 – FL maintenance

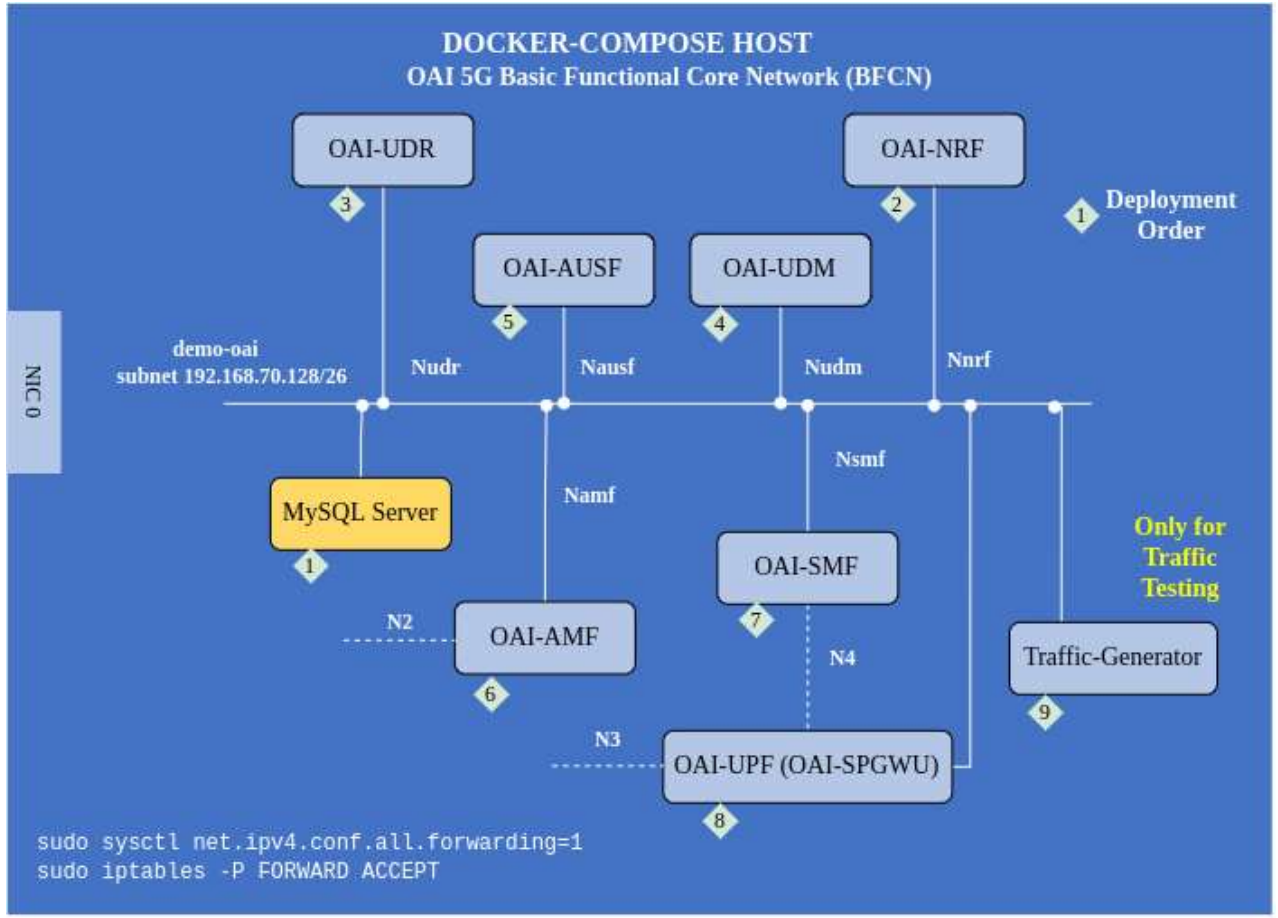


FL between NWDAFs as described in Rel-18

Figure: https://www.docomo.ne.jp/english/corporate/technology/rd/technical_journal/bn/vol26_3/003.html
3GPP TR23.700-80 Study on 5G System Support for AI/ML-based Services
3GPP TR23.700-81 Study of Enablers for Network Automation for the 5G System (5GS)

5G Core (5GC) Open Source Implementations

OpenAir Interface (OAI) 5GC



https://gitlab.eurecom.fr/oai/cn5g/oai-cn5g-fed/-/blob/master/docs/DEPLOY_HOME.md

NWDAF support: https://gitlab.eurecom.fr/oai/cn5g/oai-cn5g-nwdaf/-/blob/master/docs/TUTORIAL.md

ML/AI in 5GC: A Sample of Research

AI-Driven Scaling and Orchestration of 5GC Network

- Sheoran, A., Fahmy, S., Cao, L., & Sharma, P., *AI-Driven Provisioning in the 5G Core. IEEE Internet Computing, 25(2), 18–25, 2021.*
- Atalay, T.O., Stojadinovic, D., Stavrou, A. and Wang, H., Scaling Network Slices with a 5G Testbed: A Resource Consumption Study. *IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 2649-2654, 2022.

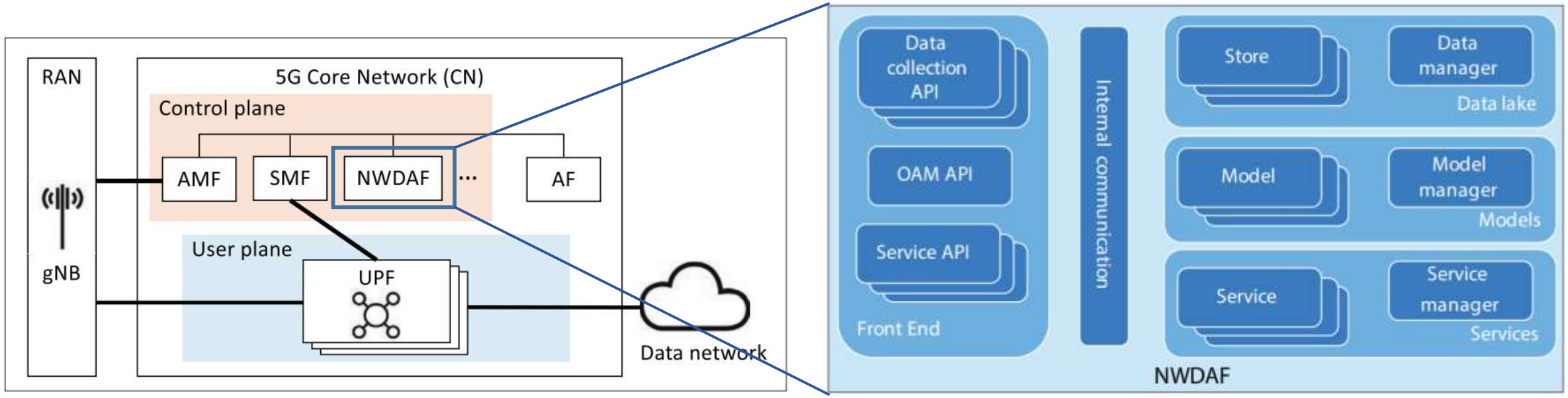
NWDAF implementation studies

- Lee, S., Lee, J., Kim, T., Jung, D., Cha, I., Cha, D., Ko, H. and Pack, S., Design and Implementation of Network Data Analytics Function in 5G. *ICTC,* pp. 757-759, 2022.
- Hossain, M.A., Hossain, A.R., Liu, W., Ansari, N., Kiani, A. and Saboorian, T., A distributed collaborative learning approach in 5G+ core networks. *IEEE Network, 2023*.

NWDAF ML/AI-based Functions

- Murudkar, C.V., Chen, K.C. and Gitlin, R.D., Network Architecture for Machine Learning: A Network Operator's Perspective. *IEEE Communications Magazine*, *60*(7), pp. 68-74, 2022.
- Jeong, J., Roeland, D., Derehag, J., Johansson, Å.A., Umaashankar, V., Sun, G. and Eriksson, G., Mobility prediction for 5G core networks. *IEEE Communications Standards Magazine*, *5*(1), pp. 56-61.

Example: Mobility Prediction in 5GC



- **Two services:** mobility prediction service and UPF area prediction for UPF re-selection
- **Data required**: mobility events streamed by AMF to NWDAF via data collection API
- **Life-cycle management**: Handles aperiodicity of mobility patterns and reduce learning delay, LCM detects concept drift and trigger retraining

---

Jeong, J., Roeland, D., Derehag, J., Johansson, Å.A., Umaashankar, V., Sun, G. and Eriksson, G., Mobility prediction for 5G core networks. *IEEE Communications Standards Magazine*, *5*(1), pp. 56-61.

Management Data Analytics Function (MDAF)

Main Functions

- OAM Data Collection Interface for Network Functions
- Provides Management Analytics Information
- Management Data Analytics Exposure for Authorised Consumers

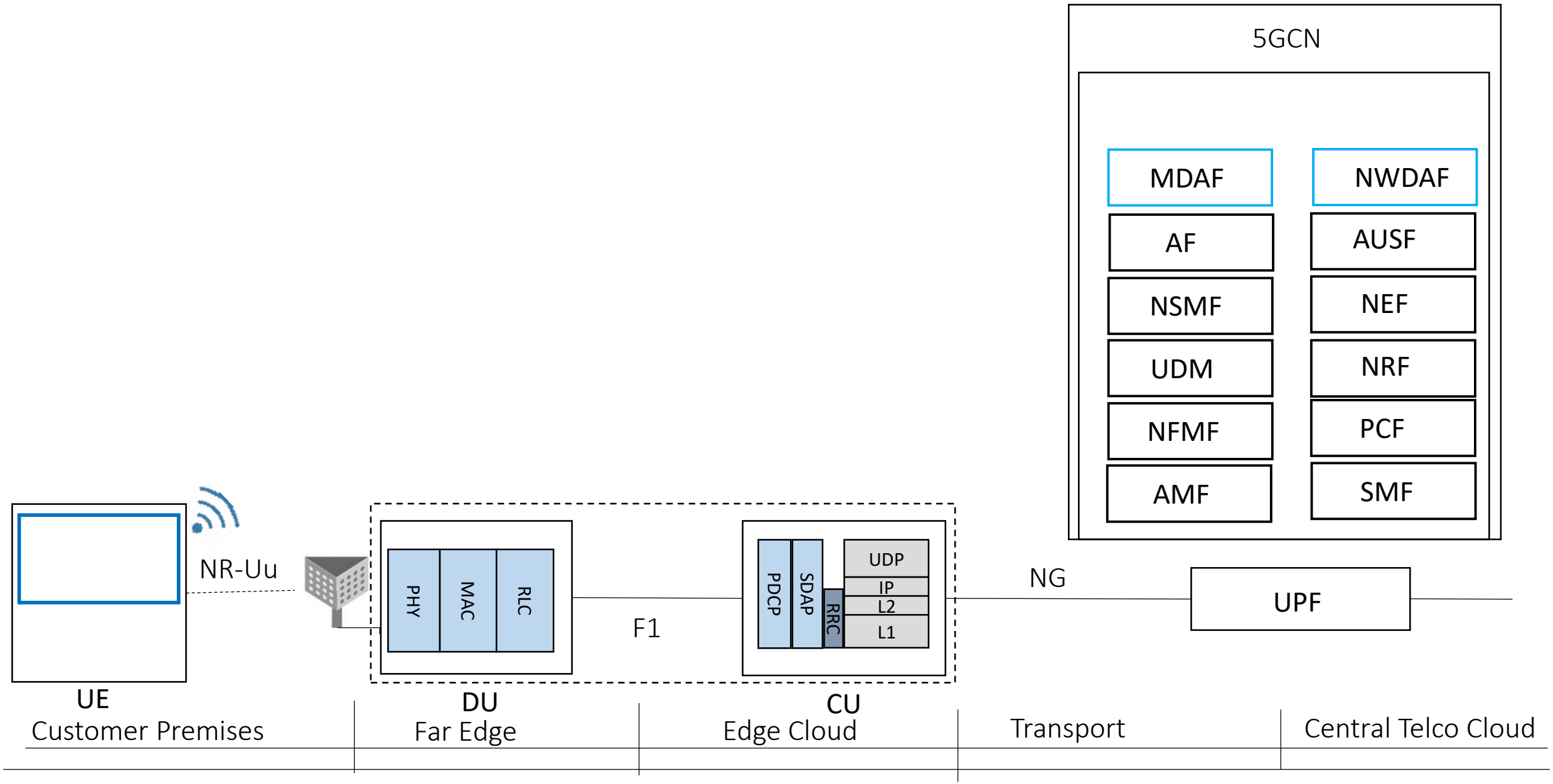
MDAF Services

- **Management Analytics Information** used to recommend appropriate management actions to operator
- Similar data collection and prediction services to NWDAF focused on OAM instead of Control domain

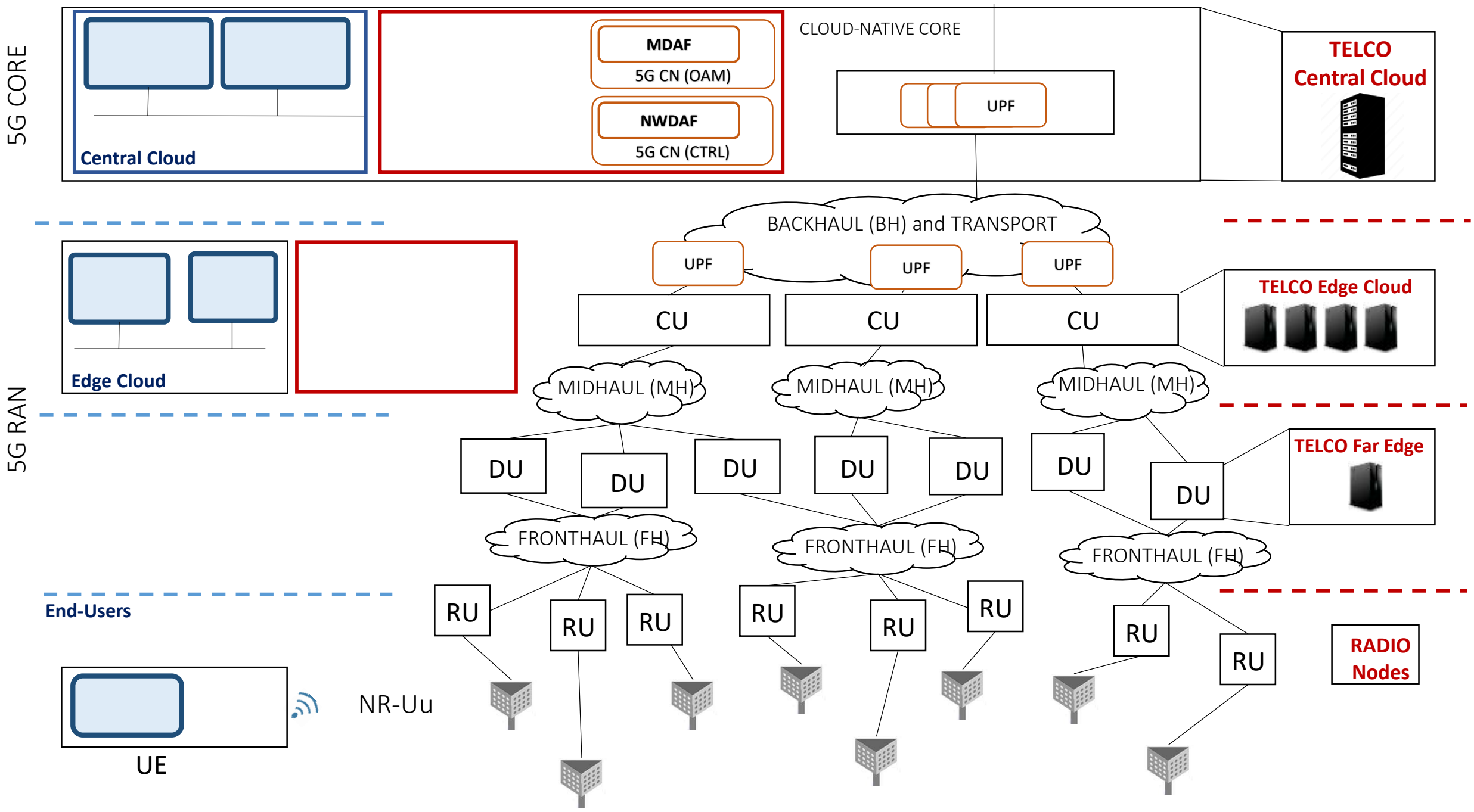
---

Ferrús, R., Sallent, O. and Perez-Romero, J., Data analytics architectural framework for smarter radio resource management in 5G radio access networks. *IEEE Communications Magazine*, *58*(5), pp. 98-104, 2020.

Pateromichelakis, E., Moggio, F., Mannweiler, C., Arnold, P., Shariat, M., Einhaus, M., Wei, Q., Bulakci, Ö. and De Domenico, A., End-to-end data analytics framework for 5G architecture. *IEEE Access*, *7*, pp. 40295-40312, 2019.

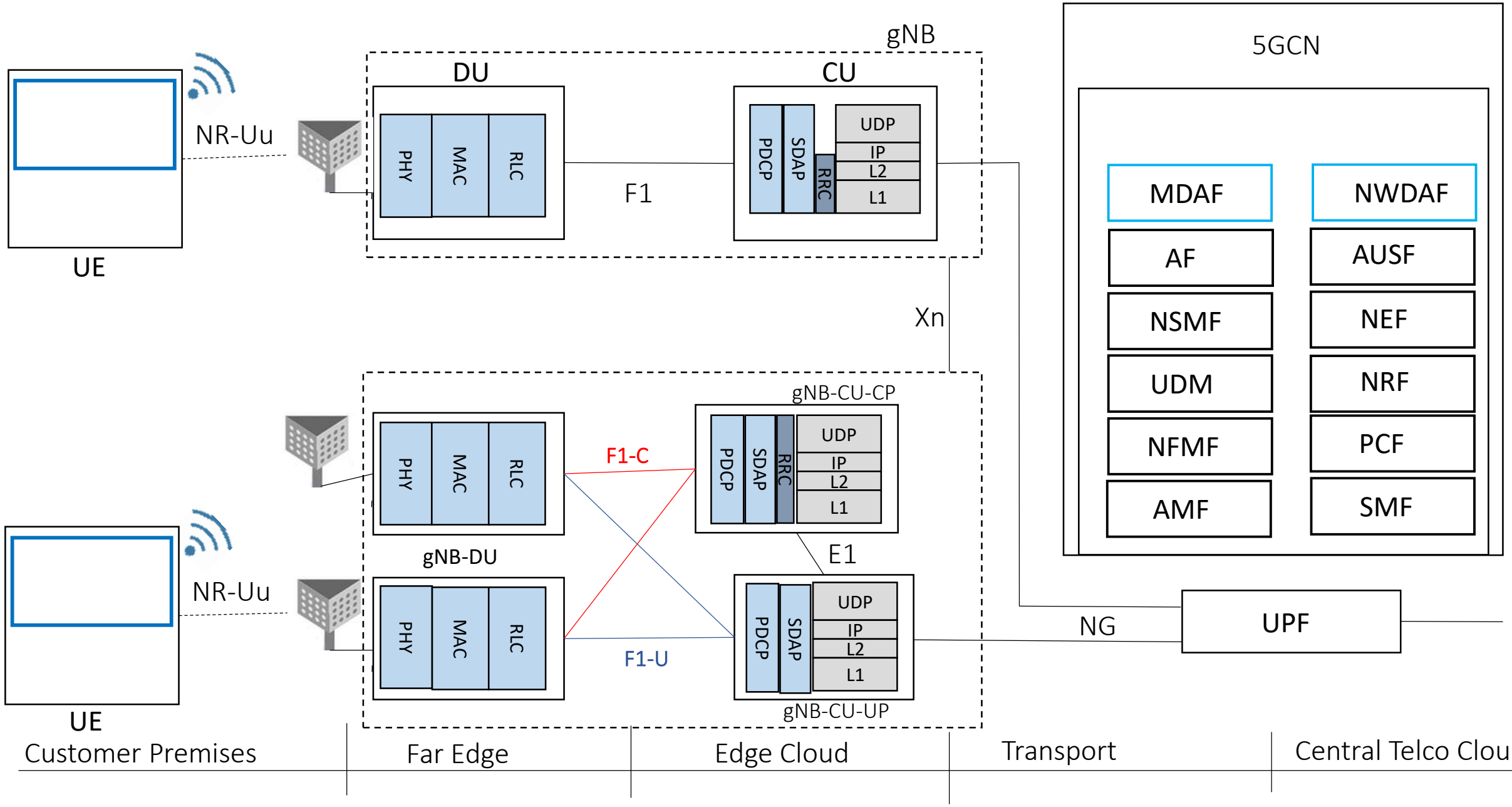


Ferrús, R., Sallent, O. and Perez-Romero, J., Data analytics architectural framework for smarter radio resource management in 5G radio access networks. *IEEE Communications Magazine*, *58*(5), pp. 98-104, 2020.


5G CORE
Central Cloud
MDAF
5G CN (OAM)
NWDAF
5G CN (CTRL)
CLOUD-NATIVE CORE
UPF
TELCO
Central Cloud
5G RAN
Edge Cloud
BACKHAUL (BH) and TRANSPORT
UPF
UPF
UPF
TELCO Edge Cloud
CU
CU
CU
MIDHAUL (MH)
MIDHAUL (MH)
MIDHAUL (MH)
DU
DU
DU
DU
DU
DU
DU
TELCO Far Edge
FRONTHAUL (FH)
FRONTHAUL (FH)
FRONTHAUL (FH)
RU
RU
RU
RU
RU
RU
RU
RU
RADIO
Nodes
End-Users
NR-Uu
UE

Outline of the talk

- AI/ML in the Core Network
- **AI/ML in the RAN**
- AI/ML for the PHY
- AI/ML at the Application Layer
- Case Study: AI/ML for 5G Smart Grids



3GPP **TS 38.401**, "NG-RAN Architecture Description," V18.2.0, July 2024.

3GPP ML/AI study for 5G RAN

Initial Technical Study Item

- 3GPP **TR 37.817**, "Study on enhancement for data collection for NR and EN-DC," V17.0.0, April 2022.

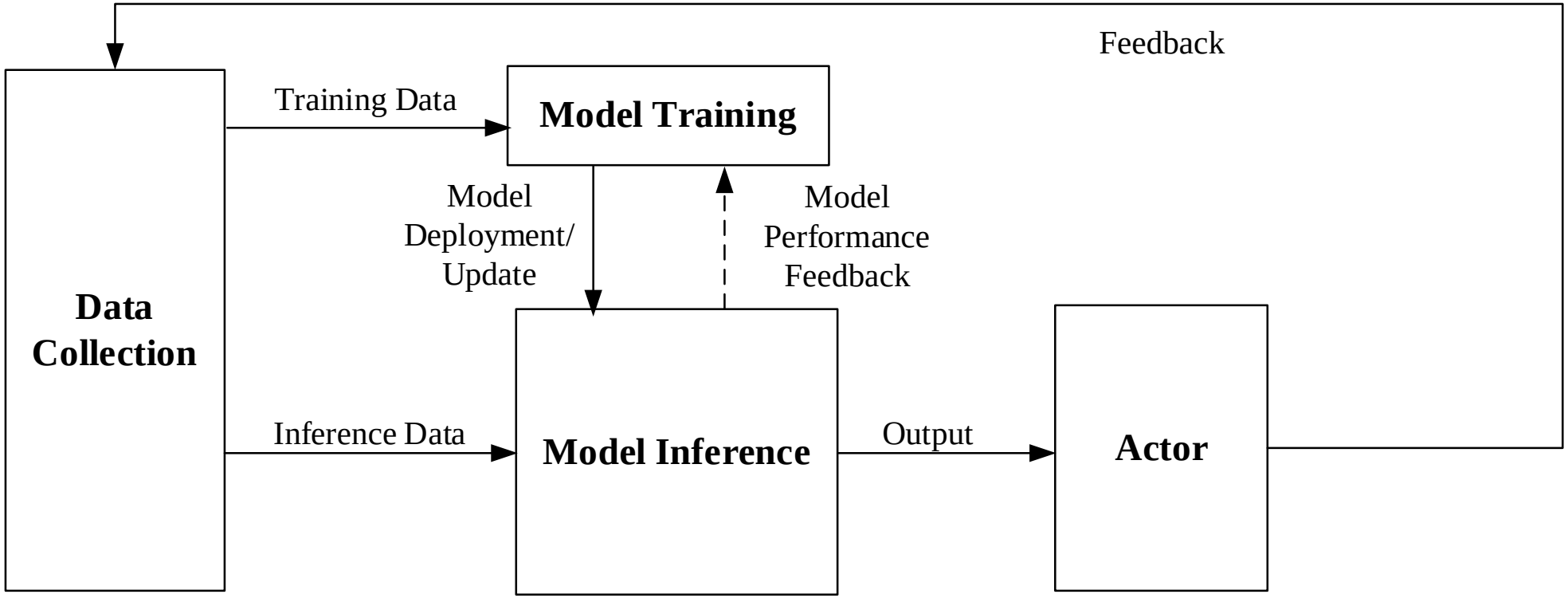
Main Outcomes

- Functional Framework for RAN Intelligence
- AI-enabled RAN for three use cases:
  - **Network Energy Saving** via traffic offloading and cell deactivation
  - **Load Balancing** via prediction-based across RAN cells and multiple-RATs
  - **Mobility Optimization** via UE cell association
- Led to Approval of Rel. 18 Study on AI/ML for 5G NR RAN

---

https://www.3gpp.org/technologies/ai-ml-nr

3GPP ML/AI study for 5G RAN



Functional Framework for RAN Intelligence

3GPP **TR 37.817**, "Study on enhancement for data collection for NR and EN-DC," V17.0.0, April 2022.

Network Energy Saving

Energy saving actions

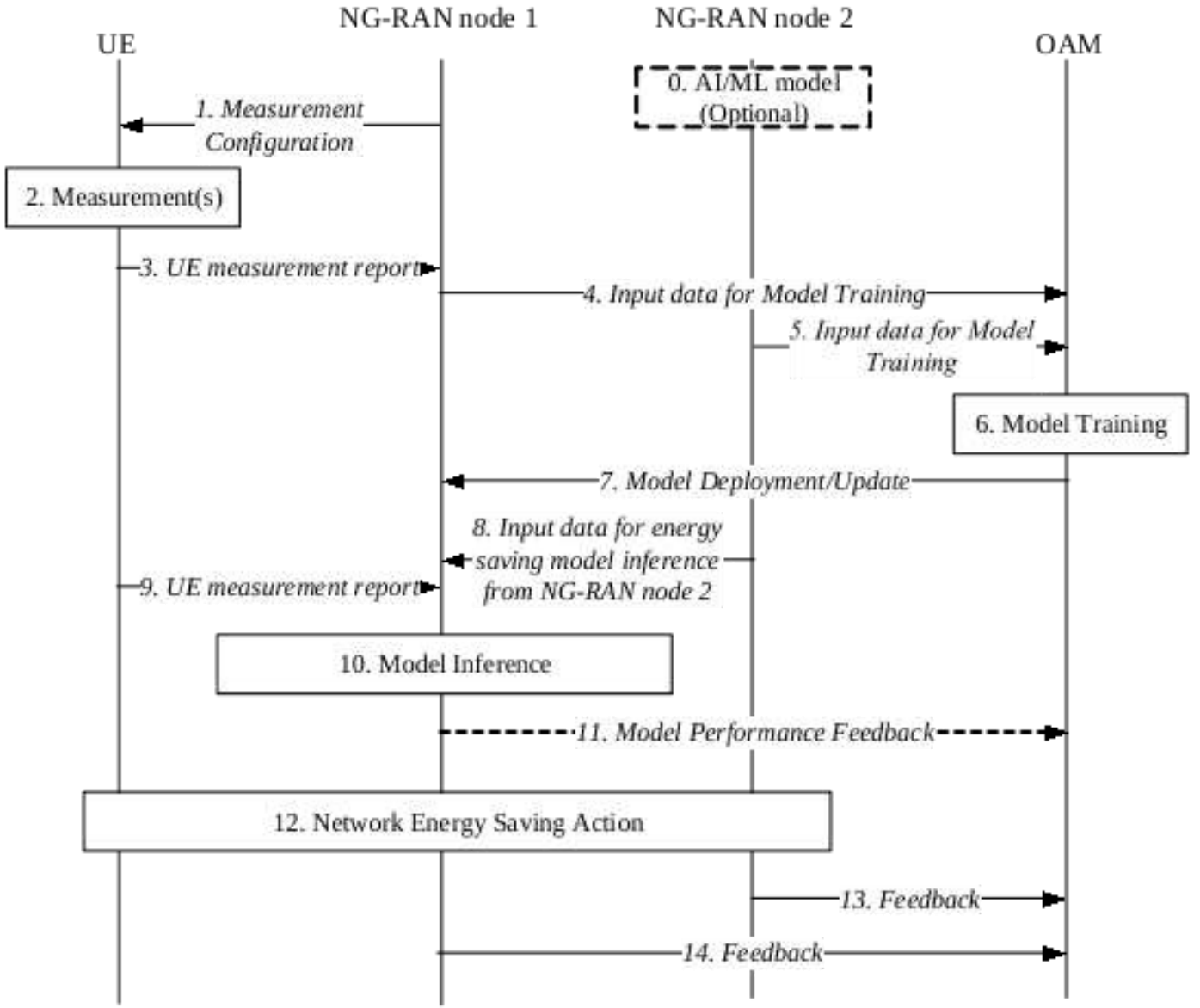
- Cell activation/deactivation
- Sector activation/deactivation
- Bandwidth reduction

Location of ML/AI model training and inference

- OAM (NWDAF/MDAF) or gNB

Example (Figure)

- Model training at OAM
- Model inference at gNB



3GPP **TR 37.817**, "Study on enhancement for data collection for NR and EN-DC," V17.0.0, April 2022.

Network Energy Saving

Model inputs

- **Local (gNB) input**: UE mobility/trajectory prediction, current/predicted energy efficiency, current/predicted resource usage status
- **UE input**: UE location information (coordinates, velocity serving Cell ID), UE measurements report (RSRP, RSRQ, SINR), including beam level information
- **Input from neighboring NG-RAN nodes**: current/predicted energy efficiency, resource usage, current energy state

Model outputs

- Energy saving strategy/recommendation, handover strategy, predicted energy strategy, predicted energy state

Model feedback

- Energy/resource status of neighboring nodes, UEs performance KPIs, gNB performance KPIs

---

3GPP **TR 37.817**, "Study on enhancement for data collection for NR and EN-DC," V17.0.0, April 2022.

Rel. 18 Work on ML/AI for NG-RAN

Follow Up Technical Study Item

3GPP **TR 38.843**, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," V18.0.0., January 2024.

Main Targets

- General Framework for Enhancing Air Interface Using ML/AI
- Main topics:
  - defining stages of AI/ML algorithms (model training, validation, testing, inference)
  - UE-gNB collaboration levels
  - required data sets for model training, validation and testing
  - model life-cycle management (LCM)

3GPP **TR 38.843**, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," V18.0.0., Jan 2024.

Rel. 18 Work on ML/AI for NR Air Interface

Three use cases

- **CSI Feedback:** Use ML/AI to reduce CSI overhead (e.g., spatial-frequency domain CSI compression), improve feedback accuracy, and enable prediction (e.g., time domain CSI prediction at UE)
- **Beam Management:** Use ML/AI to reduce beam management overhead and latency, and improve beam selection accuracy. Design methods for spatial-domain and time-domain downlink beam prediction.
- **Positioning:** Improve positioning accuracy for different scenarios including heavy NLOS conditions. Use either direct ML/AI approach (e.g., via fingerprinting) or AI/ML assisted approach (infer useful side-features)

3GPP **TR 38.843**, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," V18.0.0., Jan 2024.

Beam Management

BM-Case 1: Spatial-domain DL Beam Prediction of Set A of beams based on measurement results of Set B of beams

- AI/ML model training and inference at: 1) NW side, 2) UE side
- Set A and B relationships: 1) Sets A and B are different, 2) Set B is a subset of Set A
- Different measurement options: L1-RSRP, CIR (Channel impulse response), other info

BM-Case 2: Temporal-domain DL Beam Prediction of Set A of beams based on historical measurement results of Set B of beams

- AI/ML model training and inference at: 1) NW side, 2) UE side
- Set A and B relationships: 1) Sets A and B are different, 2) Set B is a subset of Set A, 3) Sets A and B are the same
- Similar to BM-Case 1 but using historical measurements (e.g., last L measurement reports)
- Prediction for F future instances (typically, F=1)

3GPP **TR 38.843**, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," V18.0.0., Jan 2024.

Collaboration Models between UE and gNB

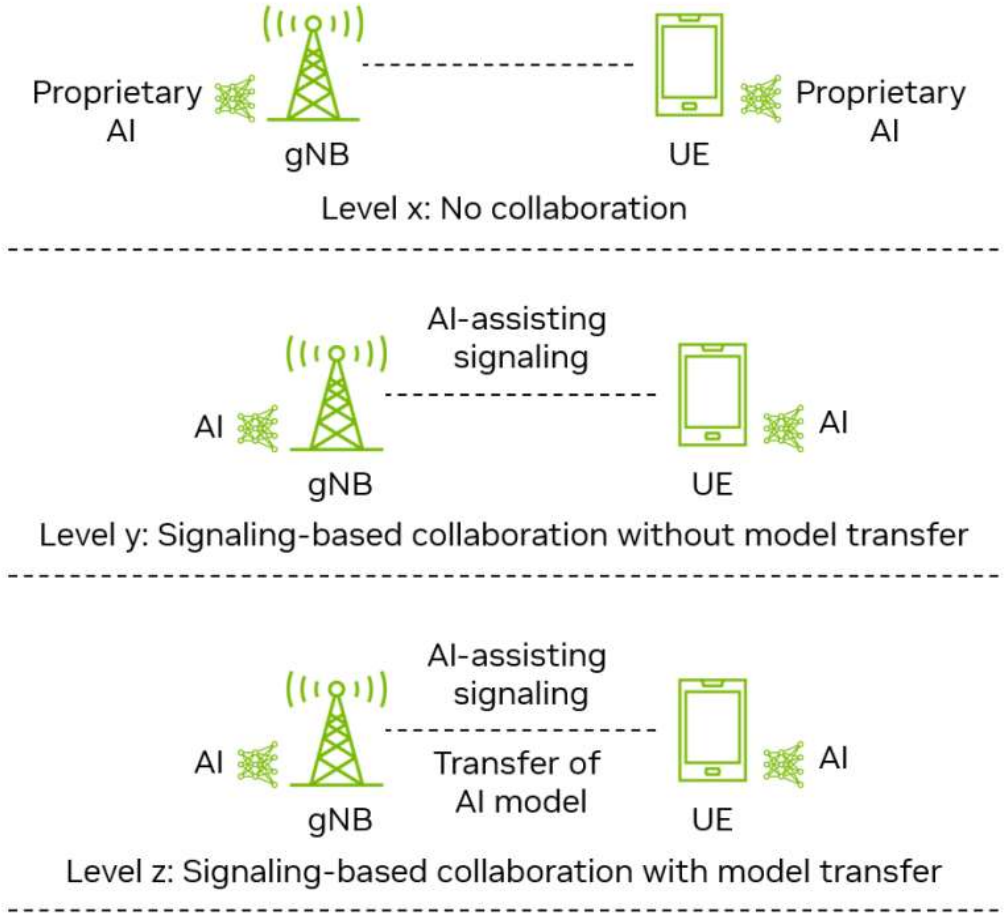
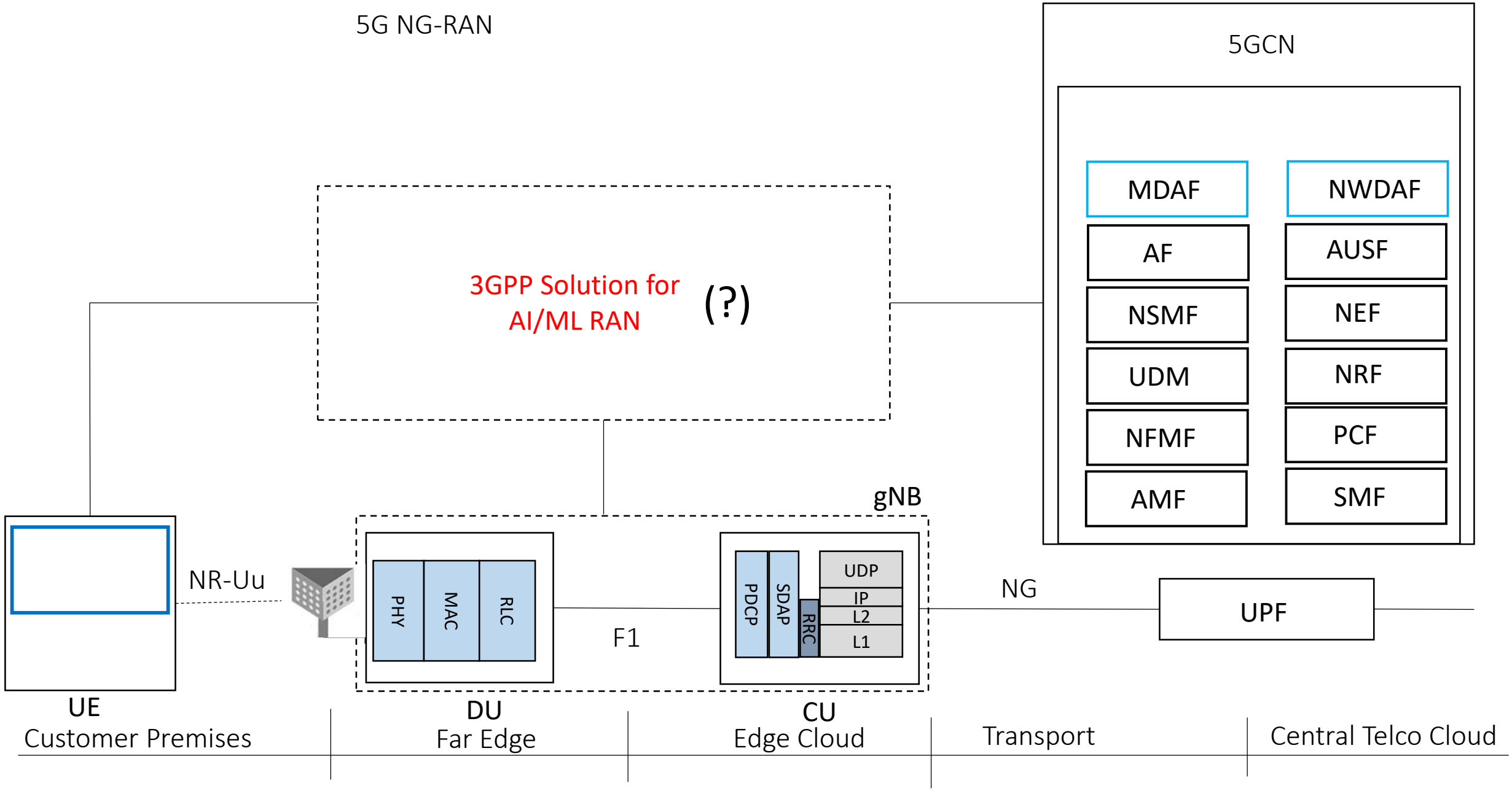


Figure: X. Lin, „Artificial Intelligence in 3GPP 5G-Advanced: A Survey," *arXiv preprint arXiv:2305.05092, also available at:* https://www.comsoc.org/publications/ctn/artificial-intelligence-3gpp-5g-advanced-survey.



3GPP **TR 38.843**, "Study on artificial intelligence (AI)/machine learning (ML) for NR air interface," V18.0.0., Jan 2024.

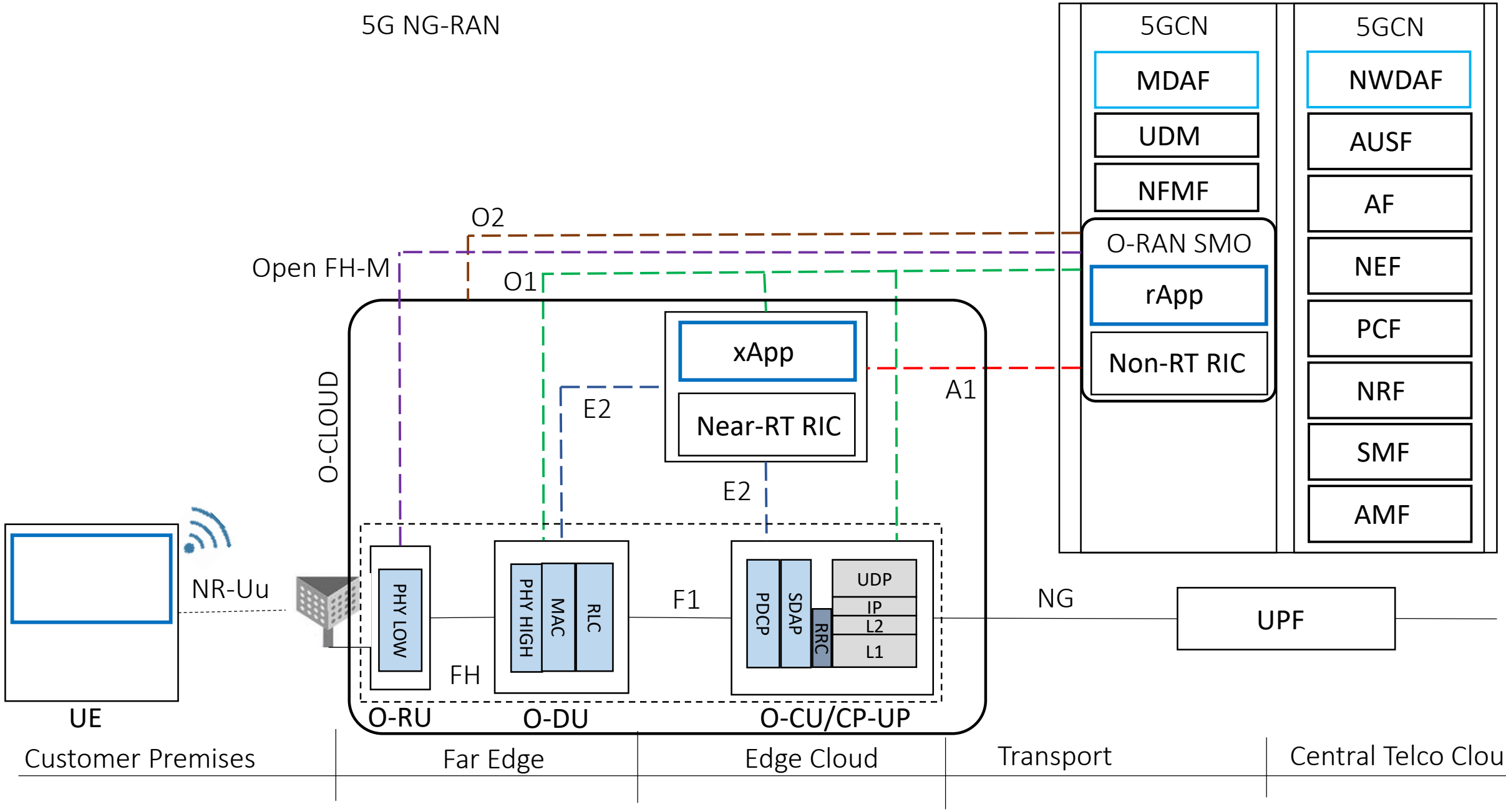
O-RAN ML/AI study for 5G RAN

O-RAN Alliance

- https://specifications.o-ran.org/specifications
- O-RAN.WG1.OAD-R003-v012.00: „O-RAN Architecture Description“, Technical Specification, 06/2024

O-RAN Architecture

- Disaggregation of NG-RAN into open, virtualised, interoperable and AI-driven architecture
- Augments 3GPP NG-RAN architecture and interfaces (F1, E1, Xn, NG) with open interfaces (A1, E2, O1, O2)
- Introduces two Radio Interface Controllers (RIC): non-real-time RIC (non-RT RIC) and near real-time RIC (near-RT RIC)



- O-RAN.WG1.OAD-R003-v12.00: „O-RAN Architecture Description“, Technical Specification, 06/2024

O-RAN Non-RT RIC

Service Management and Orchestration (SMO) Framework

- Responsible for RAN domain management
- **Non-RT RIC:** main SMO element that interfaces all other O-RAN NFs (O-NFs): near-RT RIC, O-CU-CP/UP, O-DU, O-RU implemented as VNF/PNF
- Fault, configuration, accounting, performance, and security (FCAPS)

Non-RT RIC Functions

- Controls Near-RT RIC via A1 through: 1) policy-based guidelines, 2) AI/ML model management, 3) enrichment information for Near-RT RIC
- Intelligent Non-RT (> 1s) RAN optimization control loops
- **rApps** (Non-RT RIC applications): AI/ML data-driven non-real time (> 1s) resource optimization – actions triggered via A1, O1, O2, Open FH-M

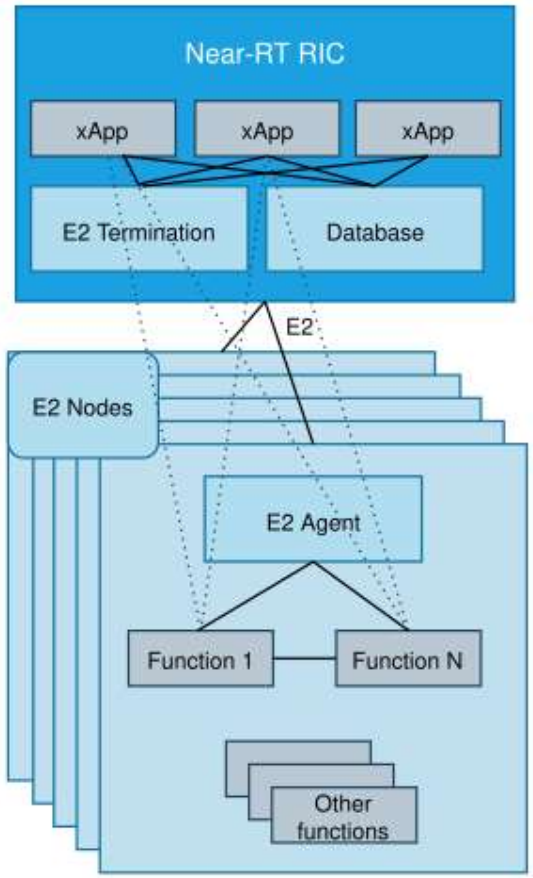
O-RAN Near-RT RIC

Near-RT RIC Functions

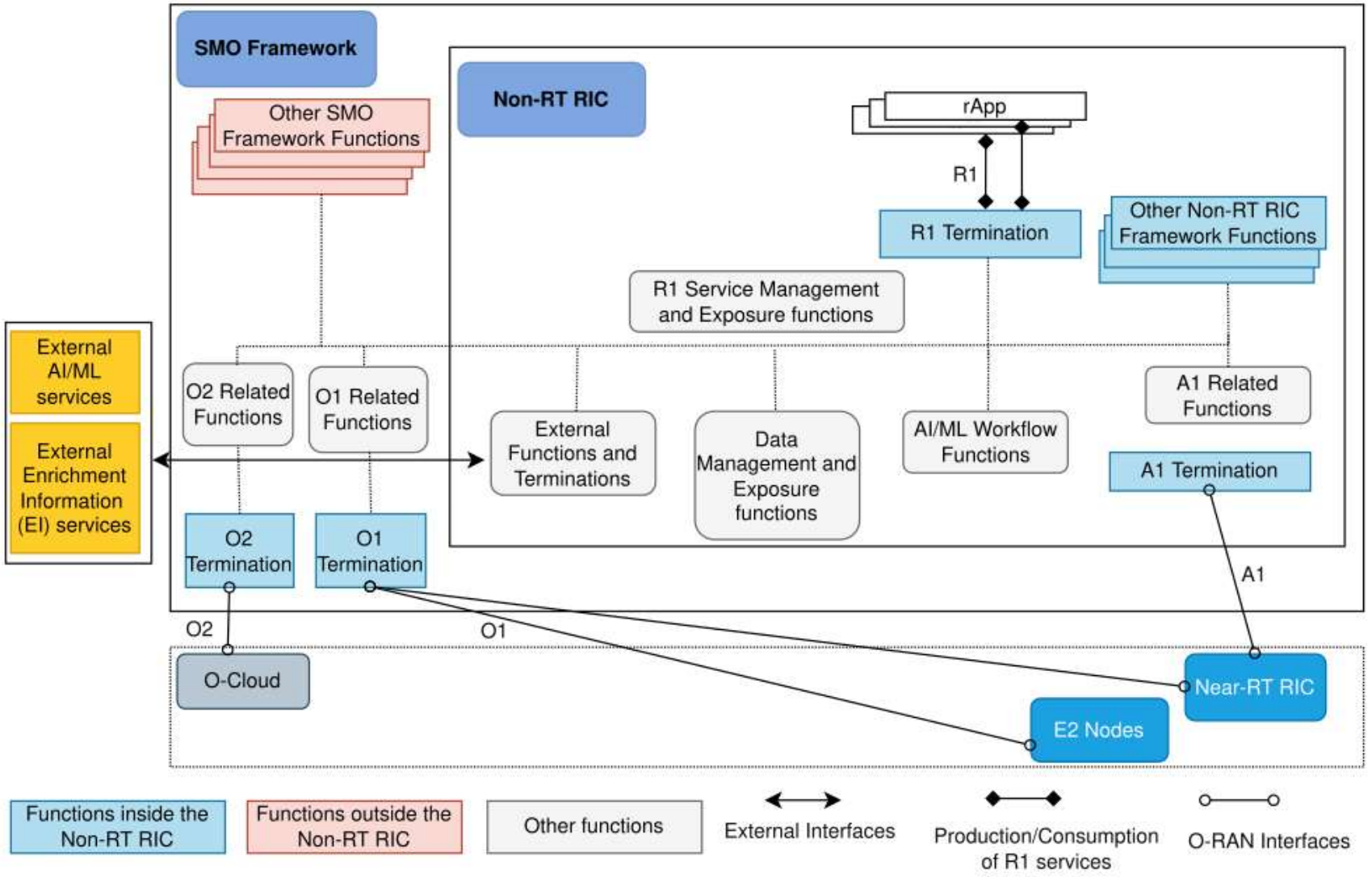
- Logical node placed between SMO and RAN nodes (O-CU/O-DU)
- Enables AI/ML data-driven near-real time (10ms – 1s) control and optimisation of RAN O-NFs (E2 nodes: O-CUs and O-DUs)
- Actions triggered via E2 interface
- **xApps** (Near-RT RIC applications): collect data in near-real time from E2 nodes and provide back value-added services to E2 nodes (O-CU, O-DU)
- In O-RAN, O-DU is split into O-DU (VNF) and O-RU (PNF) according to 7-2x lower layers split (one of the 3GPP split scenarios)

M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," IEEE Communications Surveys & Tutorials, 25(2), pp.1376-1411, 2023, https://arxiv.org/abs/2202.01032

O-RAN RICs



Near-RT RIC and RAN (E2) nodes.



Non-RT RIC generic architecture and components.

Figures: Marinova, S. and Leon-Garcia, A., Intelligent O-RAN beyond 5g: Architecture, use cases, challenges, and opportunities. *IEEE Access*, *12*, pp.27088-27114, 2024.

O-RAN Control Loops

Non-RT RIC Functions

- AI/ML data-driven non-real time (> 1s) control and optimisation of RAN using **rApps**

Near-RT RIC Functions

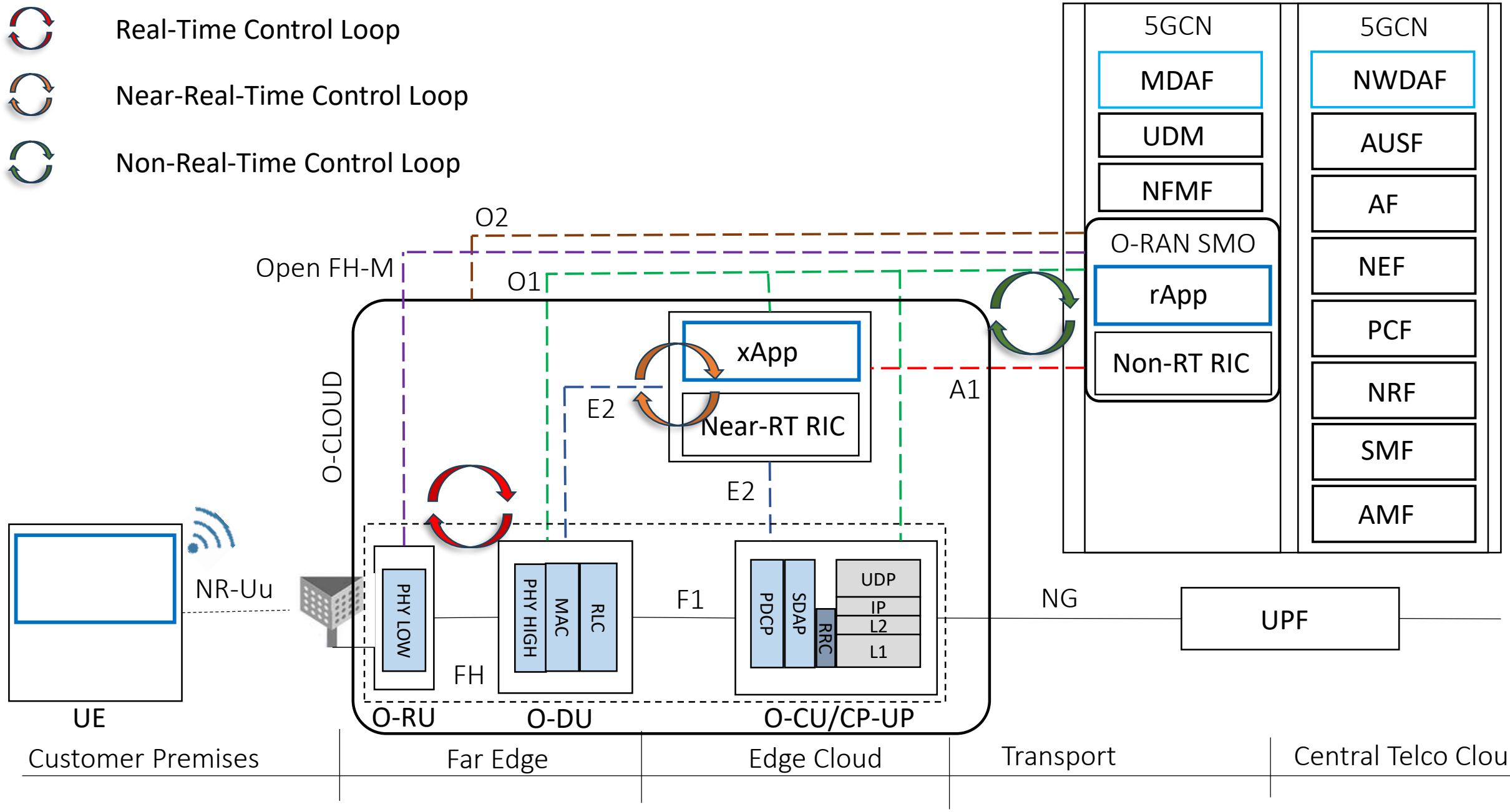
- AI/ML data-driven near-real time (10ms – 1s) control and optimisation of RAN O-NFs (O-CUs and O-DUs) using **xApps**

Real-Time RIC Functions (Work in Progress)

- AI/ML data-driven real time (< 10ms) control and optimisation of RAN O-NFs using **dApps**

M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "Understanding O-RAN: Architecture, interfaces, algorithms, security, and research challenges," IEEE Communications Surveys & Tutorials, 25(2), pp.1376-1411, 2023, https://arxiv.org/abs/2202.01032

Lacava, A., Bonati, L., Mohamadi, N., Gangula, R., Kaltenberger, F., Johari, P., D'Oro, S., Cuomo, F., Polese, M. and Melodia, T., 2025. dApps: Enabling Real-Time AI-Based Open RAN Control. *arXiv preprint arXiv:2501.16502,* https://arxiv.org/pdf/2501.16502

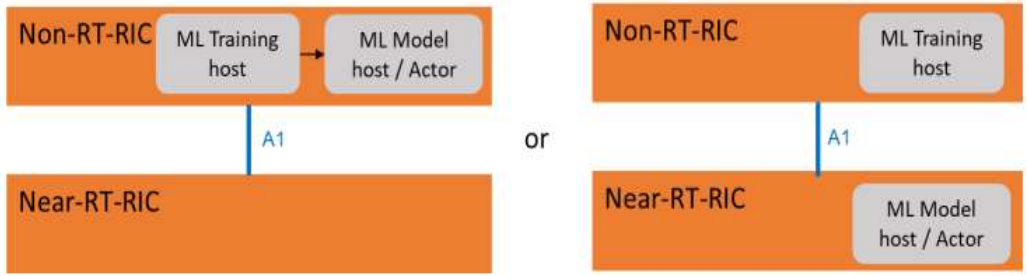


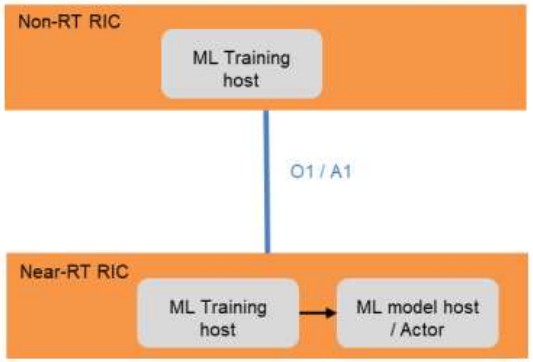
- O-RAN.WG1.OAD-R003-v12.00: „O-RAN Architecture Description“, Technical Specification, 06/2024

ML/AI Workflows in O-RAN

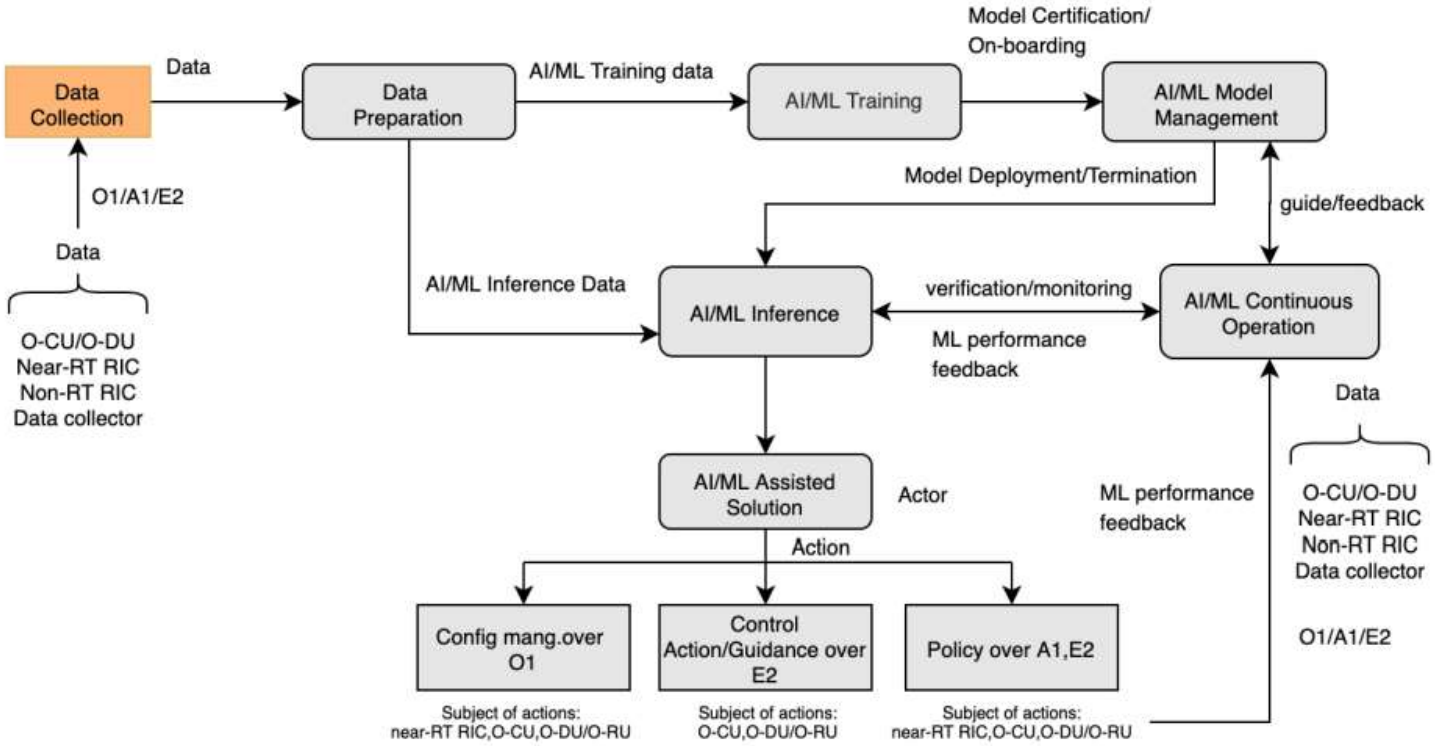
Training: Usually at non-RT RIC, exception is FL

Inference: Either non-RT RIC or near-RT RIC (if delay limited)



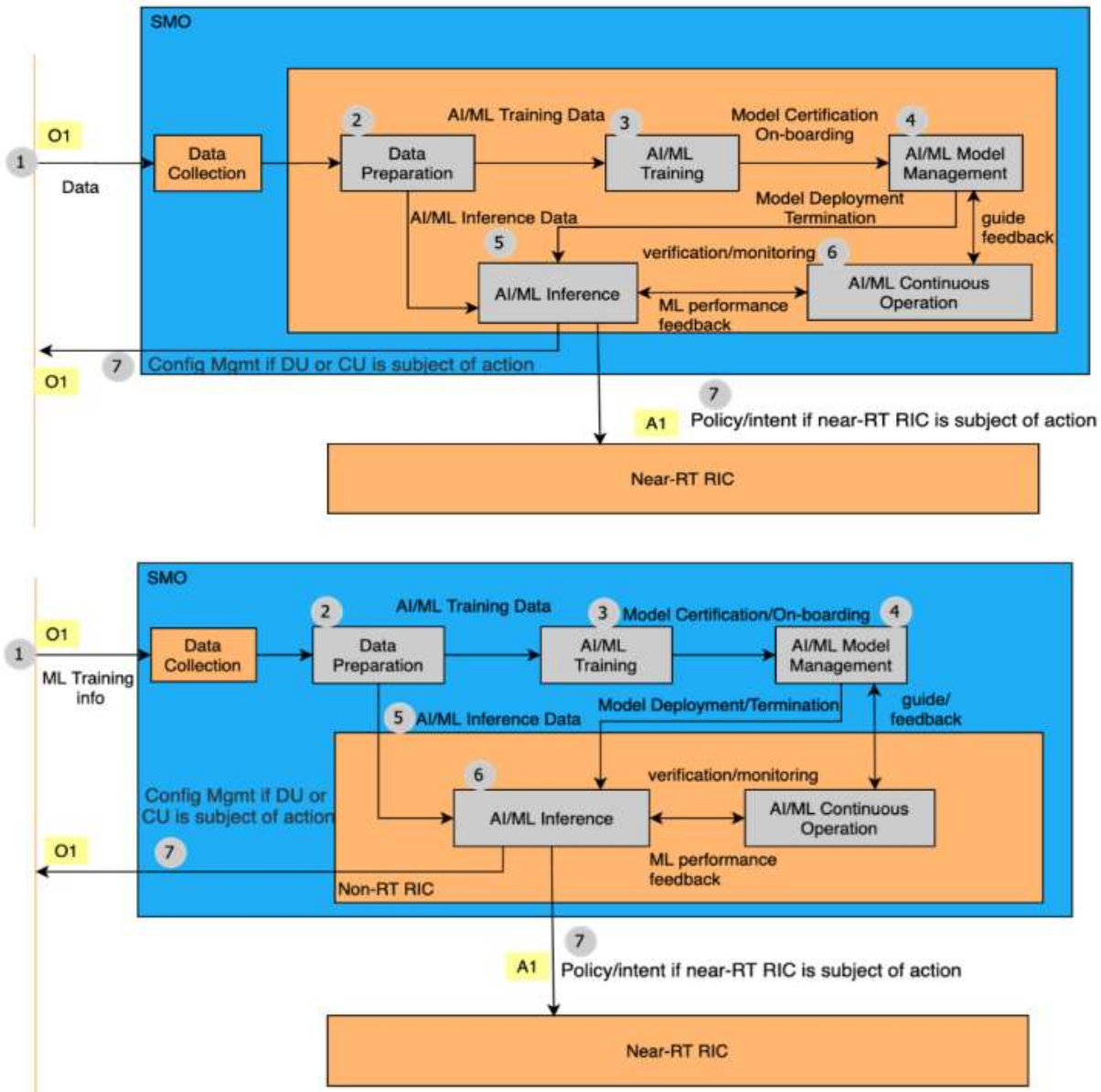


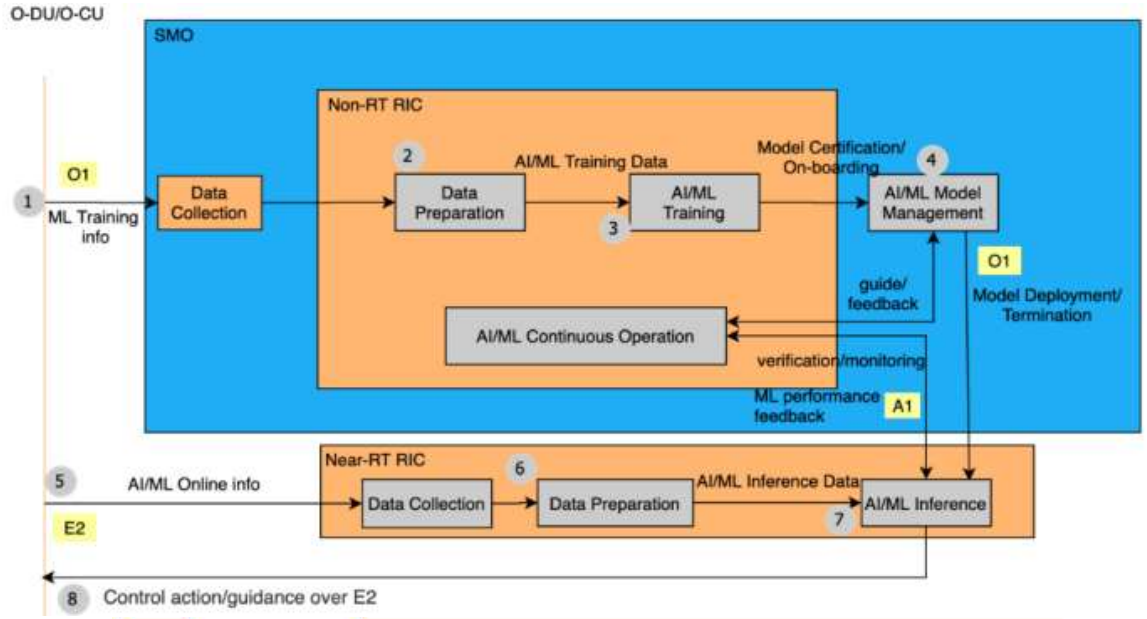
AI/ML General Lifecycle Procedure O-RAN

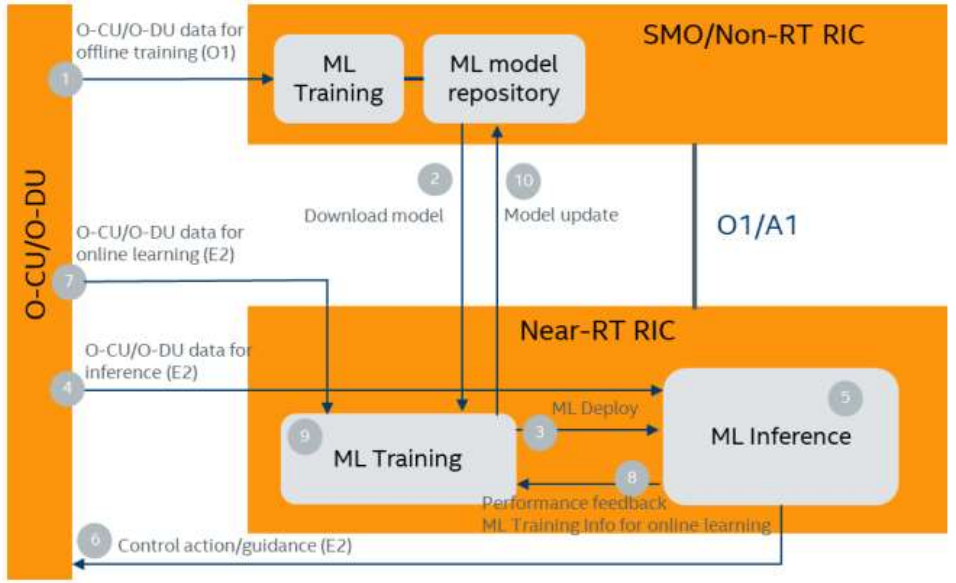


Figures: O-RAN.WG2.AIML-v01.03: O-RAN Working Group 2 AI/ML workflow description and requirements

ML/AI O-RAN Deployment Options







Figures: O-RAN.WG2.AIML-v01.03: O-RAN Working Group 2 AI/ML workflow description and requirements

ML/AI in 5G NG-RAN: A Sample of Research

O-RAN xApps and rApps implementation

- Kouchaki, M. and Marojevic, V., Actor-Critic Network for O-RAN Resource Allocation: xApp Design, Deployment, and Analysis, *IEEE Globecom Workshops*, pp. 968-973, 2022.
- Vilà, I., Sallent, O. and Pérez-Romero, J., On the Implementation of a Reinforcement Learning-based Capacity Sharing Algorithm in O-RAN, *IEEE Globecom Workshops*, pp. 208-214, 2022.

O-RAN Open Source Testbed Implementation

- Open AI Cellular (https://github.com/openaicellular) - Upadhyaya, P.S., Abdalla, A.S., Marojevic, V., Reed, J.H. and Shah, V.K., Prototyping next-generation O-RAN research testbeds with SDRs. *arXiv:2205.13178*.
- OpenRAN Gym (https://openrangym.com/) - M. Polese, L. Bonati, S. D'Oro, S. Basagni, and T. Melodia, "ColO-RAN: Developing Machine Learning-based xApps for Open RAN Closed-loop Control on Programmable Experimental Platforms," IEEE Trans. Mobile Computing, July 2022.
- Both OAI and srsRAN are developing their O-RAN compliant software

OAI: https://openairinterface.org/oai-5g-ran-project/

srsRAN:https://docs.srsran.com/projects/project/en/latest/knowledge_base/source/oran_gnb/source/index.html

AI-RAN Alliance

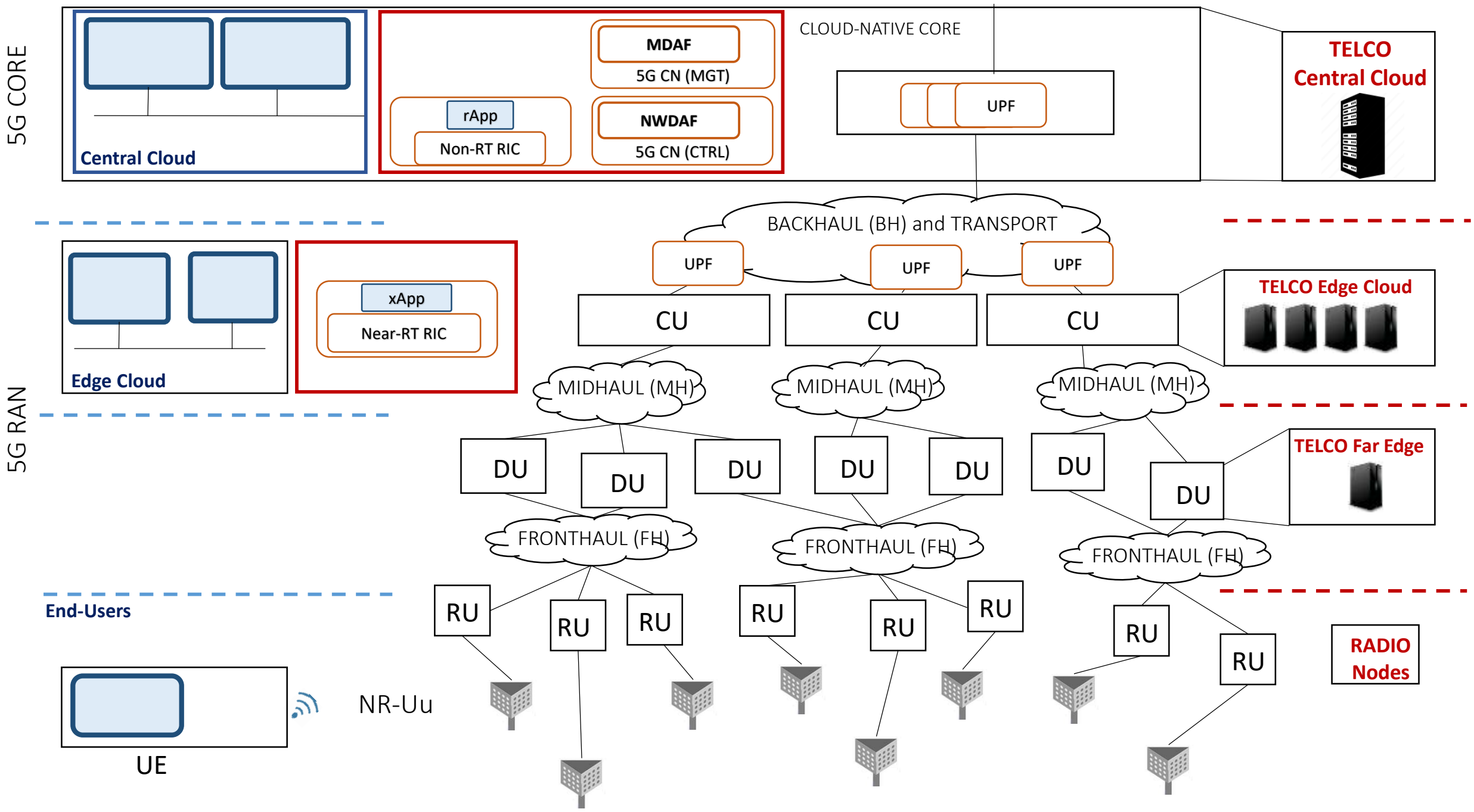
AI-RAN Alliance

"Bringing together the technology industry leaders and academic institutions, the AI-RAN Alliance is dedicated to driving the enhancement of RAN performance and capability with AI."

AI RAN Whitepaper: https://ai-ran.org/wp-content/uploads/2024/12/AI-RAN_Alliance_Whitepaper.pdf

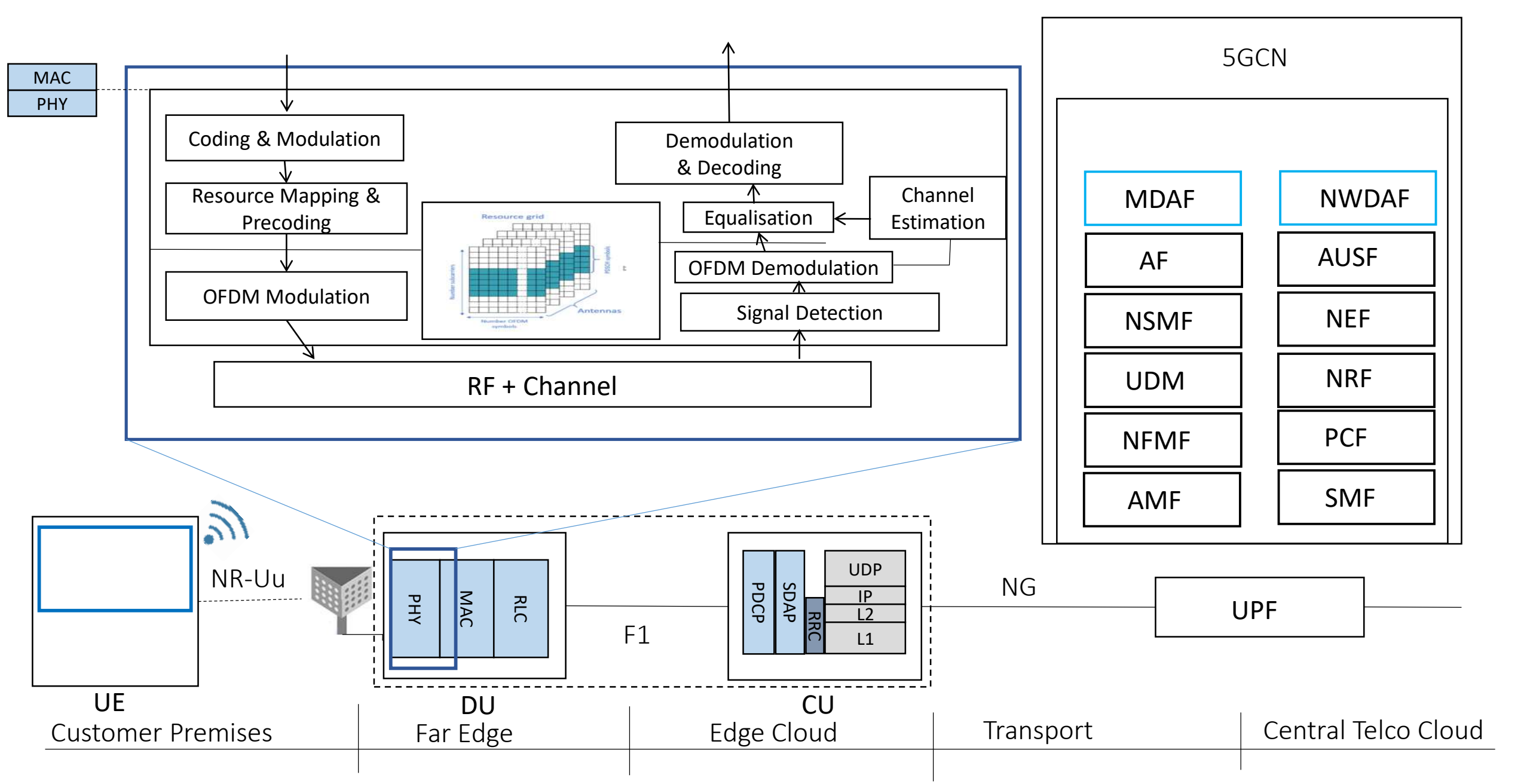
Organised in three Working Groups

- **AI for RAN** focusing on the integration of AI into RAN to significantly enhance RAN performance, such as improving spectral and operational efficiency, optimizing radio resource management, and enabling predictive maintenance.
- **AI and RAN** seeks to create a shared infrastructure between RAN and AI workloads, enabling concurrent resource utilization on this converged computer-and-communications infrastructure.
- **AI on RAN** aims to enable new RAN services to enhance AI applications running at the network edge, to be able to offer new consumer and enterprise services and applications from the edge of the network.


5G CORE
Central Cloud
MDAF
5G CN (MGT)
rApp
Non-RT RIC
NWDAF
5G CN (CTRL)
CLOUD-NATIVE CORE
UPF
TELCO
Central Cloud
5G RAN
Edge Cloud
xApp
Near-RT RIC
BACKHAUL (BH) and TRANSPORT
UPF
UPF
UPF
TELCO Edge Cloud
CU
CU
CU
MIDHAUL (MH)
MIDHAUL (MH)
MIDHAUL (MH)
DU
DU
DU
DU
DU
DU
DU
TELCO Far Edge
FRONTHAUL (FH)
FRONTHAUL (FH)
FRONTHAUL (FH)
RU
RU
RU
RU
RU
RU
RU
RU
RADIO
Nodes
End-Users
NR-Uu
UE

Outline of the talk

- AI/ML in the Core Network
- AI/ML in the RAN
- **AI/ML for the PHY**
- AI/ML at the Application Layer
- AI/ML-driven Network Management and Orchestration


MAC
PHY
Coding & Modulation
Resource Mapping & Precoding
OFDM Modulation
RF + Channel
Demodulation & Decoding
Channel Estimation
Equalisation
OFDM Demodulation
Signal Detection
Resource grid
Antennas
5GCN
MDAF
NWDAF
AF
AUSF
NSMF
NEF
UDM
NRF
NFMF
PCF
AMF
SMF
NR-Uu
PHY
MAC
RLC
F1
PDCP
SDAP
RRC
UDP
IP
L2
L1
NG
UPF
UE
DU
CU
Customer Premises
Far Edge
Edge Cloud
Transport
Central Telco Cloud

rApp

Non-RT RIC

xApp

Near-RT RIC

MAC

PHY

E2

O1

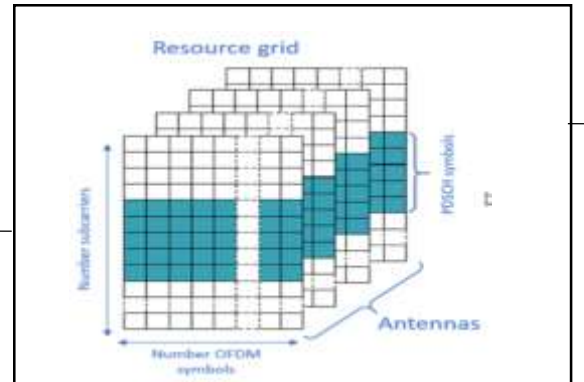
MODEL TRAINING, VALIDATION, DEPLOYMENT, LIFE-CYCLE MANAGEMENT

- Data collection, model training, model inference
- How to communicate trained models?
- Model life-cycle management
- Hardware and ML architectures (hardware, computation, energy costs)?
- Influence of functional split (3GPP, O-RAN)

ML/AI-BASED CODING/MODULATION

ML/AI-BASED DECODING/DEMODULATION

ML/AI BASED PRECODING/BEAMFORMING



EQUALISATION

ML/AI-BASED CHANNEL ESTIMATION

OFDM MODULATION

OFDM DEMODULATION

ML/AI-BASED SIGNAL DETECTION, TIMING/CFO ESTIMATION

RF

ML/AI-based Signal Detection

ML/AI-BASED SIGNAL DETECTION, TIMING/CFO ESTIMATION

Deep Learning-based Timing Offset and Carrier Frequency Offset Estimation

- First batch of baseband signal processing steps (before channel estimation)

O'Shea, T., Karra, K., and Clancy, T.C., "Learning approximate neural estimators for wireless channel state information", *Proc. IEEE 27th Int. Workshop Mach. Learn. Signal Process. (MLSP)*, pp. 1-7, Sep. 2017.

- Extensions to I/Q imbalance, phase noise and power amplifier nonlinearities

Pihlajasalo, J., Korpi, D., Riihonen, T., Talvitie, J., Uusitalo, M.A. and Valkama, M., Detection of Impaired OFDM Waveforms Using Deep Learning Receiver, *IEEE SPAWC 2022,* pp. 1-5, 2022.

- DL-based timing offset and CFO estimation (not 3GPP but IEEE 802.11 OFDM Rx)

Ninkovic, V., Valka, A., Dumic, D. and Vukobratovic, D., Deep learning-based packet detection and carrier frequency offset estimation in IEEE 802.11 ah, *IEEE Access*, *9*, pp. 99853-99865, 2021.

ML/AI-based Channel Estimation

ML/AI-BASED CHANNEL ESTIMATION

Deep Learning-based Channel Estimation

- DNN trained to map received pilots and data symbols directly into equalised data symbols (without intermediate channel estimation)

Ye, H., Li, G.Y. and Juang, B.H., Power of deep learning for channel estimation and signal detection in OFDM systems, *IEEE Wireless Communications Letters*, *7*(1), pp. 114-117, 2017.

- Channel estimation as CNN-based 2D image recovery of channel response grid

Soltani, M., Pourahmadi, V., Mirzaei, A. and Sheikhzadeh, H., Deep learning-based channel estimation. *IEEE Communications Letters*, *23*(4), pp. 652-655, 2019.

- Extensions to mmWave, massive MIMO, RIS, Thz channels, etc.
- Recommended paper on DL-based Channel Estimation Interpretability

Hu, Q., Gao, F., Zhang, H., Jin, S. and Li, G.Y., Deep learning for channel estimation: Interpretation, performance, and comparison. *IEEE Transactions on Wireless Communications*, *20*(4), pp. 2398-2412, 2020.

- Recommended paper on implementation aspects

Haq, S.A.U., Gizzini, A.K., Shrey, S., Darak, S.J., Saurabh, S. and Chafii, M., Deep Neural Network Augmented Wireless Channel Estimation for Preamble-Based OFDM PHY on Zynq System on Chip, *IEEE Transactions on Very Large Scale Integration (VLSI) Systems*, 2023.

ML/AI-based Coding/Modulation

ML/AI-BASED CODING/MODULATION

End-to-End Autoencoder-Based Deep Learning

- End-to-end learning of Communication Systems using Deep Autoencoders

O'Shea, T. and Hoydis, J., An introduction to deep learning for the physical layer. *IEEE Transactions on Cognitive Communications and Networking*, *3*(4), pp. 563-575, 2017.

- Deep Autoencoder-based PHY design in OFDM systems

Dörner, S., Cammerer, S., Hoydis, J. and Ten Brink, S., Deep learning based communication over the air. *IEEE Journal of Selected Topics in Signal Processing*, *12*(1), pp. 132-143, 2017.

- Autoencoder-based UEP codes and Rateless codes

Ninkovic, V., Vukobratovic, D., Häger, C., Wymeersch, H. and i Amat, A.G., Autoencoder-Based Unequal Error Protection Codes. *IEEE Communications Letters*, *25*(11), pp. 3575-3579, 2021.

Ninkovic, V., Vukobratovic, D., Häger, C. and Wymeersch, H., Rateless Autoencoder Codes: Trading off Decoding Delay and Reliability, *IEEE ICC 2023, Rome, Italy, May 2023*.

- Open source libraries for end-to-end learning

https://developer.nvidia.com/sionna
https://developer.nvidia.com/aerial-sdk

ML/AI-based Coding/Decoding

ML/AI-BASED CODING/MODULATION

Neural-Enhanced Belief Propagation Decoding

Neural-network inspired Belief Propagation decoder

- Nachmani, E., Marciano, E., Lugosch, L., Gross, W.J., Burshtein, D. and Be'ery, Y., Deep learning methods for improved decoding of linear codes. *IEEE Journal of Selected Topics in Signal Processing*, *12*(1), pp.119-131, 2018.

Principled Approach for Combining Data-Based and Model-Based Methods

- Shlezinger, N., Whang, J., Eldar, Y.C. and Dimakis, A.G., 2023. Model-based deep learning. Proceedings of the IEEE, 111(5), pp.465-499.

Decoding Using Graph-Neural Networks

Decoding LDPC codes using GNNs derived from code factor graph

- Cammerer, S., Hoydis, J., Aoudia, F.A. and Keller, A., Graph neural networks for channel decoding. In *2022 IEEE Globecom Workshops (GC Wkshps)* (pp. 486-491), 2022.
- Ninkovic, V., Kundacina, O., Vukobratovic, D., Häger, C., A. Graell i Amat, Decoding Quantum LDPC Codes Using Graph Neural Networks. arXiv preprint arXiv:2408.05170, to appear, IEEE GLOBECOM 2024.

Decoding Using Transformers

Transformer-based approach to decoding LDPC codes

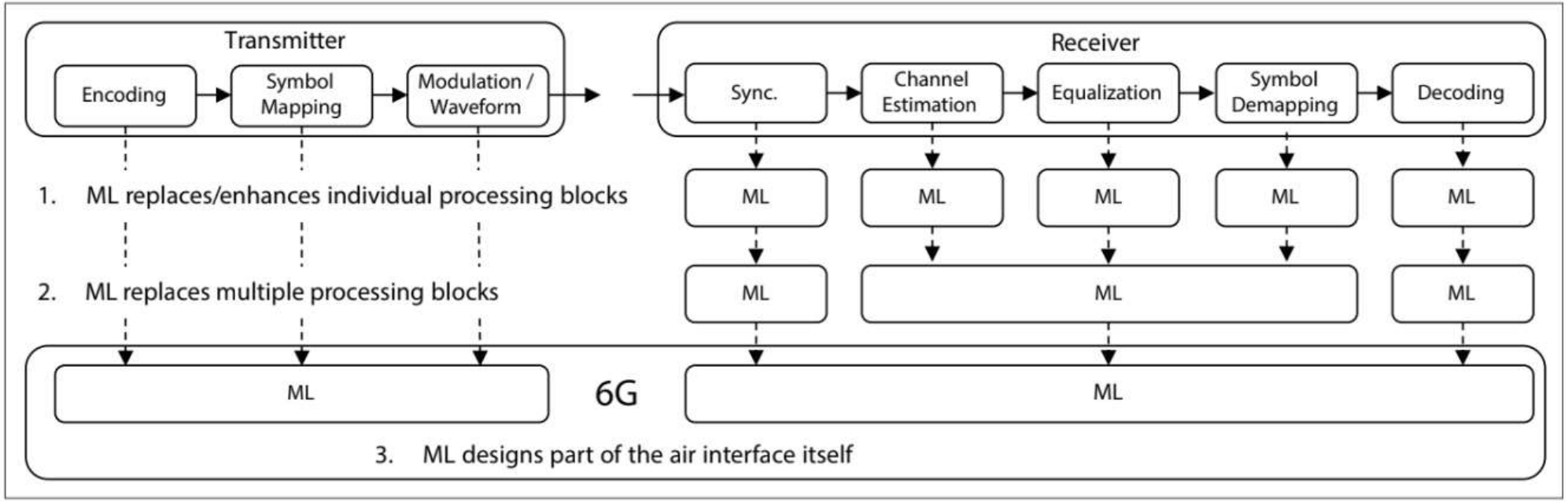
- Choukroun, Y. and Wolf, L., 2024. Learning Linear Block Error Correction Codes. arXiv preprint arXiv:2405.04050.

AI-Based PHY Design

ML/AI-BASED CODING/MODULATION

Deep Learning for the Physical Layer

Hoydis, J., Aoudia, F.A., Valcarce, A. and Viswanathan, H., Toward a 6G AI-native air interface, *IEEE Communications Magazine*, *59*(5), pp. 76-81, 2021.



Question: Which PHY blocks to replace with AI/ML?

ML/AI-based Beamforming

ML/AI BASED PRECODING/BEAMFORMING

ML for Beam Alignment

Ma, W., Qi, C. and Li, G.Y., Machine learning for beam alignment in millimeter wave massive MIMO. IEEE Wireless Communications Letters, 9(6), pp. 875-878, 2020.

ML-based Initial Beam Alignment

Sohrabi, F., Chen, Z. and Yu, W., Deep active learning approach to adaptive beamforming for mmWave initial alignment. *IEEE Journal on Selected Areas in Communications*, *39*(8), pp. 2347-2360, 2021.

Joint Learning of Beams and Alignments

Heng, Y., Mo, J. and Andrews, J.G., Learning site-specific probing beams for fast mmWave beam alignment, *IEEE Transactions on Wireless Communications*, *21*(8), pp. 5785-5800, 2022.

Overview of AI/ML Beamforming Methods

Al Kassir, H., Zaharis, Z.D., Lazaridis, P.I., Kantartzis, N.V., Yioultsis, T.V. and Xenos, T.D., A Review of the State of the Art and Future Challenges of Deep Learning-Based Beamforming. *IEEE Access,* 2022.

Beam Management

Khan, M.Q., Gaber, A., Schulz, P. and Fettweis, G., Machine Learning for Millimeter Wave and Terahertz Beam Management: A Survey and Open Challenges, *IEEE Access*, *11*, pp. 11880-11902, 2023.

Outline of the talk

- AI/ML in the Core Network
- AI/ML in the RAN
- AI/ML for the PHY
- **AI/ML at the Application Layer**
- AI/ML-driven Network Management and Orchestration

AI/ML model transfer in 5G

Use Cases and Requirements for Model Transfer

- **AI/ML operation splitting (split learning)**: keep privacy- or latency-sensitive parts in UE, offload computation- or energy-intensive parts
- **AI/ML model data distribution**: adaptive model downloading when needed (efficient unicast/multicast)
- **Distributed/Federated learning over 5G**: UEs perform local training while a central entity trains a global model by aggregating local models

TR 22.874 is extended in Rel. 18/19 to TS 22.261 (KPIs for model transfer)

- **Example:** maximum allowed downlink end-to-end latency is 1 s and experienced downlink data rate required is 1.1 Gbps for image recognition related AI/ML model distribution

---

- 3GPP **TR22.874**, “Study on traffic characteristics and performance requirements for AI/ML model transfer,” V18.2.0., December 2021.
- 3GPP **TS22.261**, “Service requirements for the 5G system,” V19.7.0, June 2024.

AI/ML model transfer in 5G

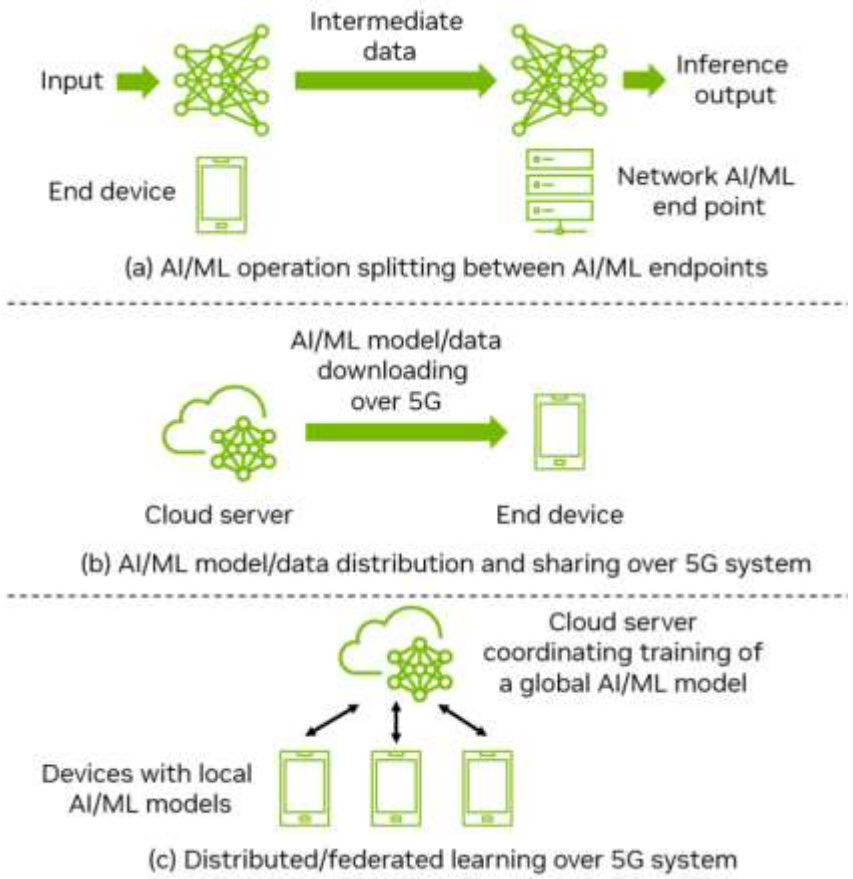


Figure: X. Lin, „Artificial Intelligence in 3GPP 5G-Advanced: A Survey,“ *arXiv preprint arXiv:2305.05092*.

- 3GPP **TS 22.261**, “Service requirements for the 5G system,” V18.6.0, March 2022.
- V. Ninkovic, D. Miskovic, M. Zennaro, D. Vukobratovic: "COMSPLIT: A Communication-Aware SPLIT Learning for Heterogeneous IoT Platforms," IEEE Internet of Things Journal, 2024.

5G System Support for AI/ML services

How to operate 5G system more intelligently for AI/ML AL services?

5G Application Function (AF) and Network Exposure Function (NEF)

- AF controls the logic of the application layer AI/ML operation
- AF interacts with 5GC to provide services to support, for example, application influence on traffic routing
- How to monitor network resource utilization related to the UE's performance such as data rate and latency?
- NEF in 5GC supports monitoring capability which allows configuration, detection, and reporting of monitoring events to authorized external party

---

- X. Lin, „Artificial Intelligence in 3GPP 5G-Advanced: A Survey," *arXiv preprint arXiv:2305.05092*.
- 3GPP **TR 23.700-80**, "Study on 5G system support for AI/ML-based services," V18.0.0, December 2022.

Outline of the talk

- AI/ML in the Core Network
- AI/ML in the RAN
- AI/ML for the PHY
- AI/ML at the Application Layer
- **AI/ML-driven Network Management and Orchestration**

AI/ML-Based Network Management

- Recent trends in GenAI models have a potential to advance long-term goals of autonomous service/network management

Two main ingredients

- Service-network interaction through **Intent-Based Networking**
  - **Intents** are in human-readable formats and suitable for LLM processing
- Network/Service management and orchestration using **AI agents**
  - Distributed set of AI agents (LLMs or others) that automate NMO

Chen, Z., Sun, Q., Li, N., Li, X., Wang, Y. and Chih-Lin, I., Enabling mobile AI agent in 6G era: Architecture and key technologies, *IEEE Network, 2024.*

AI Agents in 6G

AI agent as 6G network controller

Intent-based network management

- Translate high-level intents from operators to specific configurations and actions deployed on infrastructure

Automated network optimization

- Monitor and analyze network performance in real time, identify issues and perform optimization

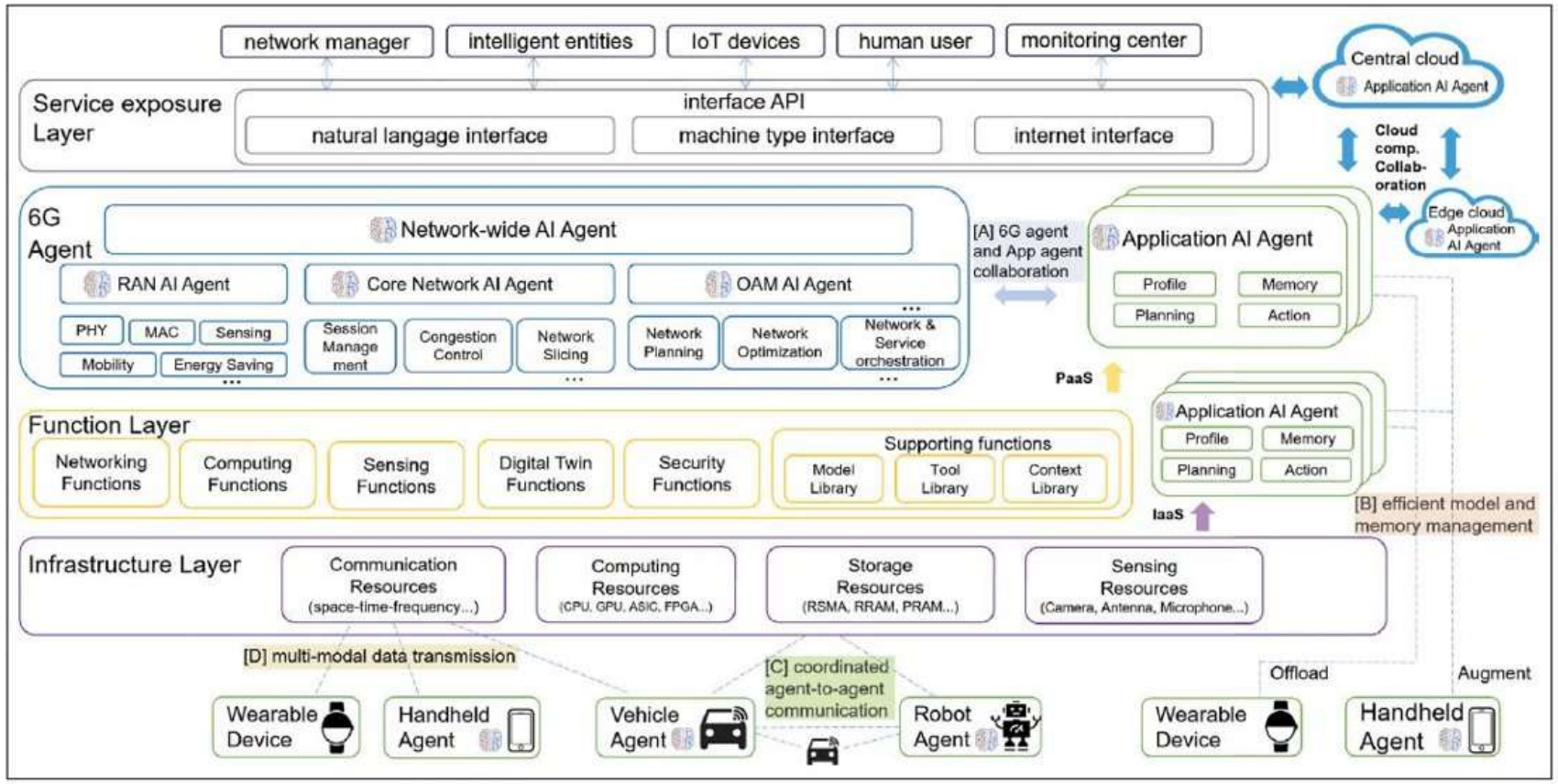
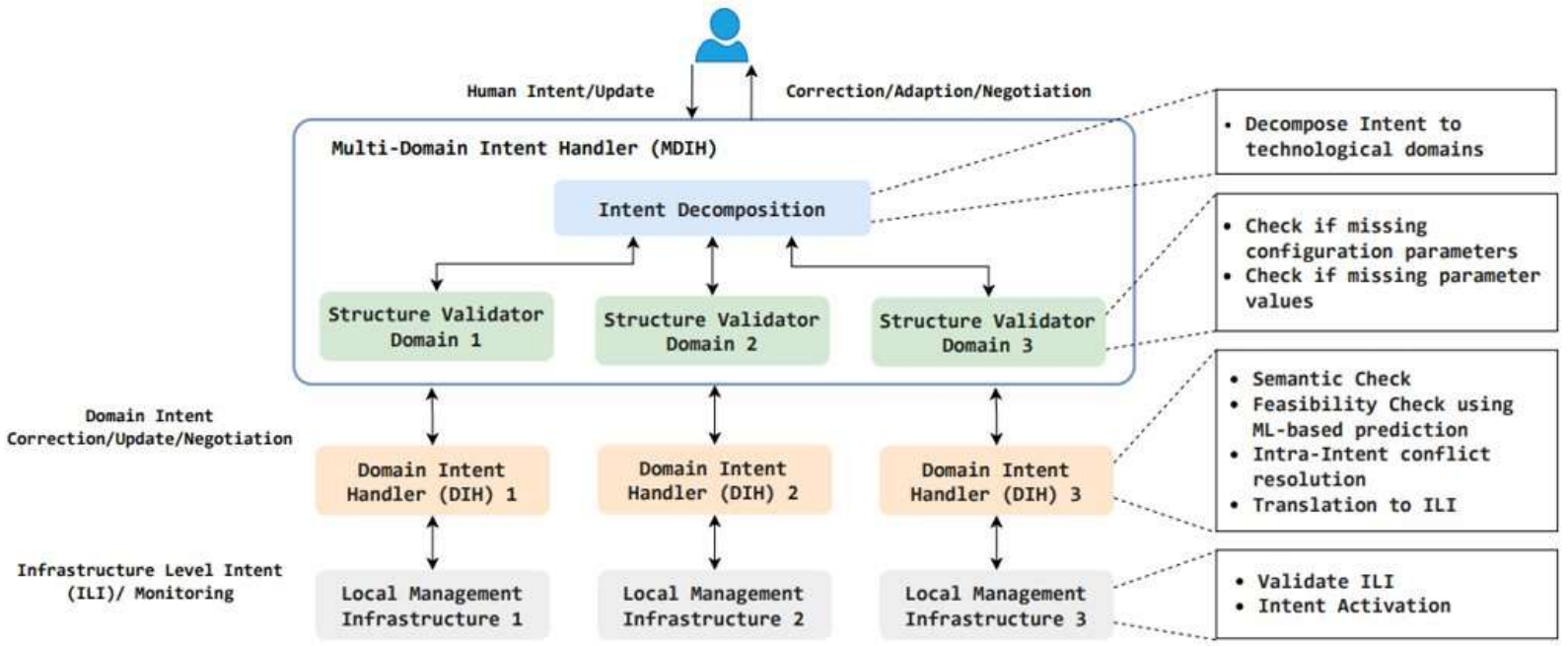


Figure: Chen, Z., Sun, Q., Li, N., Li, X., Wang, Y. and Chih-Lin, I., Enabling mobile AI agent in 6G era: Architecture and key technologies, *IEEE Network, 2024.*

AI Agents in 6G

AI agents (LLMs)

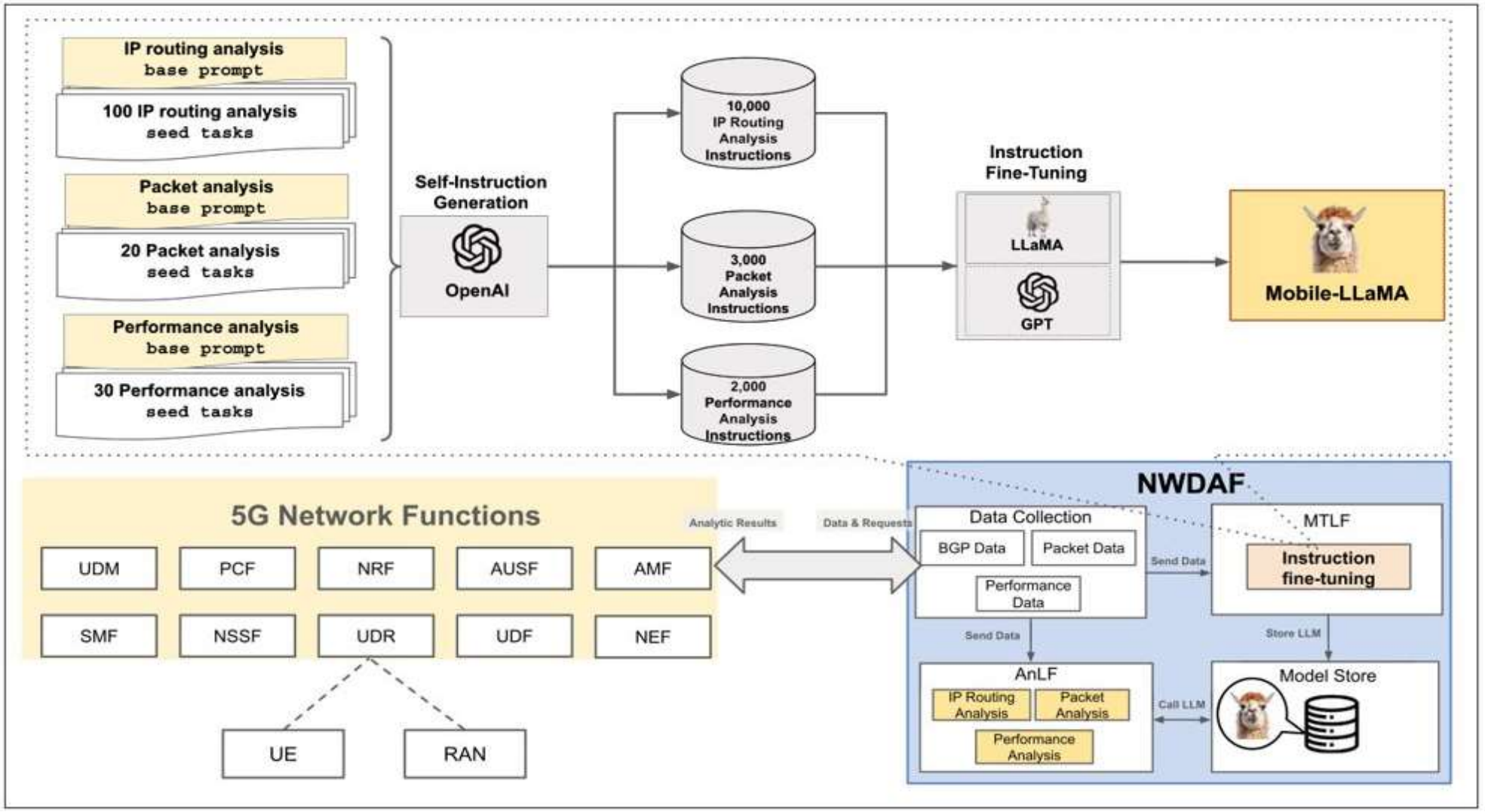
- Hierarchical levels of intent decomposition and translation
- High-level intents decomposed and translated for domain levels (RAN, CN, transport)
- Local intent processing and translation to language (commands) used by network controlers/orchestrators



1: High-level architecture design to handle natural language-based Intents LC.

Figure: Mekrache, A., Ksentini, A. and Verikoukis, C., 2024. Intent-based management of next-generation networks: An LLM-centric approach, IEEE Network, 2024.

Example: AI Agent (LLM) as part of NWDAF



Kan, K.B., Mun, H., Cao, G. and Lee, Y., Mobile-llama: Instruction fine-tuning open-source llm for network analysis in 5g networks, IEEE Network, 2024.

Thank You! Questions?