Network Architecture for Machine Learning: A Network Operator's Perspective

Chetana V. Murudkar, Kwang-Cheng Chen, and Richard D. Gitlin

A conceptual framework for RNA configuration and management enabling a broader perspective toward an ML-driven hybrid self-organizing network is discussed to improve the signaling load to attain reduced latency and improved network capacity.

Abstract

Data-driven network design suggests that substantial technology advances of 5G and 6G networks will be enabled with enhanced automation, intelligence, and user-experience-focused capabilities. Network operators need to upgrade the standard network models by applying machine learning (ML) to address the complexities of next-generation network deployments. This article explores the role of ML and its interplay with wireless communications networks to develop the next-generation network architecture. A use case scenario for self-configuration of radio-access-network-based notification areas (RNAs) for effective resource management is analyzed to exemplify the proposed architecture where a paging load reduction of 64 percent is observed in the resulting RNA clusters. A conceptual framework for RNA configuration and management enabling a broader perspective toward an ML-driven hybrid self-organizing network is discussed to improve the signaling load to attain reduced latency and improved network capacity.

Introduction

The transition into the digital society gives rise to a diverse range of applications and services that demand a radically new communication network architectural design for accommodating artificial intelligence (AI) into 5G and 6G technologies. As 5G and beyond networks get deployed, there is a need to extend the scope of network functionalities by integrating new capabilities such as self-learning via machine learning (ML). ML can leverage the data generated across the network and enable intelligent self-learning decision-making mechanisms to manage network complexities and improve network performance and efficiency. Network operators need to disrupt the conventional and traditional models used in the previous generations and develop an ML-driven network architecture to enhance network functionalities and enable new networking services and applications that can be executed in a predictive manner to successfully deploy the next-generation network. Standards organizations such as 3rd Generation Partnership Project (3GPP) have started working toward developing standardization support for ML-enabled use cases. ML-based frameworks would assist network operators to reduce the complexities in the network and improve user experience by studying and analyzing the network data collected and autonomously looking for patterns that can yield further insights.

Network operators are not going to just flip a switch to turn off the 4G network and deploy the 5G and beyond network, but will rather need to ensure a speedy but agile 5G rollout by considering the customer impact during network migration, the capital and operational expenditures, and the business opportunities that can be developed for enhanced mobile broadband, massive machine type communications, and ultra-reliable low-latency communications.

ML is expected to become a powerful technique of network association for substantially improving the network performance by accurately learning the near-real-time physical operating scenario, and the motivation and benefits behind using ML algorithms for future networks were further elaborated in [1]. The authors in [2] overviewed some of the key factors for successful AI deployment and integration in future networks. A four-layered architectural approach to induce the AI-enabled functions for intelligent 6G networks was described in [3]. A comprehensive survey on the application of AI in wireless networks from the data life cycle perspective was presented in [4].

The state-of-the-art literature has investigated the benefits, applications, and design challenges in applying ML to future networks, but an ML-driven network architecture remains to be well defined and to materialize. Network operators would resonate with an architectural view that is built on the present network models for 5G and beyond systems and tailor it with upcoming 3GPP and International Telecommunication Union — Telecommunication Standards Sector (ITU-T) standards. In this article, a systematic approach is taken toward developing the foundation for ML-driven next-generation network architecture starting with three viewpoints, including the examination of a computing-driven network infrastructure, understanding the role of networking for ML, and analyzing the role of ML in network architecture, subsequently leading to an appropriate fusion of ML and network architecture. It also refers to 3GPP and ITU-T standards and state-of-the-art literature to develop a unified layout for ML-driven network architecture.

This article explores the role of ML and its interplay with wireless communications networks such that appropriate network architecture can be

Digital Object Identifier: 10.1109/MCOM.006.2100456

Chetana V. Murudkar is with the University of South Florida and T-Mobile USA, Inc., USA; Kwang-Cheng Chen and Richard D. Gitlin are with the University of South Florida, USA.

developed to facilitate where and how such ML functionalities shall be located inside the network. New data flow paths, networking/computing entities, and interfaces are introduced to facilitate more efficient networking functionalities beyond the existing concepts. An illustrative scenario for ML-driven network architecture is demonstrated. Future directions are discussed. Finally, conclusions are drawn.

ML-Driven Network Architecture for 5G and Beyond Systems

Unlike existing static optimization methods, ML-based methods can proactively learn and dynamically reconfigure network functionalities by extracting relevant features and applying that knowledge or model for network optimization. Network operators capture an abundance of data about their subscribers, and ML can help exploit and sift that data in a variety of ways to improve customer experience. In order to get the most value from ML, an appropriate foundation of ML-driven next-generation architecture is critical so that network operators can deploy network functionalities and algorithms that can scale resources as per customer demand within acceptable levels of capital and operational expenditures.

ML-based solutions should comply with service level agreements, function stably at operating frequencies within the allotted spectrum range available to the network operator, allow interoperability supporting multiple vendors' equipment, and be sustained in a dynamically changing network environment. Therefore, it is important to develop a certain degree of standardization effectively facilitating ML-driven network deployment and management. Figure 1 envisages a new network architecture for ML in 5G and beyond networks based on three viewpoints that include the examination of a computing-driven network infrastructure, understanding the role of networking for ML, and analyzing the role of ML in network architecture, subsequently leading to an appropriate fusion of ML and network architecture.

Computing-Driven Network Infrastructure

It has become pivotal to examine the computing environment for 5G and beyond networks to explore the interplay between networking and computing in which ML can assist future networking, and networking can enable new scenarios using ML. The goal of installing ML in networking architecture is to execute networking functionalities or networked services in a predictive manner. ML relies on high-performance computing that can sit on cloud, edge (or fog), and agents, all requiring networking infrastructure to enable efficient ML.

In cloud computing, the data center, database, and so on are externally connected to the core network (CN). However, such big user data or control data can be deeply learned and analyzed and internally connected to CN in 5G and beyond network architecture. 3GPP has defined the network data analytics function (NWDAF) [5], which can provide analytics to 5GC network functions (NFs) and operations, administration, and maintenance (OAM). NWDAF comprises analytics logical function (AnLF) and model training logical function (MTLF). AnLF can perform inference, derive analytics information, and expose analytics service. Analytics information are either statistical information of past events or predictive information, and can include information such as slice and NF load level information, observed service experience information, network performance information, user equipment (UE) mobility information, UE communication information, expected as well as abnormal UE behavioral information, user data congestion information, and quality of service (QoS) sustainability. MTLF trains ML models and exposes new training services (e.g., providing a trained model). NWDAF is also expected to support edge computing applications by providing user plane performance analytics in the form of statistics or predictions to a service consumer. Edge computing enables processing of computation-intensive and delay-sensitive services and applications at the service provider's network border by moving the computing resource and data storage from CN to the edge of the network. Edge computing, from this view, will be connected to the radio access network (RAN) executing edge networking functionalities. Edge AI and analytics combines AI with edge computing. It enables processing of ML algorithms locally within the RAN network without needing to connect to the cloud. State-of-the-art edge computing that may consist of many core parallel processors can support ML for data analysis and AI. In agent computing, each smart agent typically has impressive computing power to execute complex ML and AI functionalities.

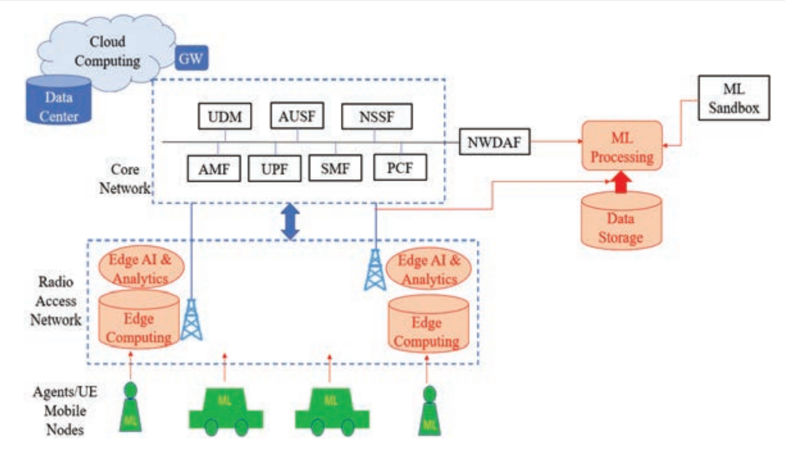

FIGURE 1. Interplay between 5G (and 4G) networking and computing, in which ML can assist future networking, and networking can enable new scenarios using ML/AI. AMF: access and mobility management function; UPF: user plane function; SMF: session management function; PCF: policy control function; UDM: unified data management; AUSF: authentication server function; NSSF: network slice selection function;NWDAF: network data analytics function.

Role of Networking for ML

The purpose of ML is not just for (big) data analysis. ML facilitates AI applications that commonly require networking to achieve safety, reliability, and overall efficiency. Agents and sensors that support smart applications generate big data, and ML can take advantage of inference from big data assisted networking functionalities. For many cases of ML in smart agents, correctly received messages of longer delay can be practically useless. For example, for an autonomous driving vehicle, a message from another agent to indicate a deer jumping into the road can be useless (as not enough time to act) if the end-to-end networking latency is several hundreds of milliseconds.

It is commonly recognized that applying ML can enhance network functionalities, and the application scenarios of ML in the network can be categorized into online and offline ML computing.

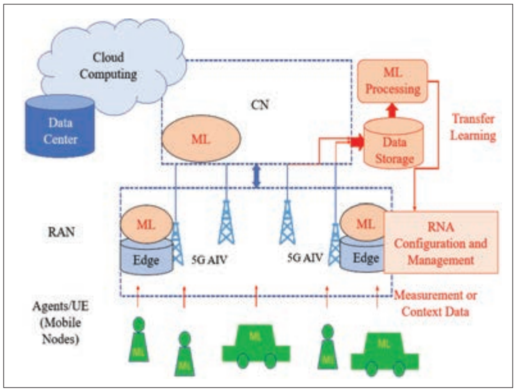

FIGURE 2. An illustration of ML-driven network architecture for RNA configuration and management. AIV: air interface variant.

Network infrastructure for ML applications would help achieve scalability and time-bounded performance. It is expected that future network architecture shall support or enable ML operating inside smart agents, which involves massive operation of autonomous vehicles, robots, smart manufacturing, tele-medicine, and more. Multiple robots or agents of AI are well known to collectively work toward human benefits, which form a multi-agent system (MAS). If a MAS has (wireless) networking capability, it is known as networked MAS. Although MAS has been widely studied, MAS with networking capability has been little covered.

Role of ML in Network Architecture

It is commonly recognized that applying ML can enhance network functionalities, and the application scenarios of ML in the network can be categorized into online and offline ML computing. In online ML, the ML functionalities are embedded into networking operating algorithms or protocols, and thus must be implemented into the corresponding network entities such as NWDAF. If the ML functionality is executed and then used to assist network functionalities, usually with a larger computational delay, it is known as offline ML computing, which can be executed in a co-located computing facility properly connected to the corresponding network entities. Offline ML also can be computed in another far-away computing facility and then apply transfer learning to the target network entities. Ongoing ITU-T studies have defined the concepts of ML sandbox and ML pipeline toward developing a conceptual architectural framework for ML in future networks [6]. An ML sandbox is an environment in which ML models can be trained, tested, and evaluated before their applications to operating networks. An ML pipeline is a set of logical nodes, each with specific functionalities, that can be combined to form an ML application in a telecommunication network.

Given the fact that ML can be executed in the agents (i.e., UE), edge (i.e., RAN), and cloud (i.e., CN or beyond gateway), ML can play different roles in executing a specific networking functionality following which ML-based networking can be categorized as ML-aware, ML-aided, and ML-enabled. ML-aware networking is where networking entities know the availability of ML functionalities, and it is optional to take advantage of the ML capability. In ML-aided networking, network functionalities are mandatorily enhanced by adopting ML technology, which can be offline ML or online ML. In ML-enabled networking, network functionalities must rely on ML, which is online ML in principle.

The proposed ML-driven network framework utilizes user data generated across the network and processes it using ML, enabling a user-centric approach such that network strategies and solutions can be tailored per user needs and feedback. This may include data extraction of the connectivity and mobility status of users, studying user behavioral patterns, profiling end user's perception of QoS, and utilizing end user's input and feedback for optimization. To facilitate ML-based networking functionalities, new traffic will be potentially included such as sensor and measurement data, context information/context-aware data, and planning and policy for ML functionalities. Sensor and measurement data will be collected for ML training. Context information data can extract location information, user experience (i.e., latency, jitter, packet drops, etc.), and private references (e.g., map or topology). Planning and policy for ML functionalities can comprise agent's policy and/or rewards, extracted features or learning models, and more, while the agent may be a network entity or a smart UE (e.g., a robot). The traditional control plane and data plane may no longer be adequate in data handling and transmission and will need a new ML plane to manage ML-based data traffic.

The realization of this new network architecture will help achieve a certain degree of standardization effectively facilitating ML-driven 5G and beyond systems. An application scenario is discussed in the next section along the lines of this new architecture as an illustration.

An Illustration of ML-Driven Network Architecture

Mobility management remains a key component of existing as well as next-generation wireless communications networks, and an ML-driven architecture can effectively maneuver signaling and data resources via mobility predictions. The network function design of ultra-low-latency mobile networks proposed in [7] indicates anticipatory mobility management enabled by online ML. The authors in [8] illustrated an NWDAF-based architecture for mobility management in the CN by proposing an ML-assisted CN-based paging method. The remainder of this article focuses on developing an adaptive mechanism for RAN-based paging in an ML-driven network architecture.

The logical/functional split between CN and RAN led to the introduction of RAN-based notification areas (RNAs). An RNA cluster may constitute cells covered by one or more 5G network nodes (base stations) enabling RAN-based paging. This will allow network operators to effectively manage network resources as it significantly reduces latency by lowering CN/RAN signaling by more than 85 percent and regulates UE power dissemination especially suitable for bursty connectivity and massive access [9].

Figure 2 is an illustration of using offline ML for RNA configuration and management to comprehend the new realization of ML-driven network architecture. ML-driven data flows, interfaces, and entities are in red. Measurement and context data are collected for ML training, stored in the data storage, processed in the ML processing unit, and transferred to the RNA configuration and management unit for RNA cluster formation and optimization. The following subsections include an overview of radio resource control (RRC) state handling and transitions, key RNA configuration factors, and a case study to demonstrate and evaluate the proposed RNA configuration technique, and provide future recommendations to further enhance the technique to attain a more robust and reliable approach.

RRC State Handling and Transitions

A typical LTE network has two RRC states, RRC connected and RRC idle. The RRC connected mode is activated during data transfer, and UE enters RRC idle mode when there is no data to be transmitted or received. A 5G network is expected to encounter a large amount of random aperiodic and keep-alive traffic generated by a plethora of autonomous applications and services supported by 5G that will cause several RRC state transitions, adversely affecting signaling and paging load, latency, power consumption, and capacity of the network. A new RRC state, RRC inactive, has been introduced in 3GPP standards to address these issues.

A UE is either in RRC connected state or in RRC inactive state when an RRC connection has been established, but if this is not the case (i.e., no RRC connection is established), the UE is in RRC idle state [10]. Transitions from RRC idle to RRC connected are expected to occur mainly when a UE first attaches to the network, while transitions from RRC inactive to RRC connected are expected to occur frequently and are optimized to be fast and lightweight in terms of signaling achieved by keeping the CN-RAN connection alive during inactivity periods allowing the UE to move around within a pre-configured area (the RNA) without notifying the network [9].

RNA Configuration Factors

One of the most important factors while configuring RNA clusters is analyzing the user activity by means of UE-gNB connections as that has a direct impact on signaling load. The RF conditions will help gauge the boundaries of RNA clusters; hence, it is important to incorporate reference signal received power (RSRP) and signal-to-noise-plus-interference ratio (SINR) conditions of the user connections. The RSRP measurements help analyze path loss, and SINR measurements can be used to ensure good cluster throughput. Another aspect that is critical in RNA cluster formation is the paging load. In LTE, paging is a CN function that is envisaged to be moved into the RAN in 5G by taking advantage of the RRC inactive state and RNA, thus allowing RAN-controlled paging initiation procedures [9]. A RAN-initiated paging vs. a CN-initiated paging procedure for a 5G network can be described as shown in Fig. 3 [11]. For M cells and N gNBs per RNA, the paging load (in terms of the number of messages) in the RAN-initiated paging is equal to the sum of M messages over radio and $(N - 1)$ messages over Xn, whereas the paging load in the CN-initiated paging is equal to the sum of M messages over radio, $(N + 3)$ messages over N2, and 3 messages over N4 and N11, where Xn is the interface between gNB-gNB, N2 is the interface between RAN and AMF, N4 is the interface between SMF and UPF, and N11 is the interface between AMF and SMF [11].

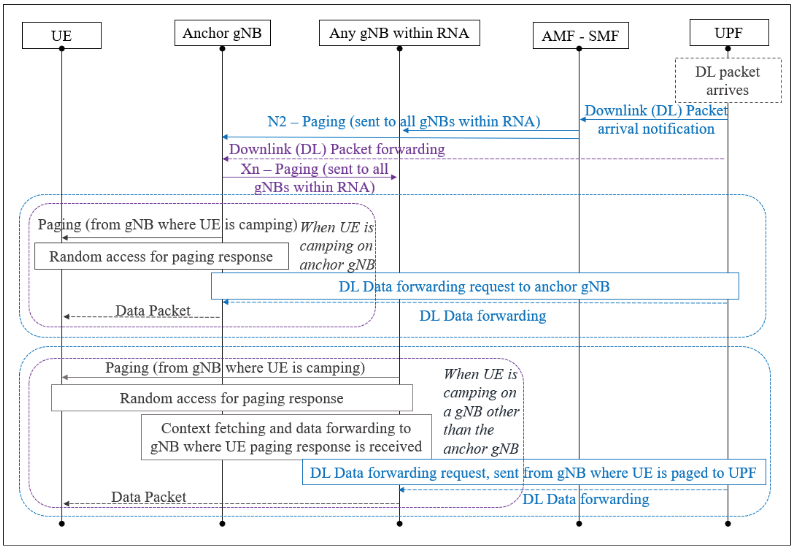
UE
Anchor gNB
Any gNB within RNA
AMF - SMF
UPF
DL packet arrives
Downlink (DL) Packet arrival notification
N2 – Paging (sent to all gNBs within RNA)
Downlink (DL) Packet forwarding
Xn – Paging (sent to all gNBs within RNA)
Paging (from gNB where UE is camping)
When UE is camping on anchor gNB
Random access for paging response
DL Data forwarding request to anchor gNB
Data Packet
DL Data forwarding
Paging (from gNB where UE is camping)
When UE is camping on a gNB other than the anchor gNB
Random access for paging response
Context fetching and data forwarding to gNB where UE paging response is received
DL Data forwarding request, sent from gNB where UE is paged to UPF
Data Packet
DL Data forwarding

FIGURE 3. RAN-initiated paging procedure (purple and black) vs. CN-initiated paging procedure (blue and black) for a 5G network.

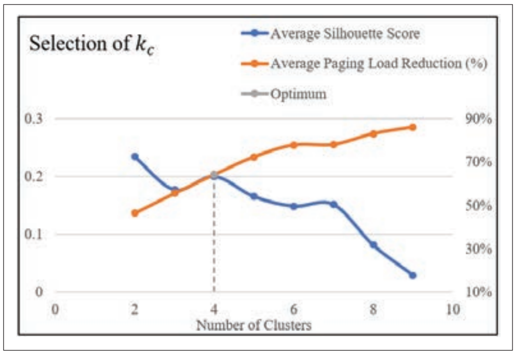

FIGURE 4. Performance evaluation for the selection of k_c.

Performance Analysis and Evaluation

This section performs a case study to demonstrate and evaluate the performance, feasibility, and potential benefits of the proposed RNA clustering mechanism.

For verification, a simulated network consisting of several users being served by multiple network nodes representing 4G/5G base stations is configured using the ns-3 simulator. The maximum transmission power of network nodes that are located outdoors is set up to 40 W and is set up to 20 W for indoor nodes. The channel bandwidth considered is 20 MHz, and the testing frequencies include 700 MHz and 2.6 GHz. The proportional fair algorithm is applied for scheduling. The radio propagation model used is Cost231, which is designed to cover an elaborated range of frequencies to predict path loss for outdoor scenarios in urban areas. The indoor scenarios are mimicked by creating a building with user-defined dimensions and attributes and applying a hybrid buildings propagation loss model that is a combination of several well-known path loss models. The users are allocated in random positions, and the mobility model assigned is 2D random

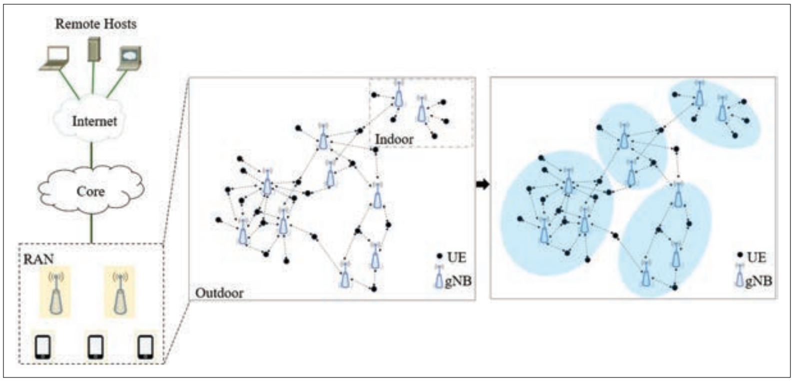

FIGURE 5. The network model and the resulting RNA clusters formed with $k_C = 4$. The arrows represent the connectivity status of the mobile UEs with the network nodes as they move around in the network.

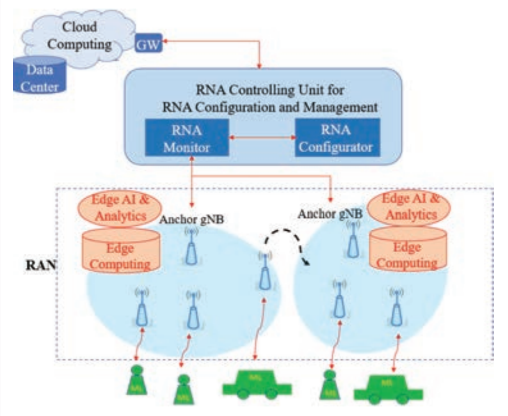

FIGURE 6. ML-based hybrid SON framework for RNA configuration and management.

walk mobility where each user moves with certain speed and direction chosen at random until a certain amount of time, after which the users randomly change their positions and directions. The A3 RSRP algorithm is implemented to trigger handovers. An A3 event is triggered when the UE perceives that a neighbor cell's RSRP is better than the serving cell's RSRP by an offset. RSRP and SINR statistics are collected during the entire simulation run for every user and its serving network nodes.

During data preprocessing, thresholds for RSRP and SINR conditions are set to consider measurements that are within an acceptable range of radio conditions so that RNA clusters are configured to meet the minimum allowable range of signal strength and throughput requirements and diminish the ping- pong effect at cluster boundaries. The processed data defines the relationship between every network node and user, and the size of the dataset is defined by the number of network nodes and users.

ML processing based on uncertainty or incomplete information can be common in wireless networks as the network data is subjected to various scenarios that may be associated with heterogeneous small- and large-scale structures relying on a diverse range of parameters in a constantly changing RF environment. The *k*-means algorithm, an unsupervised ML algorithm, performs cluster analysis that can help identify patterns and derive correlations between data samples to make the best possible sense of the data and make predictions even when labeled data samples are not available to create appropriate groups of network nodes per RNA.

The simulated network data collected is not explicitly labeled. The *k*-means algorithm is implemented in Python to operate on this dataset. It has excellent fine-tuning capabilities and can be effectively used as a subset of more complex algorithms. As the number of dimensions increases, pre-clustering steps such as spectral clustering can be applied where necessary followed by the use of *k*-means to cluster the data in a lower-dimensional subspace, making it a good candidate for cluster analysis. The *k*-means clustering is computationally efficient and can adapt to new samples, making it a scalable solution.

The *k*-means algorithm clusters data by dividing a set of samples of a dataset into *k* disjoint clusters, each described by the centroid (mean) of the samples in the cluster as described below.

1. Create a dataset, or matrix, of dimensions defined by the number of network nodes and the number of users to represent the relationship (i.e., the connectivity status) between every network node and user under nominal conditions for a given period. (The matrix entries are populated over time and are set to zeroes and ones to show the association between every network node and user and capture the user mobility status.)
2. For an initial setting of *k* clusters, choose *k* samples from the dataset to select the initial centroids (i.e., the initial cluster centers are selected using the *k*-means++ initialization method).
3. Repeat the steps below until convergence, which occurs when the centroids stop changing:
   - For all the data samples, find the closest (in the sense of Euclidean distance) centroid.
   - Create new centroids by taking the mean value of all samples assigned to each preceding centroid and compute the difference between the previous and new centroids.

The determination of the optimal number of clusters is performed by using silhouette analysis, and the comparison between the RAN-initiated and CN-initiated paging loads. The optimum number of clusters in this case study are determined with an objective to maximize the paging load reduction, subject to the constraints of the silhouette scores. Silhouette analysis [12] measures the quality of clustering by studying the separation distance between the resulting clusters to validate the consistency within them. Silhouette coefficients have a range of [−1,+1] such that the worst value is −1 and the best value is +1. A value of +1 indicates that the sample is far away from the neighboring clusters. A value of 0 indicates that the sample is on or very close to the decision boundary between two neighboring clusters, and negative values indicate that those samples might have been assigned to the wrong cluster. The silhouette coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample.

The average silhouette score provides an evaluation of clustering validity and is used to select an "appropriate" number of clusters. Silhouette analysis and the paging load reduction obtained after comparing the RAN-initiated paging over CN-initiated paging were plotted, and the optimum value for the number of clusters (i.e., the value of k_c) is found to be equal to 4, as depicted in Fig. 4, corresponding to the highest point of intersection as the higher the paging load reduction and the silhouette score, the better are the clustering results. A paging load reduction of 64 percent is observed in the resulting RNA clusters. It gives a balanced tradeoff between maximizing the average paging load reduction and maximizing the average silhouette score. The silhouette analysis prevents the number of clusters from becoming arbitrarily large, and the paging load reduction prevents the number of clusters from becoming arbitrarily low.

One limitation of the *k*-means algorithm is that it may not always succeed in optimizing the centroid locations globally and can get stuck at a local minimum. To address this, a more powerful ML algorithm, spectral clustering [13, 14], is implemented. Instead of clustering in the original space, the data is first mapped to a new space such that similarities are made more apparent using Laplacian Eigenmaps [13] to place data instances in such a way that the similarities between neighboring instances are preserved and clustering is applied to a projection of the normalized Laplacian. In the original space, a local neighborhood is created such that the instances in the same neighborhood are defined by creating an affinity matrix using the kernel radial basis function (RBF). The matrix value of a similar pair of data instances →0 and the value of a dissimilar pair of data instances →1. This has the effect that instances nearby in the original space, probably located within the same cluster, will be placed very close in the new space, whereas those that are some distance away, probably belonging to different clusters, will be placed far apart. The *k*-means clustering is then run with new data coordinates in the new space.

The *k*-means algorithm would work effectively for simple cluster formations, and spectral clustering can serve as an extension to *k*-means and would be preferred for more general problems. The application of *k*-means and spectral clustering both provide identical resulting clusters for the example network model used in this case study. For a more complex network, it is recommended to combine these unsupervised learning algorithms with deep learning. The simulated network model and the resulting RNA clusters formed with $k_c = 4$ are depicted in Fig. 5.

Future Research Directions

To improve network resilience and robustness, it is recommended that the RNA clusters should be monitored periodically and fine-tuned post initial clustering. This can be achieved by monitoring key performance indicators (KPIs) such as user throughput, traffic volume density, end-to-end latency, reliability, availability, and retainability. The computational capabilities and scalability required to effectively embed these KPIs can be achieved by adopting an ML-based hybrid self-organizing network (SON) framework.

A conceptual framework to enable self-configuration and management of RNAs is proposed as depicted in Fig. 6, which is analogous to a hybrid SON structure [15] where the centralized management system represented by the RNA controlling unit and the element management system represented by the anchor gNBs work together, in a coordinated manner, to build up a complete SON algorithm. The decisions on SON actions may be made by either the RNA controlling unit via centralized computing or the anchor gNBs via edge computing. An anchor gNB is the network node that is aware of or has the list of all the gNBs that are a part of that RNA. It is the anchor gNB that maintains the CN-RAN connection and the UE context as the UE moves around within the RNA. The initial RNA clustering mechanism illustrated in the previous subsection is implemented in the RNA configurator.

Once initial clusters are formed, the RNA monitor would monitor KPIs for each cluster. A cluster-level threshold margin can be set such that it would trigger either an addition or removal of a gNB from a cluster or trigger re-initiation of RNA clustering depending on the tolerance limit set for the threshold variations. If a gNB is moved from one cluster to its adjacent cluster to maintain an acceptable level of cluster performance, the anchor gNBs of the clusters that underwent the changes would relay this information to the RNA monitor so that the modified clusters are considered for future monitoring. When the RNA monitor detects threshold variations exceeding the tolerance limit set, the RNA monitor would trigger the RNA configurator to re-initialize and form new RNA clusters.

This framework will achieve improved signaling and paging load to attain reduced latency and improved capacity, which are two of the key requirements for emerging use cases where an appropriate configuration of RAN-based notification areas will play a significant role as it will have a direct impact on controlling the RRC state transitions.

Conclusion

In the orchestration of the 5G and beyond era, network operators need to expand their current scope of network deployment operations by integrating ML technologies. This article initiates a discussion on developing a certain degree of standardization effectively facilitating ML-driven network deployment and management. A new perspective of network architecture for ML in 5G and beyond networks is envisaged using a computing-driven infrastructure. Furthermore, an illustrative scenario is demonstrated based on the ML-driven architecture where a self-learning mechanism that can predictively configure RAN-based notification areas is proposed, demonstrated, and evaluated, enabling a user-centric smart RAN paging technique. Future research includes effective RNA configuration and management in an ML-based hybrid SON system facilitating ML-driven next-generation self-organizing 5G/6G networks that would help improve signaling load to attain reduced latency and improved network capacity.

References

[1] J. Wang *et al.*, "Thirty Years of Machine Learning: The Road to Pareto-Optimal Wireless Networks," *IEEE Commun. Surveys & Tutorials*, vol. 22, no. 3, 2020, pp. 1472–1514.

[2] U. Challita *et al.*, "When Machine Learning Meets Wireless Cellular Networks: Deployment, Challenges, and Applications," *IEEE Commun. Mag.*, vol. 58, no. 6, June 2020, pp. 12–18.

[3] H. Yang *et al.*, "Artificial-Intelligence-Enabled Intelligent 6G Networks," *IEEE Network*, vol. 34, no. 6, Nov. 2020, pp. 272–80.

[4] D. C. Nguyen *et al.*, "Enabling AI in Future Wireless Net-

works: A Data Life Cycle Perspective," *IEEE Commun. Surveys & Tutoials.*, vol. 23, no. 1, 2021, pp. 553–95.

[5] 3GPP TS 23.288, "Architecture Enhancements for 5G System to Support Network Data Analytics Services," v. 17.0.0, Mar. 2021.

[6] ITU-T Y.3172, "Architectural Framework for Machine Learning in Future Networks Including IMT-2020," June 2019.

[7] C.-Y. Lin *et al.*, "Anticipatory Mobility Management by Big Data Analytics for Ultra-Low Latency Mobile Networking," *IEEE ICC*, May 2018.

[8] J. Jeong *et al.*, "Mobility Prediction for 5G Core Networks," *IEEE Commun. Stds. Mag.*, vol. 5, no. 1, Mar. 2021, pp. 56–61.

[9] P. Marsch *et al.*, "5G Radio Access Network Architecture: Design Guidelines and Key Considerations," *IEEE Commun. Mag.*, vol. 54, no. 11, Nov. 2016, pp. 24–32.

[10] 3GPP TS 38.331, "NR; Radio Resource Control Protocol Specification." v. 15.11.0, Sept. 2020.

[11] S. Hailu and M. Säily, "Hybrid Paging and Location Tracking Scheme for Inactive 5G UEs," *Euro. Conf. Networks and Commun.*, June 2017.

[12] P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *J. Comp. Appl. Math.*, vol. 20, Nov. 1987, pp. 53–65.

[13] Ethem Alpaydin, *Introduction to Machine Learning*, MIT Press, 2014.

[14] U. von Luxburg, "A Tutorial on Spectral Clustering," *Stat. Comp.*, vol. 17, no. 4, Dec. 2007, pp. 395–416.

[15] 3GPP TR 28.861, "Telecommunication Management; Study on the Self-Organizing Networks for 5G Networks." v. 16.0.0, Dec. 2019.

Biographies

Chetana V. Murudkar [S'18] (cvm1@usf.edu) received her Ph.D. degree in electrical engineering from the University of South Florida where she is a part of the Advanced Wireless Networking group investigating new technologies that address research challenges directed toward emerging 5G and 6G systems. She is an engineer at T-Mobile USA, Inc. and her responsibilities involve design, deployment, performance assessment, and optimization of its wireless communications network. Her past work experience includes working at Sprint, AT&T Labs, and Ericsson.

Kwang-Cheng Chen [F'07] (kwangcheng@usf.edu) is a professor of electrical engineering, University of South Florida, Tampa. He has widely served in IEEE conference organization and journal editorship. He has contributed essential technology to IEEE 802, Bluetooth, LTE and LTE-A, and 5G-NR wireless standards. He has received a number of IEEE awards. His recent research interests include wireless networks, quantum computing and communications, machine learning and multi-agent systems, and IoT/CPS.

Richard D. Gitlin [LF'09] (richgitlin@usf.edu) is a Distinguished University Professor (Emeritus) at the University of South Florida and a State of Florida 21st Century World Class Scholar. He is a member of the NAE and a Fellow of the NAI and Bell Labs. He was at Bell Labs for 32 years, including serving as Senior Vice President for Communications and Networking. He is best known as the co-inventor of Digital Subscriber Line and as a pioneer in the application of advanced spatial signal processing (now known as MIMO) in wireless systems.