Supporting 6G Mission-Critical Services on O-RAN

Rafael Kaliski and Shin-Ming Cheng

Abstract

In the era of 6G, cellular networks will no longer be locked into a small set of equipment manufacturers; instead, cellular networks will be disaggregated and support open interfaces. Thus, there is an inherent need for networking functions to be softwarized and virtualized so that customers can apply different vendors' solutions. 6G mission-critical networks must be dependable and secure, ultra-reliable and low latency, and support high connectivity, all while being flexible enough to support custom user deployments. 6G will integrate Artificial Intelligence (AI) into the network architecture to meet the diverse user requirements of 3rd party solutions. A possible 6G candidate capable of supporting the requirements above is Open Radio Access Networks (O-RAN). O-RAN enables multiple levels of AI-based control for RAN Intelligent Controllers (RICs). RICs facilitate real-time sensing, reaction, policy determination, and management of radio resources. When coupled with Multi-access Edge Computing (MEC), O-RAN enables customized per-device AI service chains that can address the needs of dynamic, diverse 6G networks in real-time. This article presents an O-RAN architecture that supports split-plane multi-component cooperative AI models that utilize multiple RIC-centric and MEC-centric control loops. Through multiple example applications and O-RAN testbeds, we demonstrate the efficacy of our proposed architecture and how it can address the multitude of 6G requirements as necessitated for mission-critical Internet of Things applications.

Introduction

6G and Internet of Things (IoT) networks will require a high level of security, intelligent traffic control to enhance Quality of Experience (QoE) and Quality of Service (QoS), and reduced computational requirements due to the diverse hardware and deployment scenarios [1]. To better address these requirements, 6G necessitates data-driven dynamic network configurations, which Artificial Intelligence (AI) services control in real-time. The 6G requirements for devices, in terms of connectivity, latency, bandwidth, reliability, and services, are diverse. This implies that 6G mission-critical requirements (e.g., Extreme Ultra Reliable Low Latency Communication (URLLC), end-to-end QoS, spectrum efficiency, high connection density, energy efficiency, dependability, security, etc.) and related applications [1] will likely need to replace the prior one-size-fits-all configurations and architectures with network disaggregated architectures (as 6G is based on 5G, the user plane is separate from the control plane.) To address these requirements, independent hardware solutions, latency-specific control loops, and service chains (a service chain refers to a set of protocols, slices, and features) where features can be selectively added or removed via an AI as a Service (AIaaS) paradigm on a per-device basis will be necessitated. Compared to 5G, where Mobile Network Operators (MNOs) defined the requirements, in 6G, in part due to high density and short-range deployments, new players will drive the specific system requirements and services; for IoT, this implies cost-reduced and energy-efficient deployments. To enable service chains, softwarization and virtualization of network functions and open interfaces will be essential.

One approach to 6G is 6G Network in a box [2] (6G-NIB), where the essential network components are self-contained such that network coverage in traditional no-service coverage areas becomes possible, i.e., 6G-NIB enables truly ubiquitous networking across numerous interfaces. NIB comes in a multitude of low-cost flavors, such as ISP-NIB, Virtual-NIB, and LTE-NIB, depending on the required deployment. Unlike traditional cellular networks, which are designed for larger region-wide deployments, NIB is designed for localized deployments. Another approach to 6G is space-air-ground integrated networks (SAGINs) [3], which addresses Quality of Service (QoS) for a multitude of emerging 6G services (latency and/or reliability) via seamlessly integrating terrestrial, aerial, and satellite networks using agile microservices and edge intelligence. Finally, the approach we propose is for 6G networks to be based on Open Radio Access Network (O-RAN); O-RAN is open-standard. Via open-interfaces and function virtualization, O-RAN transforms fixed network architectures, such as 5G, into flexible dynamic deployments [4]. Multi-vendor components and 3rd party applications can utilize the open-standard interfaces to build O-RAN networks that can interoperate with different deployment configurations [5] and offer custom solutions.

Compared to 5G, the main entities introduced by O-RAN are RAN Intelligent Controllers (RICs). RICs can enable AI on the network, where Machine Learning (ML) can improve system responsiveness and maintenance, such as via dynamic spectrum sharing, network slices, and Physical layer (PHY) configurations [6]. By dividing the RIC into multiple layers, based on the latency requirements, customized and efficient RIC applications (APPs) can be developed to meet the specific needs of the service(s) being deployed via custom policies. O-RAN's RIC is divided into three layers, the Non-Real Time (Non-RT) RIC, the Near-RT RIC, and the RT-RIC. In terms of policy implementation, long-term policies are directed by the Non-RT RIC's rAPPs, implemented by the Near-RT RIC's xAPPS, and executed by the RT-RIC's zAPPs [6].

When considering AI on the network, RICs and Multi-access Edge Computing (MEC) serve two distinct yet complementary purposes. RICs offer ML-based management of control plane signaling, such as User Equipment (UE) connection metrics, while MECs offer ML-based models to manage user plane signaling. Through the addition of MEC to O-RAN, more powerful and efficient detection models can be implemented and work in concert with the RICs and the RAN. Due to the volume of users, limited computation power, and need to quickly process control plane signaling, lightweight ML models such as weak

Rafael Kaliski (corresponding author) is with National Sun Yat-Sen University (NSYSU), Taiwan.

Shin-Ming Cheng is with the National Taiwan University of Science and Technology (NTUST), Taiwan.

This work was supported by the National Science and Technology Council (NSTC), Taiwan under grants: NSTC 108-2218-E-011-036-MY3 and NSTC 112-2221-E-110-023.

Digital Object Identifier: 10.1109/IOTM.001.2300032

learners (Linear Regression, Random Forest, Support Vector Machine, Multi-layer Perceptron, etc.) and ensemble learners (ensemble learners are comprised of multiple weak learners) are best targeted for RIC r/x/zAPPs. More complex models designed for user plane data, such as Deep Learning (DL) models, Convolutional Neural Networks (CNNs), and Federated Learning (FL), are best targeted for MEC servers that have higher computing power.

Each RIC and MEC controls a subset of the RAN components via separate control loops (each RIC layer and control loop offers a different latency. A hierarchy of RICs offers more efficient, targeted to a specific latency, easier to construct ML models and solutions). In addition to the standard O-RAN control loops, i.e., Non-RT RIC, Near-RT RIC, and RT-RIC, we consider a yet-to-be-discussed fourth control loop, the MEC control loop. The fourth control loop enables UEs to communicate indirectly with the Near-RT RIC. In addition, the fourth loop enables the Near-RT RIC to manage and share resources with multiple MECs. By utilizing the fourth control loop, a UE's communication with the O-RAN's Native-ML can effect changes on their network configuration and indirectly utilize RAN-controlled resources via the MEC.

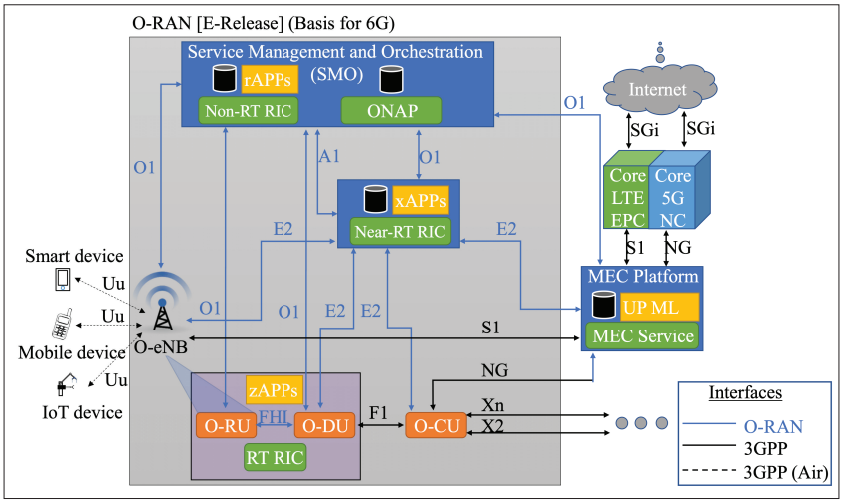

FIGURE 1. O-RAN Architecture with MEC and multi-cell connectivity. One or more O-RUs may be connected to an O-DU via the fronthaul interface (FHI). One or more O-DU may be present (one Near-RT-RIC per O-DU). Each O-RU resides in an O-eNB. O-RAN architecture is based on 3GPP New Radio (NR) option 7.2X split (PHY-low, PHY-High) for the RU and option 2 splits (control plane - user plane split) for the CU [7].

Current O-RAN designs do not emphasize cooperative multi-component ML models and O-RAN's synergy with MEC nor analyze why all four control loops and split-plane ML models are necessitated from an end-to-end behavior standpoint. The insight we gain is that no one monolithic ML model can sufficiently handle a high number of users in a resource-limited environment under tight latency constraints; rather, ML models must be designed for specific applications while considering an overall cooperative design between each model.

This article analyzes the need for three RIC closed control loops, a MEC closed control loop, and why the ML models are split-plane (user-plane (UP) ML models are separable from the control-plane ML models). Our contributions are:

- We present an O-RAN-based architecture that supports multi-component cooperative AI models via split-plane and multiple closed control loops involving the O-RAN, MEC, and UE.
- We analyze our proposed system's end-to-end behavior and how it addresses 6G's mission-critical requirements.
- We present multiple applications and empirical test-bed-based results demonstrating the efficacy of the proposed architecture.

We present an envisioned 6G compatible O-RAN architecture based on O-RAN E-Release. Then we present several AI applications utilizing O-RAN and MEC-based ML models and discuss the end-to-end behavior of malicious application detection using a custom xAPP we developed. After which, we present a couple of applications using a hybrid ML model where the UE ML model needs to communicate with the O-RAN and/or MEC ML models to change network configurations. Finally, we present our conclusion.

O-RAN Architecture

In recent years, O-RAN has gained popularity in, and support from, both academia and industry [6]. As O-RAN already addresses many of 6G's requirements, its architecture [7] can be considered as a reference basis for 6G. Figure 1 shows the network architecture (for latency purposes, O-RAN splits the user plane from the control plane). Note: Even though O-RAN is compatible with both 4G and 5G networks, some applications are only compatible with 5G, such as dynamic Resource Block (RB) geometry.

Summary of O-RAN Components and Interfaces

Service Management and Orchestration: The Service Management and Orchestration (SMO) uses the O1 interface to manage and monitor all O-RAN-connected components and services.

Open Network Automation Platform: The Open Network Automation Platform (ONAP) enables policy-driven orchestration and automation of physical and virtual network functions for services in 5G+.

Non-RT RIC: The Non-RT RIC deploys rAPPS to perform ML. rAPPS can control and optimize RAN elements and resources via the O1 interface. In addition, rAPPS can determine policies. The A1 interface is used to gather information from the Near-RT RIC and send policy-related decisions to the Near-RT RIC to enforce. The Near-RT RIC executes said policies via its xAPPs (per the xAPP's results, each policy is evaluated and conditionally performed. The Non-RT RIC's rAPP monitors each xAPP's effect on the system to determine if continued policy execution is required. Either MNOs or 3rd parties may design and deploy rAPPs, thus enabling customized policies.

Near-RT RIC: The Near-RT RIC uses the active ML model and data provided by the Non-RT RIC to determine if and how to execute a policy via an xAPP. Based on the decision, the Near-RT RIC uses the E2 interface to communicate with the RAN and MEC via the E2 nodes (connected elements are the MEC platform, O-eNB, O-RAN Centralized Unit (O-CU), O-RAN Distributed Unit (O-DU) and the O-RAN Radio Unit (O-RU). In addition, the E2 interface is utilized by each Near-RT RIC to collect Near-RT information about the attached UEs and cells. The connection status of each of the E2 node-connected elements and the attached UEs is stored in a Shared Data Layer (SDL) database, which may be queried by any xAPP. Either MNOs or 3rd parties may design and deploy xAPPs, thus enabling customized service chains and network settings.

A1 interface: The A1 interface connects the Non-RT RIC to the Near-RT RIC. It is used for policy management and data transfer (only information which will assist in model training is transferred).

E2 Interface: The E2 interface connects the Near-RT RIC with the E2 nodes as the interfaces/components are separable, and both RAN and function virtualization are simplified, effectively lowering investment costs while increasing system flexibility. The O-CU, O-DU, and O-RU control different network resources, i.e., connectivity, media access, and protocol functionality, respectively.

O1: The O1 interface enables the management of all O-RAN components associated with O-RAN network functions.

Fronthaul Interface: The Fronthaul (FHI) interface connects the O-DU and O-RU to the RT-RIC. The FHI provides control and user plane synchronization and management functionalities.

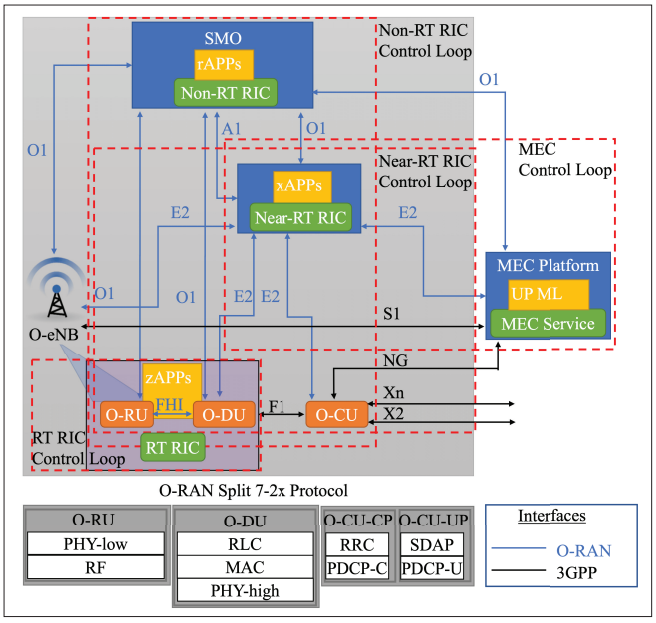

FIGURE 2. Proposed O-RAN architecture with control loops and protocol stacks. Note that the outer control loops overlap with the inner control loops, as they need to interact. As network disaggregation is a requirement of 6G, split-plane protocols are by design.

O-eNB: The O-eNB terminates the O1, S1, E2, and FHI (indirectly NG (core), X2, and Xn) interfaces as well as the relevant radio protocol stacks. The O-eNB acts as a communication point between the O-RAN system and the mobile devices via the Uu interface. As such, the O-eNB effectively enables real-time data collection by each RIC.

MEC and RIC ML Synergy

Current O-RAN designs focus on RAN and cloud connectivity. The inherent limitations placed on ML by the RIC are two-fold:

1. The RIC handles control plane data. Thus analysis of user-plane data is separate.
2. RIC-based ML APPs are limited by the computing power available; thus, resource-intensive ML models remain prohibitive.

Even though cloud computing may offer sufficient resources to run resource-intensive ML models, it suffers from higher latency.

By combining MEC with RIC, O-RAN can not only support user-plane ML models, yet O-RAN can also utilize resource-intensive ML models by partitioning the ML model into a hybrid ML model, where the RIC-based (control plane) ML models interact with the MEC-based (user plane) ML models, via the E2 and NG interfaces, to provide a richer set of low latency services not previously possible via the RIC alone.

AI-Enabled Applications Feedback Loops

Our proposed O-RAN solution utilizes four ML control loops to meet the diverse set of service requirements, as shown in Fig. 2. O-RAN Release E (5G-based) provides three different RICs, each of which targets a different latency granularity: Non-RT RIC (> 1s), Near-RT RIC (10ms → 1s), and RT RIC (< 10ms). The 5G-based O-RAN architecture's latency requirements are achieved via three control loops that handle three different latencies. To enable 6G support, we also consider a non-RIC control loop which enables the UE to communicate indirectly with the Near-RT RIC via the MEC. The control loops are summarized below:

Non-RT RIC Control Loop: The Non-RT RIC control loop consists of the Non-RT RIC and the Near-RT RIC. It is responsive to second-level granularity latency operations via the A1 and O1 interfaces. Long-term policy decisions are typically handled by this loop. By utilizing network slicing, differentiated services can be provided. The Non-RT RIC can monitor each attached O-eNB via the O1 interface. Based on the Non-RT RIC's rAPP results, policies are updated and monitored via the Near-RT RIC's A1 interface. The Near-RT RIC's xAPPs implement the policies.

Near-RT RIC Control Loop: The Near-RT RIC control loop consists of the Near-RT RIC and the RAN components (O-eNB, O-DU, O-CU). The Near-RT RIC is targeted for ten milliseconds to one-second latency operations and uses the E2 interface to manage the attached E2 nodes. The Near-RT RIC is suitable for implementing near-term policies, such as resource management and load balancing, via xAPPs.

RT RIC Control Loop: The RT RIC control loop consists of the O-RU and the O-DU. The RT RIC control loop uses the FHI interface to manage radio parameters and status. The RT RIC enables AI control of lower-layer RAN functions requiring sub-millisecond control, such as the PHY layer, interference management, modulation and code settings, and diverse adaptive operating environments. The real-time policies are executed via zAPPs.

MEC Control Loop: The MEC control loop consists of the Near-RT RIC and the MEC server. The Near-RT RIC is monitored and managed by the MEC via the E2 interface, i.e., the MEC ML model determines policies for the Near-RT RIC to implement. Although the MEC is not part of the 5G O-RAN architecture [5], we note it is essential to consider MEC for 6G as RIC ML solutions are limited in the complexity they can achieve as often RAN computing power is limited. Thus RIC ML solutions are best targeted toward control-plane problems, such as load balancing. Yet when we consider user-plane problems, the complexity of ML models can increase dramatically, e.g., DL and CNN solutions often required for computer vision applications necessitate high-end computing solutions that include a high-end Central Processing Unit (CPU) and possibly a Graphical Processing Unit, which is not available on the RIC. An example application we explore later in this article is monitoring network traffic and UEs via a malware detection application. When an infected device is detected, it is isolated.

The MEC control loop may also be utilized for hybrid ML models (UE and MEC) as necessitated by applications such as FL. In FL, UEs execute their local model, which must coordinate with the MEC to update the global model (often the central FL model uses DL or CNN), which is then pushed out to the UEs so they can seek out improved solutions.

Security Concerns and Countermeasures

As O-RAN is an open-standard system, developers and telecommunication providers can access its source code and develop new features. Unfortunately, this exposes O-RAN's components and interfaces.

The main threats O-RAN encounters can be categorized as:

- **Openness threats:** O-RAN suffers from weak verification of data source identity [8], thus leading to potentially malicious data sources and components.
- **Data threats:** By analyzing network traffic, illicit packets can be constructed to influence the data and parameters used by the ML models via the E2 interface [8].
- **Flow threats:** UE-originating and external network-originating attacks, such as DNS server (located at the network edge), Distributed Denial of Service (DDoS), and malware attacks, can impact the RAN's and network's performance.

The aforementioned threats can be addressed by combining data and behavior pattern analysis with RIC xAPPs (control plane) and MEC-based ML APPs (user plane). Thus to ensure O-RAN integrity and stability, RIC and MEC-based intelligent system monitoring (for rogue components) and network intrusion detection (for interfaces) are necessitated.

O-RAN Assisted Services

Mission Critical O-RAN assisted services is deployed using information gathered from the UE (in this section, we assume the UE does not have an ML model). The collected information is interpreted by and acted on by either the RIC and/or the MEC ML models.

Traffic Steering and Localized QoS

For this application, we consider the Mission Critical requirement of addressing End-to-End QoS in dense networks [1] via V2X and massive Machine Type Communication (mMTC). Vehicles within the vicinity of a traffic accident consider it a high priority, while those further away do not. With the increasing number of connected vehicles, the impact due to bursty network traffic originating from the scene of an accident is region-wide and can cause a corresponding network QoS degradation impacting all vehicles.

By considering the network traffic type, origin, and destination of each network flow, the QoS requirements of each network flow can be dynamically modified, i.e., we can mitigate the impact on region-wide network QoS by utilizing multiple QoS settings per network flow. This can be achieved by steering high-priority network traffic to local resources for immediate distribution, then forwarding it as normal-priority network traffic to regional resources [9], as shown in Fig. 3.

To implement this scheme, in addition to dual connectivity with the regional BS and the local RSU, a MEC server is necessitated to dynamically modify the QoS settings of the forwarded network flows to normal priority. This type of dual connectivity and dynamic QoS modification is not supported in traditional cellular networks. Yet with O-RAN, the locations of the transmitters and receivers can be gathered by the xAPP via the MEC control loop, the MEC ML model, and the Near-RT control loop the QoS of the network traffic can be dynamically modified (the O-CU performs QoS modification).

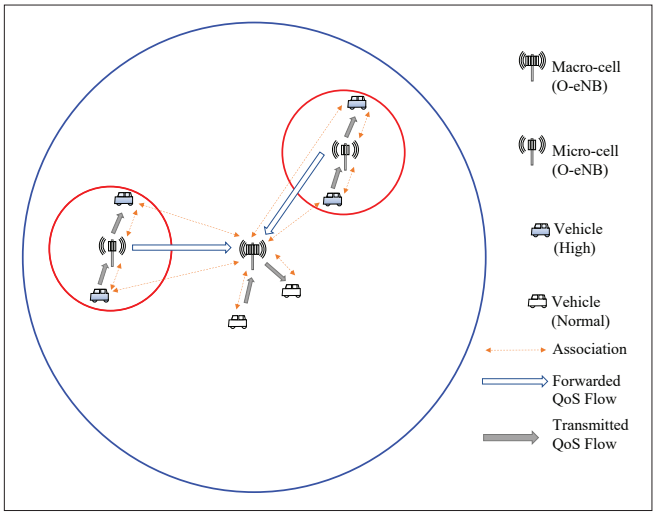

FIGURE 3. Localized QoS: High-priority to-be-broadcast QoS flows are transmitted to the Micro-cell/Road Side Unit (RSU), which acts as a localized resource, and re-broadcast within the Micro-cell. At the same time, the high-priority traffic is forwarded to the Macro-cell/Base Station (BS) for regional broadcast as a normal priority QoS flow.

Proactive Cell Association and Uplink Single Frequency Network

For this application, we consider the Mission Critical requirement of addressing Extreme URLLC [1], such as in critical mMTC. URLLC remains a daunting challenge for wireless networks. The seemingly conflicting goals of low latency and high reliability can be addressed via open-loop feedback-less protocols and proactive multi-cell association [10]. In proactive cell association a UE may associate with one or more cells, as shown in Fig. 4; as the number of cells UE associates with increases, assuming sufficient wireless resources, the outage probability decreases as both Selective Combining and Multi-User Detection increase the number of network paths which may be decoded.

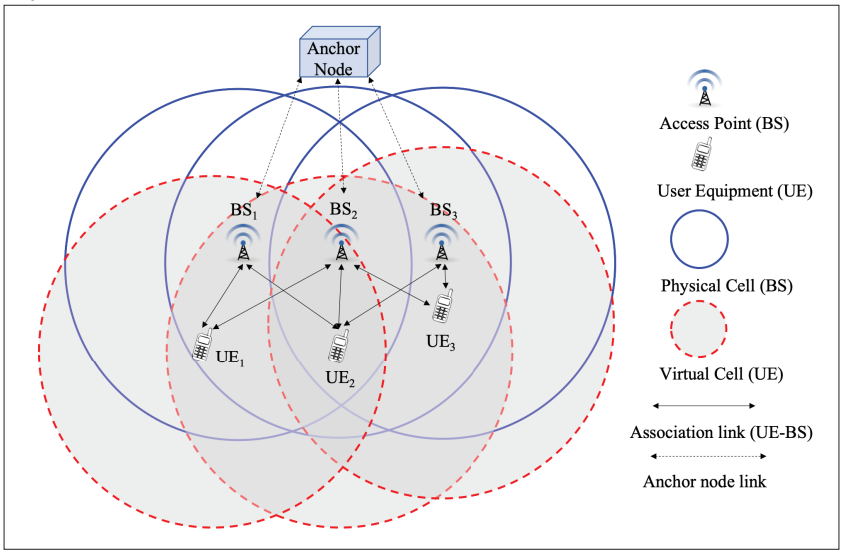

FIGURE 4. Proactive Cell Association: Each UE connects to one or more BS(s), dependent on the Signal-to-Interference-Plus-Noise Ratio (SINR). The set of signals received by each BS comprising the uplink Single Frequency Network (SFN) is forwarded to the anchor node for Multi-User Detection and Selective Combining. By associating with multiple BSs, the number of network paths a UE's signal transmits on increases, as does the UE's maximum reliability.

The dilemma in proactive multi-cell association comes from the risk of an unacceptable outage probability and how Selective Combining (SC) and Multi-User Detection (MUD) are handled via a multi-cell anchor. As proactive communication uses the same RB among multiple BS, an uplink Single Frequency Network (SFN) must be established.

Traditional 5G, per standard, lacks the requisite hardware configuration for multi-cell MUD and selective combining. Considering that multiple O-eNB PHYs must synchronize to process the received signal, O-RAN's function virtualization can effectively support multi-cell MUD and SC via a hardware anchor by utilizing the Near RT RIC and RT RIC control loops and via custom proactive cell xAPPs and zAPPs.

Dynamic Spectrum Sharing

For this application, we consider the Mission Critical requirement of addressing End-to-End QoS in dense networks [1] via delay-tolerant networks for IoT. A primary UE's (a primary UE is guaranteed access to the entire spectrum) wireless spectrum, while scarce, is often underutilized. While long-term underutilized spectrum, i.e., whitespace, can be repurposed to usable spectrum by either using cognitive radio or resold to secondary UEs (secondary UEs do not have guaranteed access to the entire spectrum), the same cannot be said for short-term underutilized spectrum as it would require the network to be reconfigured nearly every radio frame. If the transmit power of the BS is reduced while ensuring that the primary downlink UEs (inner UE and outer UE) can still receive their data at the same modulation and coding scheme, then the spare power-domain spectrum (difference from the BS's max transmit power to the two primary UEs' power shown in Fig. 5) can be repurposed via power domain Non-Orthogonal Multiple-Access (NOMA) as a delay tolerant network [11] used by a third UE (virtual UE).

This scheme is not normally possible in traditional, standard cellular networks as NOMA whitespace delay tolerant networks are

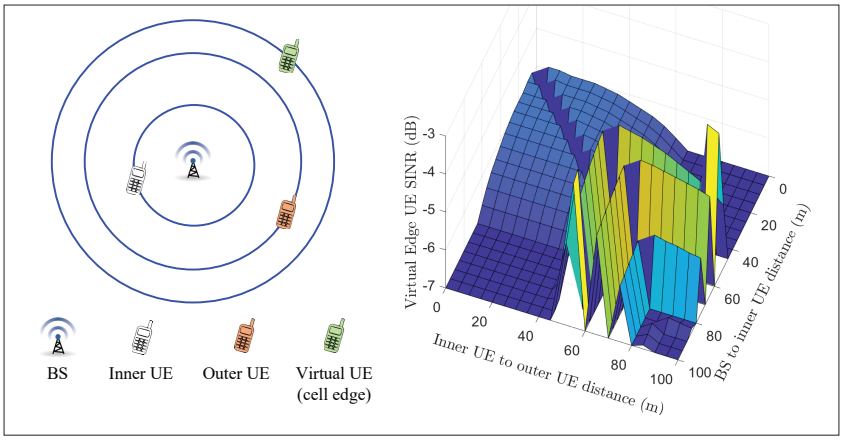

FIGURE 5. Assuming a two-layer Non-Orthogonal Multiple-Access (NOMA) scheme (left) (two UEs share an RB), it can be observed that if each UE's (inner UE and outer UE) transmit power is decreased while maintaining the same modulation and coding scheme relative to the BS, then spare power for an additional UE becomes available (right). The spare power can then be repurposed as a Delay Tolerant Network for a delay-tolerant third virtual, located at the cell edge, UE to use.

	Delay Spread	2.1µs	1.1µs	0.1µs
RB Latency	Cyclic Prefix	Capacity (bits/RB)		
Normal (1ms)	4.69µs	248.1	321.6	933.2
Low (0.5ms)	2.34µs	101.1	248.1	933.2
Very Low (0.25ms)	1.17µs	25.6	101.1	933.2

TABLE 1. RB Capacity when comparing RB Latency vs. Delay Spread: When the RB latency decreases, or the delay spread increases, the RB capacity potentially decreases.

not commercially deployed. Via O-RAN's Near-RT RIC and RT RIC control loops, the necessary PHY (dynamic power control for the primary and secondary UEs) and Medium Access Control (MAC) (computation of the spare RB capacity and RB assignment) can be supported via corresponding xAPPs and zAPPS.

IoT Gateway Malware Detection

For this application, we consider the Mission Critical requirement of network security as it is essential to ensure a network's reliability [1] in dependable mission-critical networks. IoT devices tend to lack malware detection capabilities. Thus to protect IoT devices against threats, such as Botnets like Mirai,[1] a lightweight [12] (due to the high number of potential IoT devices) Network Intrusion Detection System (NIDS) is required for each IoT device/IoT gateway. The corresponding malware detection ML model would be deployed as a per-device service on the attached MEC server so it can quickly isolate infected IoT devices and block incoming internet infections.

Traditional cellular networks do not support MEC-initiated device isolation. With O-RAN near-RT RIC and the MEC control loops, infected devices and incoming malware can quickly be isolated via a MEC ML model and an xAPP.

RIC-Based IoT Malware Detection

In RIC-based IoT Malware detection, we deploy executable file monitors on each IoT device. All executable file signatures, extracted via static analysis, are streamed from each IoT device to a malicious software detection xAPP on the Near-RT RIC that detects suspicious executable files using pre-trained ML models.

For malware detection, we focused on Mirai attacks; in general, Mirai attacks are independent of O-RAN and can attack any IoT-related environment. For our dataset, we used an open-source Mirai botnet tool to perform a telnet scan and sent the network traffic through our MEC server, which captured network traffic. After which, we labeled the dataset network flows as Mirai or benign. The dataset we used has 53,390 Mirai samples and 54,000 benign samples.

The dataset's feature values were extracted by static analysis and then trained by a Support Vector Machine (SVM). When the Near-RT RIC receives the binary file signatures from each IoT device, the xAPP analyzes them. The xAPP informs the O-CU and the MEC to isolate the suspected IoT device(s) when malicious programs are detected. Through experiment deployment of our proposed O-RAN with the MEC system, the detection rate and defense against Mirai malware achieved an accuracy of 98.83%, which can effectively protect IoT devices.

End-to-End Malicious Application Detection Analysis

Malicious attackers can penetrate the vulnerabilities of device systems via weak passwords. By using cracked passwords, a malicious user can obtain access to the device and then send malicious programs through curl or wget to infect the device. Fortunately, with the help of the O-RAN and the MEC server, an xAPP can collect information from the control plane about the UE, while the MEC collects information from the user plane about the UE's data flows to the core network. By monitoring the network traffic, not only can the end-to-end transmission status immediately be known, but also an ML model on the xAPP and MEC can be used to identify whether there are malicious programs and immediately inform the Near-RT RIC to block the infected UEs.

Hybrid AIoT O-RAN Services

Hybrid Artificial Intelligence of Things (AIoT) O-RAN services use separate ML models; one or more models are deployed on the IoT device, and one or more ML models are deployed on either the MEC and/or a RIC to achieve mission-critical objectives.

QoS Flow and RB Latency Minimization

For this application, we consider the Mission Critical requirement of addressing End-to-End QoS in industrial networks [1] via Industrial IoT (IIoT). In 5G the three most common RB geometries are: {0.25ms × 720kHz, 0.5ms × 360kHz, and 1ms × 180kHz} (each RB geometry corresponds to a different RB latency.) In addition, in 5G, the RB geometry is inherently coupled to the QoS class/settings. A shortfall with this inherent coupling is lower latency RBs are more susceptible to the effects of delay spread [13], as shown in Table 1. Consequently, for environments experiencing high delay spread, the QoS flow's latency may increase (due to inter-symbol interference, RB capacity may decrease as a carrier's delay spread increases.) A solution to decrease a QoS flow's latency while not increasing the RB latency more than necessary is to conditionally perform RB geometry migration, i.e., decouple the QoS flow's QoS class from its RB geometry.

By integrating ML, the UE can intelligently inform the base station as to which RB geometry would offer it the highest average bitrate while minimizing the RB's latency. The Near-RT RIC and RT RIC, via UE feedback and the MEC, can dynamically modify the RB geometry in response to the UE's input (UEs cannot directly control radio configurations in O-RAN; thus MEC ML is required to notify the xAPP as to the change request.) After which, the BS can reallocate radio resources among the QoS flows.

Unfortunately, dynamic reconfiguration of the PHY and its RB geometry is not possible in 5G. Yet with O-RAN, the MEC, the Near-RT RIC, and the RT RIC, dynamic on-the-fly, PHY reconfiguration becomes possible via the MEC, Near-RT RIC, and RT RIC control loops via the MEC ML model, xAPPs, and zAPPs.

Intelligent Traffic Safety

For this application, we consider the Mission Critical requirement of addressing low latency [1] Vehicle-to-everything (V2X) via an Intelligent Traffic Safety application. With the popularity of smart devices increasing every day, pedestrians and drivers are becoming more and more distracted. Thus there is an increasingly higher risk of traffic accidents. Computer vision can assist in the detection of distracted pedestrians and drivers and notify them and others around them of an impending accident. Unfortunately, there is no mechanism for 5G networks to notify both the drivers and the pedestrians without specialized software being installed on either a UE or a vehicle.

In [14], cameras placed at intersections act as sensors, which subsequently forward the captured video to a MEC server for analysis and message generation via a 3rd party application. Due to the high reaction time and the number of cameras, current MEC-hosted solutions are unlikely to expediently notify drivers and pedestrians. This is due to the requisite computer vision models, such as MobileNetV2, which can exhaust the MEC server's resources (the bandwidth required for the high-resolution video and/or the processing power required to process the video[2]) when a high number of cameras are connected to a MEC server.

One possible solution is for a hybrid ML solution where camera-enabled AIoT devices process the video locally for potential accidents, then notify an ML model located on a MEC server (UEs and AIoT devices cannot directly send broadcast emergency messages) to transmit an emergency message via the Commercial Mobile Alert System (CMAS) (CMAS messages are received by all cellular users and do not require additional software.) For this system to be successfully deployed, the AIoT device must be able to indirectly inform the distracted driver/pedestrian and others via low latency connectivity. The MEC control loop is required as the MEC ML model needs to generate CMAS messages via the Near-RT RIC's xAPP.

Rogue Base Station Detection

Rogue BSs (RBSs) present a cybersecurity threat as they can convince victim UEs to connect to them via a stronger signal strength. RBSs can subsequently capture sensitive information the UEs provide and possibly deploy Denial of Service (DoS) attacks against UEs, causing them to disconnect from legitimate BSs.

In [15], a UE deployed ML model detects RBSs. The ML model uses the stability of the received Synchronization Signal Block (SSB) signals of the connected and neighboring BSs; an RBS can have a stronger signal than the surrounding environment. If the suspect RBS signal has a higher signal strength than surrounding valid BSs and the sample standard deviation is too high, due to lower quality hardware, the user is alerted. The ML model is trained using an initial dataset collected from local telecom carriers. The O-eNB E2 interface conveys the RBS O-RAN xAPP-trained detection model to each UE (each UE uses the received ML model for RBS detection). Each UE provides signal strength updates to the O-eNB, which the xAPP uses to retrain the RBS detection model. We deployed our RBS system using a Next Unit of Computing (NUC) (Intel i5-6500 3.2GHz with 24GB RAM), a UE (Google Pixel 3), and a server-class computer (i7-12700, 16GB RAM, 1TB storage) using Kubernetes to deploy O-RAN. For our official BSs, we used three local telecom carriers. Compared to Random Forest (RF) and K-nearest neighbors (KNN), we found that Support Vector Machine (SVM) offered our ML models the highest accuracy; all of our ML models achieved an accuracy of > 99%.

The synergy between the UE and the RIC demonstrates how O-RAN supports deployments of control plane ML models. When coupled with MEC, our proposed system offers a holistic approach to cybersecurity, where both the UE and MEC play an active role in the RIC's behavior.

Conclusion

In this article, we presented several 6G requirements as they pertain to mission-critical IoT services. We presented O-RAN, its architecture, its interfaces, and how combining it with MEC offers a potential solution as it already meets many of 6G requirements by design (network disaggregation, open interfaces, AI, 5G compatible) and it enables support for hybrid ML models, which enable diverse deployments. We then introduced multiple applications where AI is only on the network (RIC and possibly MEC) and where AI is both on the UE and the network. We discussed system behavior and data flow via both malware and RBS detection applications and presented results from a O-RAN with MEC testbed and a O-RAN with UE testbed.

Acknowledgement

The authors would like to thank Cheng-Feng Hung, currently a Ph.D. candidate in the Computer Science and Information Engineeering (CSIE) department at National Taiwan University of Science and Technology (NTUST). He assisted us in reviewing and commenting on the O-RAN setup, testbed, and cybersecurity research summary.

References

[1] N. H. Mahmood *et al.*, "White Paper on Critical and Massive Machine Type Communication Towards 6G," University of Oulu, 6G flagship, white paper, June 2020; http://urn.fi/urn:isbn:9789526226781.
[2] P. P. Ray *et al.*, "A Vision on 6G-Enabled Nib: Requirements, Technologies, Deployments, and Prospects," *IEEE Wireless Commun.*, vol. 28, no. 4, 2021, pp. 120–27.
[3] X. Hou *et al.*, "Edge Intelligence for Mission-Critical 6G Services in Space-Air-Ground Integrated Networks," *IEEE Network*, vol. 36, no. 2, 2022, pp. 181–89.
[4] A. Garcia-Saavedra and X. Costa-Perez, "O-RAN: Disrupting the Virtualized RAN Ecosystem," *IEEE Commun. Standard Mag.*, vol. 5, no. 4, Dec. 2021, pp. 96–103.
[5] M. Polese *et al.*, "ColO-RAN: Developing Machine Learning-based xApps for Open RAN Closed-loop Control on Programmable Experimental Platforms," *IEEE Trans. Mobile Computing*, 2022, pp. 1–14.
[6] A. S. Abdalla *et al.*, "Toward Next Generation Open Radio Access Networks: What O-RAN Can and Cannot Do!" *IEEE Network*, vol. 36, no. 6, 2022, pp. 206–13.
[7] O-RAN ALLIANCE Working Group 1, "O-RAN Architecture Description 7.0," O-RAN ALLIANCE, Technical Specification, 2022; https://www.o-ran.org.
[8] M. Polese *et al.*, "Understanding O-RAN: Architecture, Interfaces, Algorithms, Security, and Research Challenges," *IEEE Commun. Surveys & Tutorials*, Jan. 2023.
[9] R. Kaliski and Y.-H. Han, "Socially-Aware V2X QoS for NOMA Dual-Connectivity," *2021 IEEE 94th Vehic. Tech. Conf. (VTC2021-Fall)*, 2021, pp. 1–5.
[10] K.-C. Chen *et al.*, "Ultra-Low Latency Mobile Networking," *IEEE Network*, vol. 33, no. 2, 2019, pp. 181–87.
[11] R. Kaliski, "Green NOMA M2M," 2021 *IEEE Global Commun. Conf. (GLOBECOM)*, 2021, pp. 01–07.
[12] C. A. Fadhilla, M. D. Alfikri, and R. Kaliski, "Lightweight Meta-Learning BotNet Attack Detection," *IEEE Internet of Things J.*, vol. 10, no. 10, 2023, pp. 8455–66.
[13] R. Kaliski, "5G QoS Flow Migration Over URLLC Relays," *Int'l. Symp. Personal, Indoor and Mobile Radio Communications (PIMRC)*, 2021, pp. 998–1004.
[14] I. Lujic *et al.*, "Increasing Traffic Safety with Real-Time Edge Analytics and 5G," *Proc. 4th Int'l. Wksp. Edge Systems, Analytics and Networking*, ser. EdgeSys '21. New York, NY, USA: Association for Computing Machinery, 2021, p. 19–24.
[15] J.-H. Huang *et al.*, "Developing xApps for Rogue Base Station Detection in SDR-Enabled O-RAN," *Proc. IEEE Infocom 2023 NG-OPERA*, May 2023, pp. 1–6.

Biographies

Rafael Kaliski [M] (rkaliski@ieee.org) obtained his Ph.D. degree in Electrical Engineering from National Taiwan University in 2017. He is currently an assistant professor in the department of Computer Science and Engineering at National Sun Yat-sen University (Taiwan). His research interests include wireless networks, multimedia, resource allocation, cybersecurity, and optimization (Game Theory, Artificial Intelligence, and mathematical).

Cheng-Feng Hung [M] (D10915002@mail.ntust.edu.tw) received his B.S. degree in information technology and applications the college of science and engineering from the National Quemoy University, Kinmen, Taiwan, in 2019. He is currently a Ph.D. candidate in the Department of Computer Science and Information Engineering, National Taiwan University of Science and Technology, Taipei. He visited the Warsaw University of Technology in 2022. His research interests is mobile network security.

Shin-Ming Cheng [M] (smcheng@mail.ntust.edu.tw) received his B.S. and Ph.D. degrees in Computer Science and Information Engineering from the National Taiwan University, Taipei, Taiwan, in 2000 and 2007, respectively. Since 2012, he has been on the faculty of the Department of CSIE, National Taiwan University of Science and Technology, Taipei, where he is currently a professor. He was also a joint appointment research fellow with the Research Center for Information Technology Innovation, Academia Sinica, Taipei. His current interests are mobile network security and IoT system security. Recently, he investigates malware analysis and AI robustness. He has received IEEE Trustcom 2020 best paper awards.

Footnotes

[1] Mirai employs Distributed Denial of Service (DDoS) attacks on the victim.
[2] The higher resolution the video is, the more accurate the detection. This comes at the cost of higher bandwidth consumption and higher processing demand.