IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 14 December 2023, accepted 29 December 2023, date of publication 5 January 2024,
date of current version 19 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350513

RESEARCH ARTICLE

Interference-Aware Intelligent Scheduling for Virtualized Private 5G Networks

BERK AKGUN, DEEPAK SINGH MAHENDAR SINGH, SAMATHA KOTLA, VIKAS JAIN, SAKSHI NAMDEO, RUPESH ACHARYA, MURUGANANDAM JAYABALAN, ABHISHEK KUMAR, VINAY CHANDE, ARUMUGAM KANNAN, JALAJ SWAMI, YITAO CHEN, JOHN BOYD, AND XIAOXIA ZHANG

Qualcomm Technologies Inc., San Diego, CA 92121, USA

Corresponding authors: Berk Akgun (bakgun@qti.qualcomm.com), Deepak Singh Mahendar Singh (dmahenda@qti.qualcomm.com), and Vikas Jain (vjain@qti.qualcomm.com)

ABSTRACT Private Fifth Generation (5G) Networks can quickly scale coverage and capacity for diverse industry verticals by using the standardized 3rd Generation Partnership Project (3GPP) and Open Radio Access Network (O-RAN) interfaces that enable disaggregation, network function virtualization, and hardware accelerators. These private network architectures often rely on multi-cell deployments to meet the stringent reliability and latency requirements of industrial applications. One of the main challenges in these dense multi-cell deployments is the interference to/from adjacent cells, which causes packet errors due to the rapid variations from air-interface transmissions. One approach towards this problem would be to use conservative modulation and coding schemes (MCS) for enhanced reliability, but it would reduce spectral efficiency and network capacity. To unlock the utilization of higher efficiency schemes, in this paper, we present our proposed machine-learning (ML) based interference prediction technique that exploits channel state information (CSI) reported by 5G User Equipments (UEs). This method is integrated into an in-house developed Next Generation RAN (NG-RAN) research platform, enabling it to schedule transmissions over the dynamic air-interface in an intelligent way. By achieving higher spectral efficiency and reducing latency with fewer retransmissions, this allows the network to serve more devices efficiently for demanding use cases such as mission critical Internet-of-Things (IoT) and extended reality applications. In this work, we also demonstrate our over-the-air (OTA) testbed with 8 cells and 16 5G UEs in an Industrial IoT (IIoT) Factory Automation layout, where 5G UEs are connected to various industrial components like automatic guided vehicles (AGVs), supply units, robotics arms, cameras, etc. Our experimental results show that our proposed Interference-aware Intelligent Scheduling (IAIS) method can achieve up to 39% and 70% throughput gains in low and high interference scenarios, respectively, compared to a widely adopted link-adaptation scheduling approach.

INDEX TERMS 5G, private network, industrial Internet of Things, machine learning, intelligent scheduling, open RAN, RAN virtualization.

I. INTRODUCTION

Industrial operators are seeking flexible manufacturing solutions and agile logistics operations so that their businesses can be more responsive to shifts in customer and supply chain issues, and reduce downtime. To achieve that, network infrastructure in industrial ecosystems needs to migrate from wired to wireless, but without compromising on the quality of connectivity that industries have come to expect from wired systems. Fifth Generation (5G) New Radio (NR) emerges as a reliable alternative, not only bringing ultra-reliable wireless connectivity to help industrial operators achieve untethered and flexible processes, but it also does so with a global ecosystem and without proprietary solutions.

5G NR supports industrial operators with a broad choice of spectrum options so they can deploy their own private

The associate editor coordinating the review of this manuscript and approving it for publication was Stefano Scanzio.

5G NR networks for the optimized reliability and data security that a private network can provide. 5G NR also brings ultra-reliability and supports deterministic delay with variety of novel techniques such as Coordinated Multipoint (CoMP), Packet Data Convergence Protocol (PDCP) packet duplication, and time-sensitive networking [1], [2], which are not readily available in other wireless technologies like WiFi. This helps industrial systems maintain synchronization and avoid disruption in mission-critical applications.

Traditional RAN architectures, which are widely adopted in earlier generations of wireless networks starting with 2G, have relied on monolithic solutions. Now, Open RAN (O-RAN), Virtualized RAN (vRAN), and Cloud RAN (C-RAN) revolutionize the wireless networks by turning the traditional proprietary baseband units at the base of radio towers into a fully flexible, scalable, intelligent, and cost-efficient solution. Earlier efforts such as C-RAN helped consolidate the baseband functions on a smaller number of sites across the mobile network. Such a network implementation where radio signals from geographically distributed antennas are collected by remote radio heads and transmitted to the cloud platform through an optimal transmission system has been discussed in [3] and [4]. In this paper, we focus on O-RAN functional split architecture where various network functions run as virtual instances deployed on off-the-shelf components and can be effectively utilized to address challenging 5G/6G problems. In particular, we highlight the network capabilities that we developed to help industries build flexible, scalable, and intelligent high performance 5G NR networks for industrial and enterprise requirements. Towards this goal, we developed the Next Generation Radio Access Network (NG-RAN) research platform, based on the disaggregated O-RAN framework which promotes a broad connectivity ecosystem with standardized interfaces among network functions [5], [6], [7], [8]. This NG-RAN research platform is disaggregated and enables dense network deployments with open, intelligent, virtualized, and interoperable architecture which accelerates technology innovation and showcases leading-edge end-to-end wireless technology solutions of Qualcomm Technologies.

Note that developing intelligent NG-RAN systems with Machine Learning (ML) based solutions is considerably complex, technically challenging, and open to different interpretations. For these reasons, 3GPP focuses on use cases with relatively *simpler* ML solutions that can still deliver tangible gains while being fully interoperable across multi-vendor networks [9], [10]. With this in mind, in this paper, we propose an ML-based interference prediction technique and demonstrate how to integrate such a technique into NG-RAN research platform. First, we developed a Real-Time RAN Intelligent Controller (RIC) component which resides in the Distributed Unit (DU) and runs the ML inference application for predicting over-the-air interference. This machine learning based algorithm utilizes relevant RAN measurements to determine the best Modulation and Coding Scheme (MCS) while maintaining the required target block error rate (BLER). We developed the real-time closed loop control operation between the Real-Time RIC and DU scheduler towards achieving the end-to-end Interference-aware Intelligent Scheduling (IAIS) operation. With this scheme, we aim at addressing challenging dense private network deployments in which network traffic increases with diverse service demands, including quasi-periodic traffic profiles. Such profiles include Extended Reality (XR) applications with periodic file arrivals and Industrial Internet-of-Things (IoT) applications where IoT devices are deployed to perform the same tasks continuously and generate the same amount of data periodically. This causes periodic and intense interference to and from neighboring cells, reducing the network performance. With the flexible configuration options of 5G NR and with Channel State Information Reference Signal (CSI-RS) based Downlink (DL) scheduling, this problem can be mitigated up to some extent [11]. However, these methods often result in conservative or aggressive scheduling especially under the presence of quasi-periodic and bursty interference. On the other hand, IAIS technique that we developed utilizes the channel more efficiently in dynamic environments.

To show that, we established an indoor over-the-air (OTA) testbed with a realistic Industrial IoT (IIoT) Factory Automation layout. Our end-to-end OTA private network deployment comprises of 8 FR1 radio cells realized using the NG-RAN research platform and 16 UEs that provide 5G NR radio link connectivity to a variety of end industrial devices such as a conveyor belt system, a Qualcomm® Robotics RB5 Development Kit with an integrated camera [12], [13], Automated Guided Vehicles (AGV), and programmable logic controller (PLC) components. In this network, we show performance gains of up to 70% in terms of effective throughput. Also, our experimental evaluation indicates that deviation of user-experienced Signal to Interference and Noise Ratio (SINR) is limited compared to a baseline solution, leading to more efficient channel utilization.

The main contributions of this study are as follows:

- We demonstrate our advanced scheduling method, called IAIS, that exploits ML-based interference prediction in a private network environment and significantly enhances system capacity.
- We showcase a realistic factory automation use case in an indoor OTA testbed with Qualcomm Technologies' next generation disaggregated and virtualized RAN research platform, satisfying the connectivity requirements of the IIoT applications that run on our factory automation environment.
- Through empirical analysis by utilizing our OTA testbed at sub-6 GHz frequency, we verify the efficiency of IAIS in terms of effective throughput and deviation in user experienced SINR. Our experiments show that IAIS can improve the throughput by up to 70%, compared to CSI-RS based DL Link Adaptation (LA) scheduling.

The rest of the paper is organized as follows. Section II provides an overview on RAN disaggregation. In Section III,

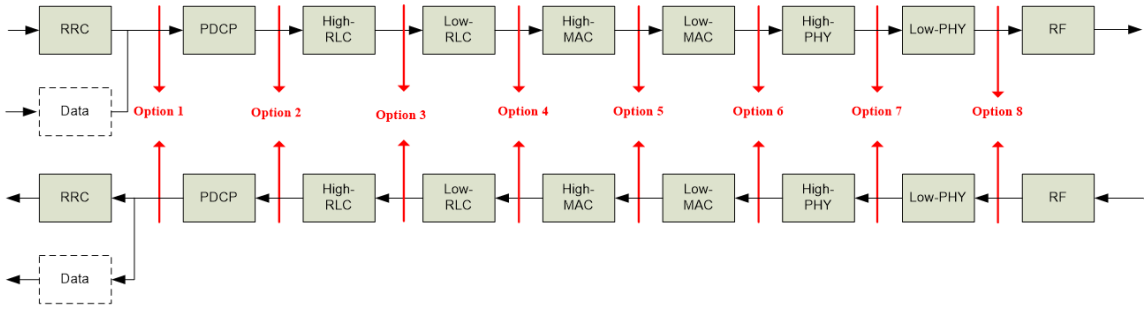


FIGURE 1. Functional split options between the central and distributed units [14].

we present the in-house developed NG-RAN research platform. IAIS algorithm is formulated and analyzed in Section IV-C with a comparison to the baseline model. Section V describes our private 5G network OTA testbed. We share our OTA performance evaluation in Section VI. The paper is concluded in Section VII.

II. RAN DISAGGREGATION OVERVIEW

RAN disaggregation, as the name implies, refers to partitioning conventional monolithic or all-in-one base station solutions into modular network functions (NF) that interoperate through open and standardized interfaces. Disaggregating RAN into modular network functions, in conjunction with software-defined networking (SDN) principles (e.g., with control plane and user plane separation), enables creation of flexible, scalable, and cost-effective radio networks which diversifies RAN vendor ecosystem and promotes innovation.

Network function virtualization (NFV) is another important paradigm in the context of evolving 5G RANs, as it decouples the NF from the underlying hardware and leverages standard virtualization technologies for deployment of these NFs on the commercial off-the-shelf (COTS) equipment (e.g., x86 based hardware servers). NFV enables flexibility in network deployment, efficient utilization of hardware infrastructure resources with adaptation to dynamically changing NF workloads, and faster new service deployment cycle with reduced capital expenditure (CAPEX) and operational expenses (OPEX).

RAN disaggregation, coupled with virtualization, enables cloud-native based flexible, resilient, and highly reconfigurable deployments that can support diverse use-cases and different traffic profiles with integration of data-driven AI/ML optimized closed-loop control for RAN.

3GPP technical report (TR) 38.801 [14] identified various functional splits for partitioning or disaggregating the RAN protocol functions into Central Units (CU) and DU. Fig. 1 (reproduced from [14]) highlights these functional splits, each of which have varying implications on deployment complexity in terms of transport bandwidth, latency, and network performance.

3GPP technical specification (TS) 38.401 [8] for 5G NR defined overall architecture of NG-RAN, which consists of set of gNB's connected to the 5GC through the NG interface. In addition to supporting conventional monolithic architecture, each gNB is logically disaggregated into CU and one or more DUs at the Option 2 functional split (i.e., at the PDCP). Option 2 split is also referred to as high-layer split (HLS) with CU and DU connected via F1 interface [15], [16].

CU is further logically disaggregated into CU-CP (CU-Control Plane) and CU-UP (CU-User plane) entities. This logical split decouples the control plane and user plane in accordance with the SDN principles and enables deployment flexibility and independent scaling. CU-CP hosts the RRC [17] and the control plane part of PDCP [18] and is responsible for mobility management, radio bearer, and admission control. CU-UP hosts the user plane part of PDCP protocol and Service Data Adaptation Protocol (SDAP) [19] and is responsible for user plane packet processing operations (e.g., packet re-ordering, duplicate detection, ciphering, etc.) and traffic flow Quality of Service (QoS) management.

DU hosts Radio Link Control (RLC) [20], Medium Access Control (MAC) [21], and PHY layers and can support multiple cells. DU is responsible for lower layer functions such as cell radio resource management, scheduling, HARQ retransmissions, ARQ, and PHY-layer operations (e.g., baseband encoding, modulation, FFT operations, etc.). RF functions such as Digital-to-Analog Conversion (DAC), signal conditioning, and OTA transmission/reception of radio signals are also implemented as part of the DU.

O-RAN Alliance [22] and Small Cell Forum (SCF) [23] further extended the functional disaggregation with specification of lower layer split (LLS) option 7-2 and option 6, respectively, which relocates one or more PHY layer functions along with RF functions to a separate entity, namely Radio Unit (RU). O-RAN WG4 [6] provides the fronthaul specification for LLS option 7-2, and SCF FAPI provides the MAC-PHY interface specification for LLS option 6 [24], [25]. Fig. 2 shows comparative view of the functional disaggregation architectures as specified in 3GPP, O-RAN Alliance, and SCF.

Separation of certain PHY-layer and RF-layer functions in RU enables potential cost reduction with low complexity radio units and pooling gains with centralized management of remote radio units. However, LLS introduces increased fronthaul transport bandwidth and relatively more latency-stringent requirements compared to HLS alone.

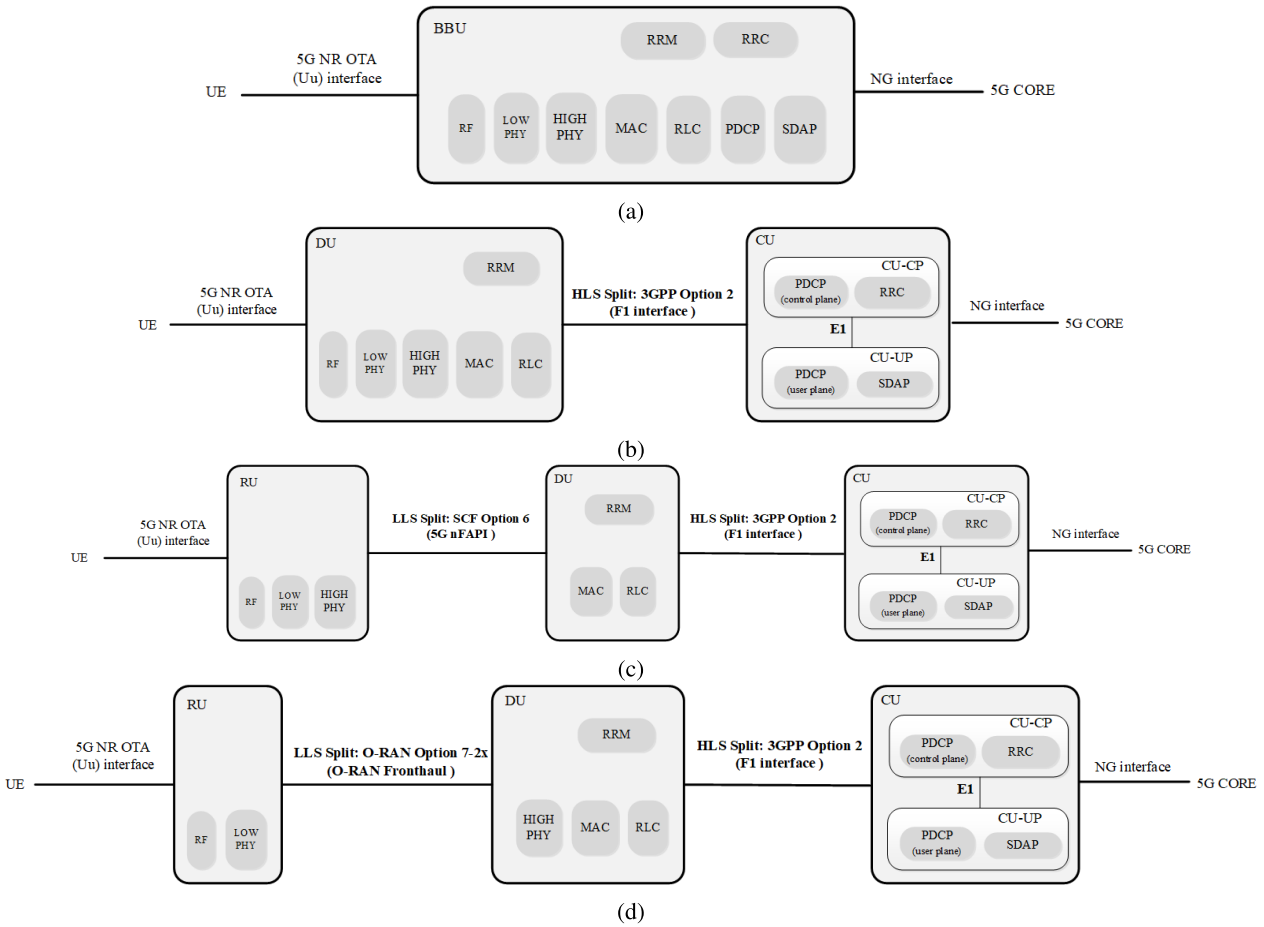


FIGURE 2. RAN split architectures for 3GPP, O-RAN, and SCF. (a) Conventional monolithic RAN architecture, (b) 3GPP defined 5G NR RAN HLS split architecture, (c) SCF defined 5G NR RAN HLS-LLS split architecture, (d) O-RAN defined 5G NR RAN HLS-LLS split architecture.

Optimal choice of LLS is dependent on deployment scenarios, requirements for supported end-to-end user traffic services, and feature requirements (e.g., multi-TRP joint transmission and reception that benefit from joint PHY baseband operations with Option 7-2 LLS).

O-RAN Alliance introduced RIC function which represents programmable components hosting AI/ML aided algorithms. This enables closed-loop control and optimization of various functions in the RAN. RIC processes measurement data provided by the RAN nodes and leverages AI/ML algorithms to optimize various aspects of RAN, e.g., handovers, scheduling policies, RAN slicing, etc. O-RAN Alliance specifies non-real-time (non-RT) RIC [26], which is a component of Service and Management Orchestration (SMO) framework and aids the near-real-time (near-RT) RIC for RAN optimization on a time scale larger than 1 second. The near-RT RIC [27] is typically deployed at the edge of the network and interacts with DU's and CU's for RAN optimization control loops on a time scale between 10 msec and 1 sec. The near-RT RIC hosts multiple applications or microservices (xApps) that encapsulate customized RAN optimization logic, and communicates through standardized interfaces and service models. Fig. 3 shows high-level O-RAN component architecture and the various RIC control loops associated with non-RT and near-RT RIC components.

CU and DU functions can be deployed as virtualized network functions (VNF) on COTS servers at the network edge (with hardware acceleration for some of the PHY-layer functions). RU functions, being RF centric, are typically considered as physical network functions (PNF) and are generally implemented on FPGAs (Field Programmable Gate Arrays) or ASICs (Application Specific Integrated Circuits) and are deployed close to RF antennas.

III. NEXT GENERATION RAN (NG-RAN) RESEARCH PLATFORM

Fig. 4 shows the various components along with the interfaces that comprise the NG-RAN research platform. The NG-RAN research platform is disaggregated, 3GPP interface compliant, virtualized, and implements Option 2 HLS (i.e., the mid-haul between CU and DU) and Option 6 LLS (i.e., the front-haul between DU and RU). NG-RAN CU, NG-RAN DU, and NG-RAN RU refer to the CU, DU, and RU components of the NG-RAN research platform, respectively, for the rest of this paper.

NG-RAN CU realizes 3GPP-defined CU-CP and CU-UP functions as a virtualized NF and is deployed as containerized

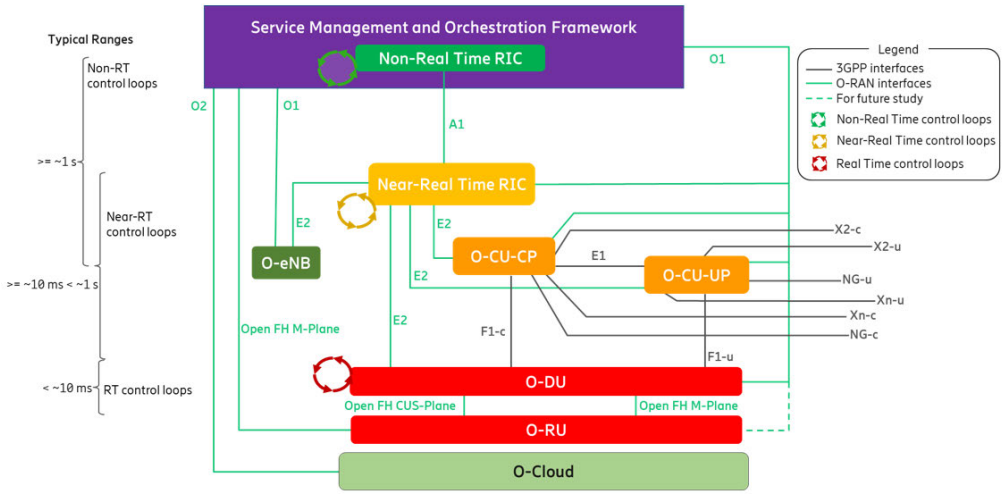


FIGURE 3. O-RAN architecture and RIC control loops [5].

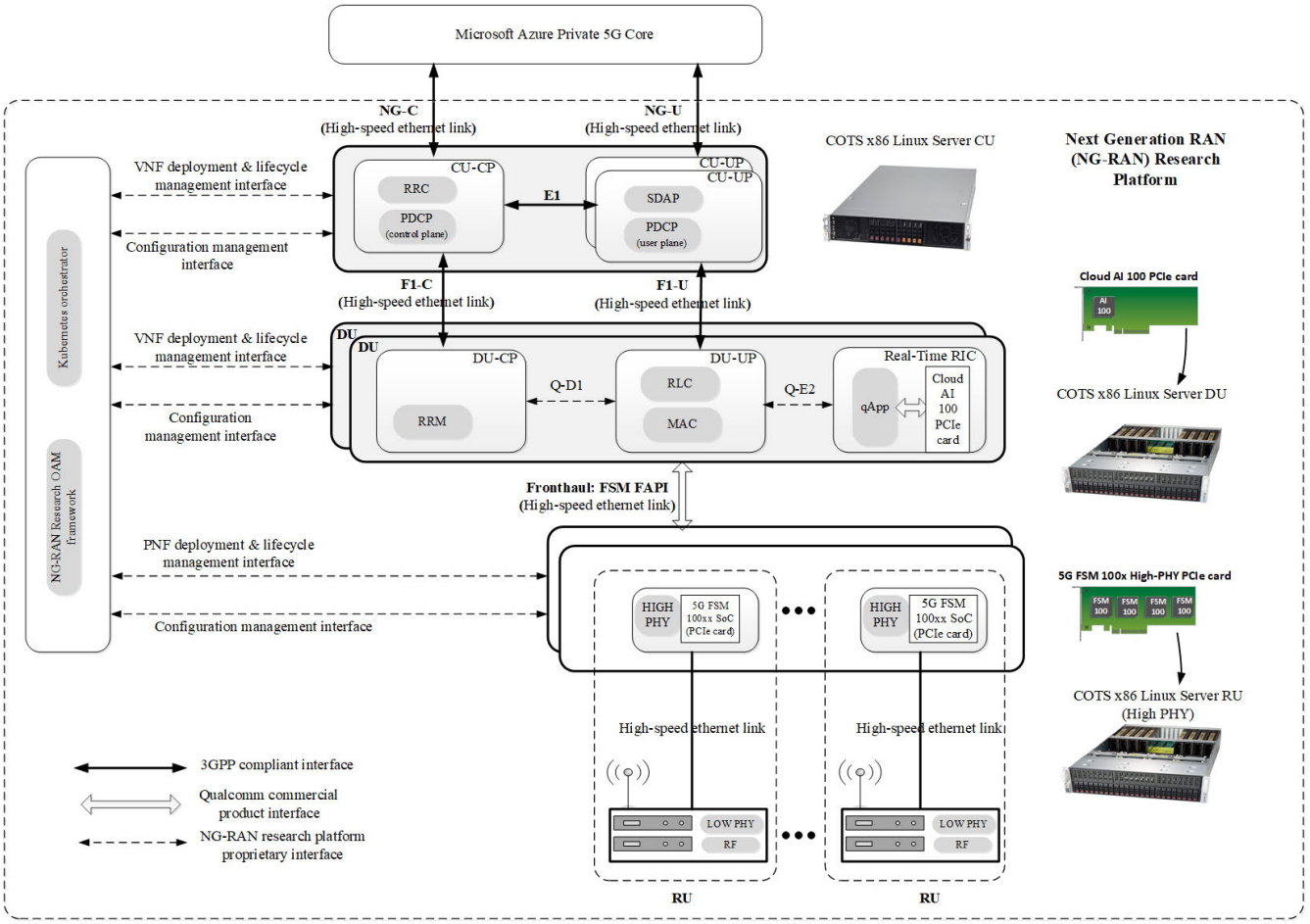


FIGURE 4. Next Generation RAN (NG-RAN) research platform.

workloads that run on on-premises COTS x86-based Linux servers. CU-CP and CU-UP communicate with each other using 3GPP-defined E1 protocol [28]. CU is connected to on-premises Microsoft Azure Private 5G Core platform over high-speed ethernet network and communicates using 3GPP-compliant NG-C / NG-U protocols [29], [30].

NG-RAN DU is further disaggregated into DU-Control Plane (DU-CP) and DU-User Plane (DU-UP) functions.

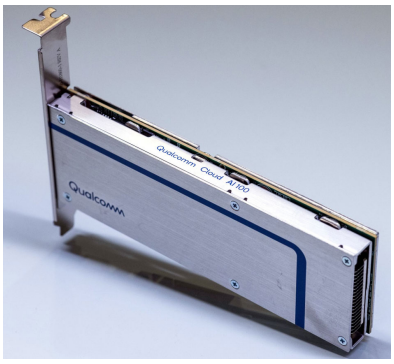


FIGURE 5. Qualcomm Cloud AI 100 PCIe card for high efficiency AI inference in Real-Time RIC.

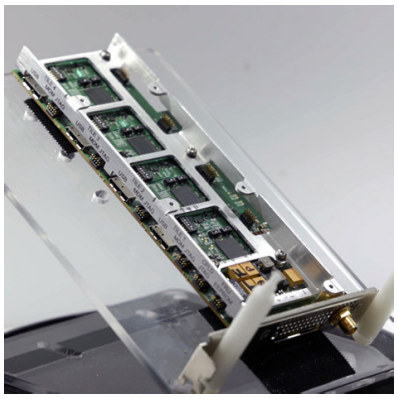


FIGURE 6. Qualcomm FSM100xx based PHY baseband PCIe card.

DU-CP is responsible for cell-level radio resource management and communicates with CU-CP using 3GPP-compliant F1-C protocol [15]. It acts as a controller for DU-UP function. DU-UP is responsible for RLC and MAC protocols and user plane data path over 3GPP-defined F1-U interface with CU-UP [16]. DU-CP and DU-UP are realized as virtualized NF and deployed as containerized workloads that run on the on-premises COTS x86-based Linux servers. They communicate with each other using Qualcomm® proprietary Q-D1 protocol. DU supports the FSM FAPI protocol (i.e., API for the Qualcomm® FSM100xx small cell modem platform product [31]) over high-speed ethernet that provides the fronthaul transport to the RU.

NG-RAN DU implements our proprietary Real-Time RIC component that can host multiple applications, called as qApps. These applications contain customized AI/ML algorithms and provide closed-loop control for optimizing RAN functions on a time scale of less than 10 msec. The Real-Time RIC is embedded within the DU and interacts with the DU-UP component via our proprietary Q-E2 interface (O-RAN E2-like interface). In particular, it receives the necessary RAN measurements to perform optimization and communicates the results to DU-UP. Real-Time RIC utilizes the Qualcomm® Cloud AI 100 PCIe (Peripheral Component Interconnect Express) card [32] for performing AI/ML inference required for RAN optimization. Please refer to Fig. 5.

NG-RAN RU realizes HIGH-PHY function [5] using Qualcomm FSM100xx based PHY baseband PCIe card, which hosts up to four 5G FSM 100xx SoC's [31]. This card can provide HIGH-PHY baseband service for up to four FR1 radio cells. The HIGH-PHY baseband PCIe card (please refer to Fig. 6) is hosted in the on-premises COTS x86-based Linux server and serves as the endpoint for fronthaul (high-speed ethernet) termination with DU. NG-RAN RU utilizes in-house developed FPGA-based components for LOW-PHY and RF functions that are connected to the HIGH-PHY over high-speed ethernet links.

The deployment of VNF containerized workloads for the NG-RAN research platform is performed using Kubernetes orchestrator. Lifecycle management of all the network functions along with the configuration management is performed using in-house developed NG-RAN research OAM framework.

IV. INTERFERENCE-AWARE INTELLIGENT SCHEDULING

A. PROBLEM STATEMENT

5G network traffic increases with diverse service demands that have periodic traffic profiles. In dense private network deployments, this causes periodic and intense interference to and from neighboring cells. Such examples include bursty XR-like traffic and IIoT applications. For instance, a common XR traffic profile consists of file arrivals every 16.66 milliseconds, corresponding to roughly 60 frames per second (fps) video-like traffic and the total offered load would be in the order of 100 Mbps [33]. Furthermore, most use cases in IoT environments are periodic transmissions of messages containing sensor measurements, status, or simple commands. Various IIoT traffic examples were discussed in [34] and [35]. Some common deterministic and periodic traffic scenarios include a pressure sensor sending values to a machine PLC, emergency stop signals from hand-held controllers being sent to another PLC, or periodic controller-to-actuator messages. Besides, there could be non-deterministic but still periodic traffic patterns such as client-server messages between non-synchronized, non-latency-critical applications, e.g., high-resolution cameras sending images to a pattern recognition server.

Let us consider a scenario where a gNB-to-UE link (DL or UL) experiences bursts of interference inline with the aforementioned traffic patterns that is (a) quasi-periodic (learnable), (b) low duty-cycle (occasional), and (c) intense (catastrophic). For simplicity, let us say that the interference levels fall into two categories (High and Low). The sketch in Fig. 7(b) depicts an example where the interference strength is varying over time. A common solution adopted by various cellular wireless technologies (e.g., 4G LTE, 5G NR, etc.) to resolve the issues arising from these scenarios is the DL link adaptation (LA) technique. Briefly, a DL reference signal called CSI-RS is transmitted by a DU, and UEs perform channel and interference estimation based on this signal. These estimations are later reported as

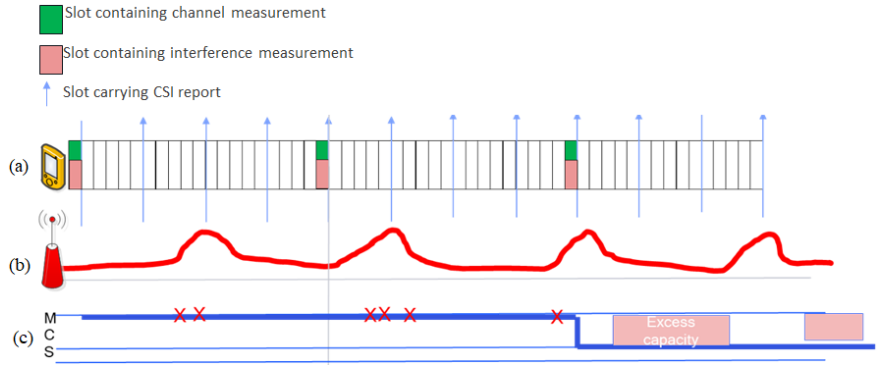


FIGURE 7. (a) Periodic CSI-RS measurements and (c) MCS scheduling under bursty interference where the sketch in (b) represents the interference signal strength. Two suboptimal scenarios arise due to quasi-periodic, low duty-cycle, and intense interference, causing packet decoding errors or under-utilization of the resources.

channel quality information (CQI) reports over a feedback channel to the scheduler residing on DU, which utilizes it to adapt attributes such as MCS, Rank, etc. when scheduling over-the-air transmissions. In Fig. 7, we consider periodic CSI-RS resources and periodic feedback messages, and they are shown with colored boxes and arrows, respectively, in Fig. 7(a). Note that processing and reporting delays are not shown in these figures.

We focus on Enhanced Mobile Broadband (eMBB) like interference patterns with the quasi-periodic arrivals of fixed size files. We focus on two scenarios. First, the interference periodicity is considered to have periodicities larger than the CQI reporting period in order to elicit conditions where interference arrival is a surprise. Second, we consider scenarios with interference periodicities smaller than the CQI reporting period in order to elicit conditions where interference arrival and departure makes the interference conditions change rapidly within a CQI reporting period. Under such bursty interference, the following suboptimal scenarios will be observed when performing baseline LA technique based on periodic CSI-RS measurements.

- Aggressive scheduling: A scheduler that uses CSI measurements that are obtained during low interference will be aggressive when scheduling during high interference bursts. This leads to packet decoding failures and HARQ retransmissions. In Fig. 7, this scenario is depicted with the red cross-marks on MCS curve (Fig. 7(c)), representing packet decoding errors. This happens since the first and second channel measurements are performed under low interference, and the scheduler assigns a high MCS for the subsequent transmissions.
- Conservative scheduling: A scheduler that uses CSI measurements that are obtained during high interference will be conservative when scheduling even after interference has reduced. This leads to inefficient utilization of the channel. In Fig. 7, this scenario is shown with the excess capacity region in MCS curve. Basically, in the third CSI-RS measurement occasion, the channel measurement is performed under high interference, and the scheduler sets a low MCS for all subsequent transmissions.

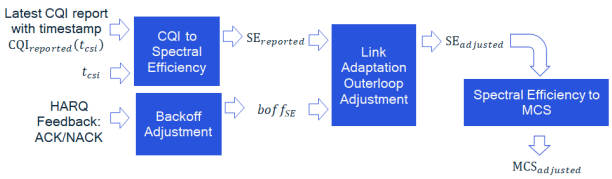


FIGURE 8. Downlink link adaptation operation with outerloop adjustments.

B. BASELINE SOLUTION

To address these suboptimal scenarios, the CQI report processing in 5G NR is augmented with the link adaptation *outerloop* mechanism, where MCS is adjusted based on widely-adopted Hybrid Automatic Repeat Request (HARQ) feedback, which is a special physical layer message that indicates the success or failure of a DL transmission. Briefly, for each DL packet, UE sends an acknowledgement message: ACK if the packet is successfully decoded or NACK if the packet decoding is failed. In the case of a NACK, gNB retransmits the same packet, increasing the reliability of the system. In HARQ feedback, the exponential backoff method is utilized to multiplicatively decrease the scheduled MCS with each corresponding NACK and to gradually increase the MCS with each ACK. Please refer to the following equation, where bler denotes the BLER target, where $0 \leq \text{bler} < 1$. Also, let Δ^{down} and Δ^{up} denote the backoff adjustment step sizes for each ACK and NACK, respectively. Then,

$$\text{boff}_{SE} = \begin{cases} \text{boff}_{SE} - \Delta^{down}, & \text{if HARQ feedback: ACK} \\ \text{boff}_{SE} + \Delta^{up}, & \text{if HARQ feedback: NACK} \end{cases} \tag{1}$$

where boff_{SE} denotes the backoff element in terms of spectral efficiency (please refer to [11] and [36] for the details). This backoff element is then utilized by the scheduler to adjust the MCS. Briefly, this value is subtracted from the spectral efficiency reported by the UE, which is calculated

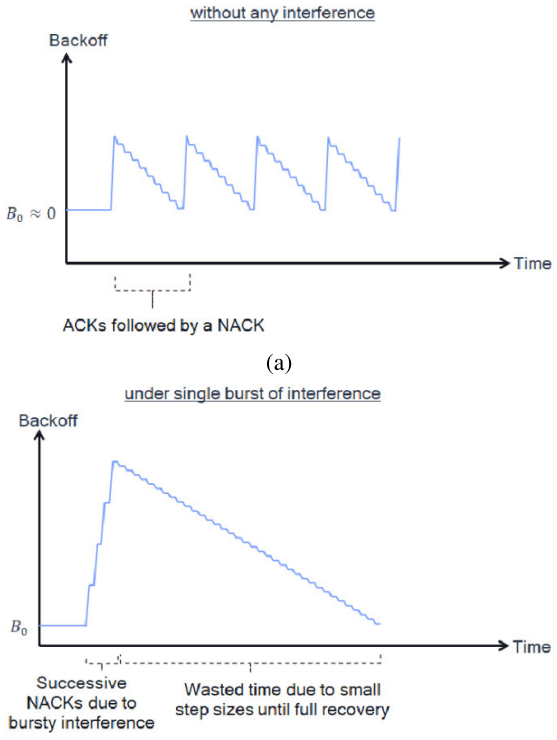


FIGURE 9. Issues with link adaptation outerloop mechanism under bursty interference. (a) With the absence of interference, the backoff parameter ideally remains around 0. (b) Under a quasi-periodic and bursty interference, we observe a sharp increase in backoff due to successive NACKs.

based on CSI-RS measurements. Let $\text{CQI}_{reported}(t_{csi})$ denote the Channel Quality Indicator (CQI) reported by the UE corresponding to CSI-RS at time t_{csi}. The spectral efficiency $\text{SE}_{reported}(t_{csi})$ is then calculated via a look-up table. Let $f_{cqi2se}(x)$ represent this conversion where x is the reported CQI. Then, $\text{SE}_{reported}(t_{csi}) = f_{cqi2se}(\text{CQI}_{reported}(t_{csi}))$ and $\text{SE}_{scheduled}(t_{csi}) = \text{SE}_{reported}(t_{csi}) - \text{boff}_{SE}$, where $\text{SE}_{scheduled}(t_{csi})$ is the adjusted/scheduled spectral efficiency that is later converted to the largest MCS supported by $\text{SE}_{scheduled}(t_{csi})$. Please refer to Fig. 8 for the high-level block diagram of the DL LA procedure.

Δ^{down} in (1) is also referred to as recovery step-size where $\Delta^{down} \geq 0$ and $\Delta^{up} = \Delta^{down} \times (1 - \text{bler})/\text{bler}$. In the absence of a bursty interference, boff_{SE} ideally remains around 0. Please refer to Fig. 9(a) where ACKs gradually reduce the backoff with a step-size of Δ^{down} until a NACK multiplicatively increases it. On the other hand, under a bursty interference, i.e., Fig. 9(b), multiple packet decoding failures happen, causing multiple NACKs. This results in a sharp increase in boff_{SE}. The resources are inefficiently utilized until the system recovers to its ideal operating point where boff_{SE} is around 0. That is, even after the interference is not present anymore, the backoff remains very high, eventually leading to low MCS and hence low throughput. In addition to that, under quasi-periodic patterns, another burst of interference may hit again before the system comes back to its ideal operating point. This causes another sharp increase in boff_{SE}. This process repeats itself due to quasi-periodic nature of the interference, eventually leading to a suboptimal and non-zero equilibrium point for boff_{SE} and causing inefficient utilization of the resources.

C. PROPOSED APPROACH: IAIS

In this section, we explain our proposed *Interference-aware Intelligent Scheduling*, IAIS, technique that addresses performance degradation issues caused by a quasi-periodic, low-duty cycle, and catastrophic interference. Since such an interference pattern is somewhat learnable due to its periodic nature, we study the performance of different machine learning methods to predict various types of interference and to avoid the two suboptimal scenarios mentioned in the previous section. Due to its superior performance in sequence prediction and in learning long-term dependencies in datasets, Long Short-Term Memory (LSTM) based neural networks are considered in this study to predict the interference patterns and proactively utilize this information during MCS adjustment during scheduling. Therefore, the idea that we explore in this paper is to predict interference patterns with an LSTM network and proactively make this information utilized by the scheduler during MCS adjustment step. Note that although the principle can be applied to DL and uplink (UL), we only cover DL here to focus on the key concepts.

Fig. 10 highlights the interactions between the various entities and relevant inputs - outputs for the key functional blocks. Briefly, we implement the LSTM-based neural network (NN) model on the Qualcomm Cloud AI 100 that is hosted in NG-RAN DU. This NN model is executed as part of the qApp which resides within the Real-Time RIC component (refer to Section III) and communicates with the DU DL scheduler over the Q-E2 interface. The overall IAIS technique has the following salient functional blocks:

- SINR predictor: This functional block is essentially the NN model within the Real-Time RIC that takes as an input the sequence of relative timestamped sequence of PDSCH SINR values measured by the UE and outputs a predicted SINR sequence for a consecutive range of future timestamps. Note that the length of the input SINR sequence and the length of the output SINR sequence is denoted by T and N, respectively, in Fig. 10.
- MCS adjustment computation: This functional block resides within the DL scheduler link adaptation module and takes as an input the predicted SINR values from SINR predictor and computes the scheduled MCS (taking into consideration the outerloop backoff as well).

Motivation for interference and SINR prediction: Principal motivation for extensive standardized support of CSI reporting in 5G NR is scheduling and link adaptation. A single CSI report, which is delayed and relatively sporadic (periodic, semi-persistent, or aperiodic), cannot be expected to uniformly and correctly represent an interference that

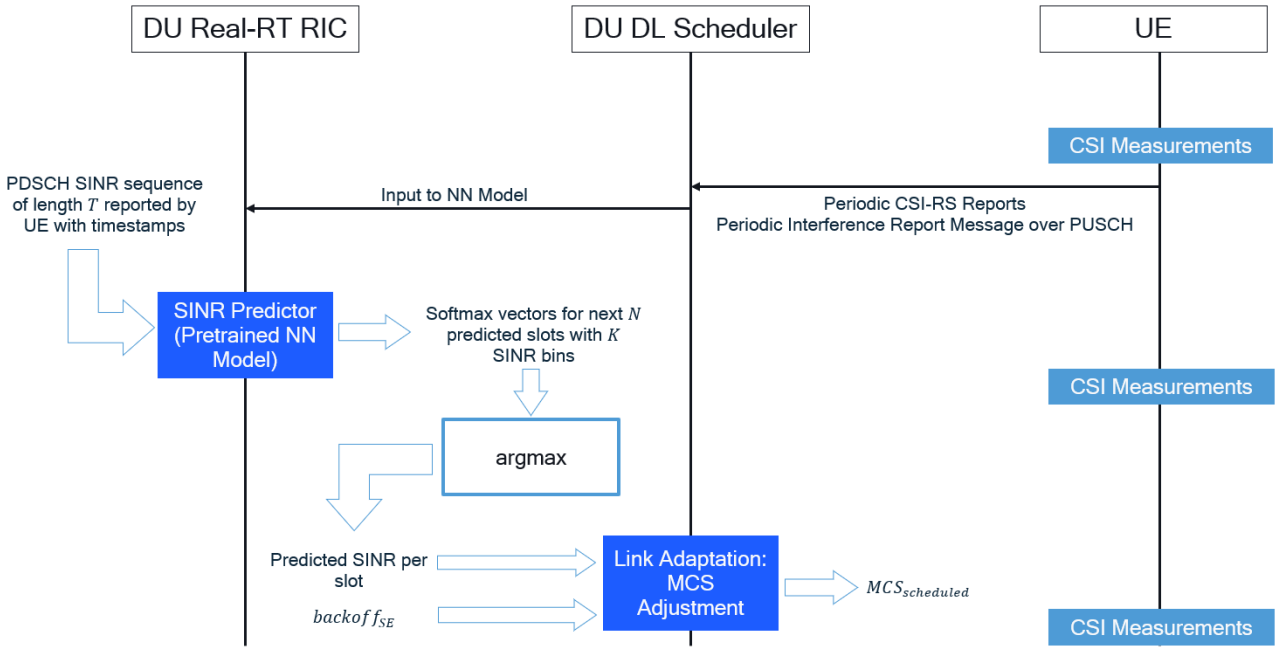


FIGURE 10. End-to-end call flow for IAIS.

varies over multiple slots. This results in a mismatch between reported CSI and actual observed SINR. As argued in Section IV-A, this may result in too aggressive or too conservative MCS selection. The source of the mismatch, namely interference variation, and hence the size of the mismatch can be large from slot to slot. A good predictor that learns the interference process or the SINR process can predict the SINR in a slot where it can potentially have a lower mismatch in predicted value and the true SINR, thereby permitting better scheduling.

The nature of the SINR process is such that a significant variation comes from interference variation, which in turn is strongly governed by the presence and absence of transmissions at the interfering nodes. This presence or absence of transmissions is due to arrivals and departures of traffic demand at the interfering links. Further, the interference process is a superposition of interferers who have different strengths, primarily arising from long term effects such as pathloss and shadowing.

Classical predictors, based on linear models such as autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) [37], are more suited where the primary sources of randomness in the underlying process are gradually varying rather than abrupt. For example, capturing a 25 dB variation in SINR time series due to presence or absence of the nearest interferer in two consecutive slots, via classical ARMA/ARIMA approach may need large order of coefficients. ML Models such as LSTM, on the other hand, appear to be more suitable to capture the gradual as well as the predictably abrupt changes in the time-series.

The predictor is a batch predictor, simultaneously predicting the likelihood (a predicted probability distribution) of quantized value of the SINR in a window of future slots. In this design the quantization of the SINR prediction converts a regression problem to a classification problem. The LSTM based NN Model for the predictor is trained on cross-entropy loss minimization.

The training phase of the SINR predictor module is based on an offline data collection procedure. The details of this procedure will be explained in Section VI. Data collection for training simply consists of running various traffic profiles on the UE under consideration, applying various interference profiles on the UEs in the neighboring cells, and recording the $(\text{SINR}_{measured}(t), t)$ pairs at the worker UE, where t denotes the timestamp and $\text{SINR}_{measured}(t)$ denotes the UE-measured SINR at time t.

The call flow schematic depicted in Fig. 10 starts with the periodic Interference Report messages that carry a set of SINR measurements and timestamps associated with PDSCH measurements for the time window since the last Interference Report message. The DU, via the Q-E2 interface, transports the same information to the RIC where the inference engine is running the SINR predictor. The NN model in Fig. 10 is a set of LSTM cells followed by two fully connected layers. The dimensions involved in the NN Model are summarized in Table 1. In particular, this NN module outputs N number of softmax vectors with K bins each. These bins correspond to a range of SINR values as shown in Table 1. Then, via argmax selection criteria, per-slot SINR predictions are provided for the next N slots by choosing the bins with the highest associated probability for each slot.

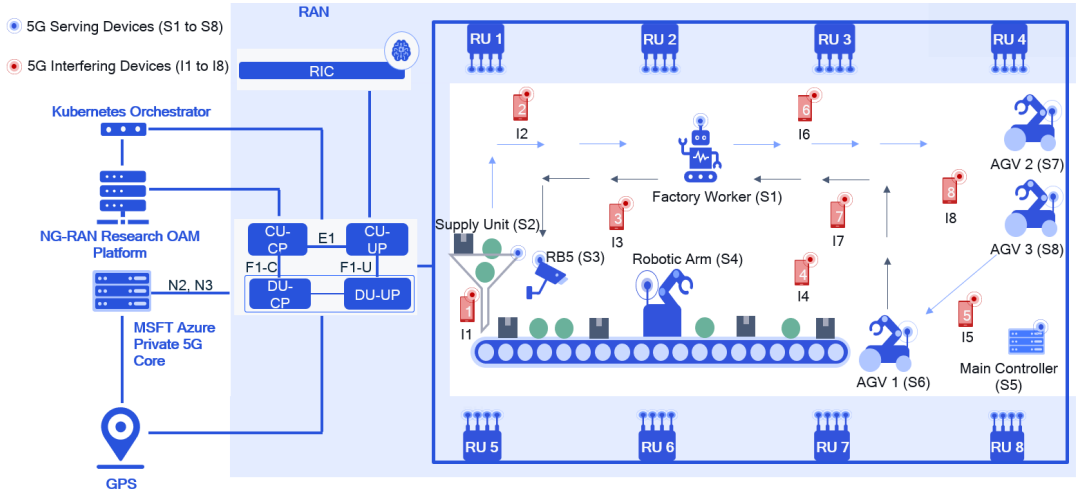


FIGURE 11. IIoT Factory Automation operation with radio network layout of 8 cells and 16 5G UEs.

TABLE 1. Summary of NN model parameters.

Model Parameters	Size Range
Sequence Length for inference (LSTM Lookback) T	40-400
Input:{SINR,encoded timestamp}	SINR:[-15,35] dB with ≈50 bins
Output: Softmax Vectors	100 predicted slots with ≈50 bins
Size of NN model	Hidden size : 32-256
Number of LSTM Layers	1
Number of Fully Connected Layers	2

The time-complexity of such an NN model varies approximately linearly with each of the following: The sequence length or lookback, the output tokens which is a product of the prediction window, and the number of bins. The time-complexity of the LSTM component varies roughly as square of the hidden size. The space-complexity of such a model varies linearly with sequence length and hidden size.

Within the MCS adjustment block, the following operations are performed. The first step is to convert the predicted SINR value $\text{SINR}_{predicted}(t_0)$ for time t_0 to the predicted spectral efficiency $\text{SE}_{predicted}(t_0)$ with a look-up table, i.e., $\text{SE}_{predicted}(t_0) = f_{sinr2se}(\text{SINR}_{predicted})$, where $f_{sinr2se}()$ denotes the conversion from SINR to spectral efficiency. Then, the similar steps as in Section IV-B are applied. Particularly, the adjusted/scheduled MCS is calculated through $\text{SE}_{scheduled}(t_0) = \text{SE}_{predicted}(t_0) - \text{boff}_{SE}$ and $\text{MCS}_{scheduled}(t_0) = f_{se2mcs}(\text{SE}_{scheduled}(t_0))$, where $f_{se2mcs}()$ denotes another look-up table that converts spectral efficiency to scheduled MCS.

V. PRIVATE NETWORK DEPLOYMENT: OTA TESTBED

Our private 5G network is deployed in a warehouse setting in one of the Qualcomm buildings. The high-level diagram of the deployment is shown in Fig. 11. This standalone (SA) 5G private network consists of a commercial Microsoft Azure Private 5GC and the NG-RAN research platform, as described in Section III. The 5GC provides Access and Mobility Management Functions (AMF), Session Management Functions (SMF), and User Plane Functions (UPF). The NG-RAN research platform is connected to Microsoft Azure Private 5GC over 3GPP-compliant N2 [29] and N3 [30] interfaces that support control plane (related to UE context management, PDU session/resource management, etc.), signaling, and exchange of user data between UPF and RAN. The 3GPP N6 interface with UPF terminates in the Qualcomm intranet infrastructure for data network connectivity.

Our private network deployment consists of 8 Radio Units (RU) that provide 8 cell radio network with each of these cells operating in FR1 band, TDD mode, and with 100MHz carrier bandwidth. The HIGH-PHY baseband functionality for these 8 cells is achieved using two Qualcomm FSM100xx based PHY baseband PCIe card hosted in RU servers. The 8 RUs are spread across the OTA testbed to ensure necessary RF coverage for all the 16 UEs that are part of the test-bed. Each RU or cell is configured with DDSU TDD pattern, and the DU scheduler performs both FDM (Frequency Division) and TDM (Time Division) scheduling operation to provide end-to-end connectivity for all the 16 UEs. The 5G UEs, shown Fig. 13, have 2 transmitter (Tx) and 4 receiver (Rx) antennas with 2×4 Multiple Input Multiple Output (MIMO) capability and can support 100 MHz bandwidth operations.

A. IIOT FACTORY AUTOMATION

Our IIoT Factory Automation operation developed on the private 5G network test-bed, represents a typical assembly line scenario, where an Automatic Guided Vehicle (AGV) moves newly minted parts for sorting and another set of AGVs supply the sorted parts for assembly (refer to Fig. 11). The newly minted parts are placed into a conveyor belt, where a camera running an ML algorithm identifies them and instructs a downstream delta robot on how to sort them. The sorted parts are then supplied to an assembly worker by AGVs. All these different units are connected to our

FIGURE 12. NG-RAN RF Radio Unit research platform.

FIGURE 13. Qualcomm® Snapdragon X60 5G Modem based User Equipment [38].

private 5G network through 5G UEs. The various endpoints connected to these 5G UEs are shown in Fig. 11 and are as follows:

- Multiple Qualcomm Robotics RB5 Development Kits (controlling AGVs),
- A Qualcomm Robotics RB5 Development Kit (camera with image classification capabilities) [13],
- A PLC to control the delta robot and the conveyor belt,
- Worker in assembly line that uses the parts supplied,
- A main controller that orchestrates the whole factory operation,
- 5G UEs serving as endpoints for security/monitoring camera traffic.

The 5G UE is connected to Factory Worker, referred as S1 (see Fig. 11), which requires ultra reliable communication, and hence the focus of IAIS performance. Four of the 5G UEs run such a traffic to emulate security/monitoring camera, and they are the interferers to S1. Among these UEs, two of them run traffic equivalent to a 60 fps video stream, and the other two run traffic equivalent to a 30 fps video stream.

VI. OTA PERFORMANCE EVALUATION

A. KEY PERFORMANCE INDICATORS

The following key performance indicators (KPIs) help in analyzing the benefits of IAIS in bursty interference scenarios:

Effective DL Throughput: Effective DL throughput is the accumulated throughput in a time window over the actual number of DL slots used for transmission. The averaging window is chosen as 500 msec. This KPI shows how efficiently the available bandwidth is being used by the scheduler.

Deviation of user experienced SINR: For the baseline approach (CSI-RS based DL Link Adaptation), the deviation of user experienced SINR is calculated as the difference between the CQI reported SINR after outerloop adjustment and the actual PDSCH SINR measured by the S1 worker UE. For the IAIS approach, the deviation is calculated as the difference between the predicted PDSCH SINR by the NN model and the actual PDSCH SINR measured by the S1 worker UE. Lesser deviation (close to zero) indicates the interference prediction is close to ideal and the channel is being used effectively. Positive values indicate that the interference prediction is optimistic, resulting in higher number of packet errors due to aggressive scheduling. Negative values indicate that the interference prediction is pessimistic, resulting in under-utilization of the channel due to conservative scheduling. Tightly bounded distribution around zero suggests that the scheduling is done effectively resulting in efficient channel utilization. A wider distribution of SINR deviation indicates that the scheduling is either conservative or aggressive resulting in inefficient channel utilization.

MCS: DL throughput is directly impacted by the MCS scheduling. Accurate interference prediction helps in MCS adjustment computations resulting in effective utilization of the channel. Choosing a higher or lower MCS than the ideal MCS for a particular DL slot can result in ineffective utilization of the channel due to packet errors or wasted channel capacity.

B. OTA TEST METHODOLOGY

The fully operational OTA network with 8 cells and 16 5G UEs is subjected to interference among the cells, as they all occupy the same 100 MHz bandwidth. In this section, we describe test scenarios where the S1 worker UE is subjected to DL interference from up to 4 other neighboring cells. We evaluate the performance of this S1 worker UE in the presence of semi-persistent bursty interference from up to 4 cells. This interference, therefore, leads to PDSCH BLER at the worker UE. This impact is observed by variations in worker UE's PDSCH SINR. In our OTA analysis, we evaluate and compare the performance of the worker UE's DL throughput while maintaining reliability requirements. The baseline is conventional 5G CSI-RS based DL link adaptation which is compared with IAIS. The specific 5G configuration on the worker and interfering UEs is not covered in this

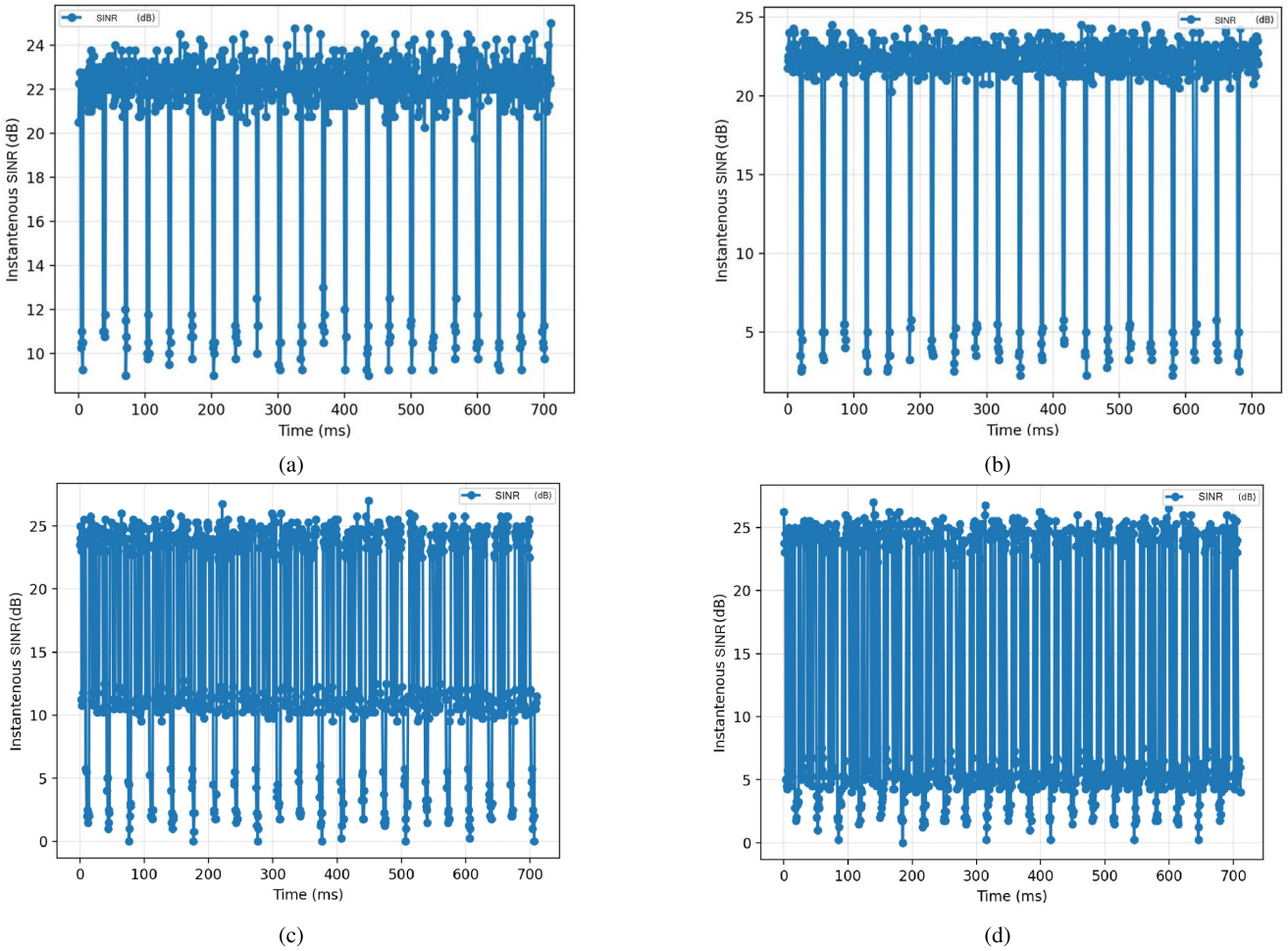


FIGURE 14. Impact of interferers on PDSCH SINR at the worker UE. (a) PDSCH SINR vs time in single interferer scenario from farther RU, (b) PDSCH SINR vs. time in single interferer scenario from nearer RU, (c) PDSCH SINR vs. time in Scenario A: Interference from Cells 2, 3, 5, and 8, (d) PDSCH SINR vs. time in Scenario B: Interference from Cells 2, 3, 6, and 7.

manuscript. However, the DL channel configurations of these UEs are such that the worker UE's DL transmission is impacted by the DL traffic on up to four interferers. S1 worker UE is served by Cell 1. We consider Cells 2, 3, 5, 8, 6, and 7 as Interferers 0, 1, 2, 3, 4, and 5, respectively. Note that even though the set of possible interferers includes six cells, only up to four of them will be active in a given scenario. In particular, in our deployment, interference on worker UE from Cells 6 and 7 (interferers 4 and 5), is stronger than interference on worker UE from Cells 5 and 8 (interferers 2 and 3) as the worker UE is closer to RU6 and RU7 compared to RU5 and RU8. Having six possible interferers with different strengths where we select up to four of them at a given time allows us to create and study various interference scenarios.

1) BASELINE

As explained in detail in Section IV-B, the baseline approach is the CSI-RS based DL link adaptation that is typically utilized in 5G RAN systems. Note that CSI-RS allows for more accurate noise estimation at the UE compared to UL-based reference signals that rely on channel reciprocity between DL and UL, since it enables capturing both interference and noise. Here, UE measures CSI and reports CQI to the DU. Then, the scheduler makes innerloop adjustments through CQI to SE mapping as per 3GPP specifications 38.214 Table 5.2.2.1-3 [11]. SE adjustments are further made based on outerloop backoff depending on HARQ feedback, and MCS is chosen as per 3GPP specifications 38.214 Table 5.1.3.1-2 [11].

2) IAIS

The NN model running in Real-Time RIC is trained (using offline training methods) to enable it to learn the interference pattern on the worker UE. The scenario for training that is chosen on our OTA deployment is as follows:

- Resource utilization: The worker UE is scheduled with 100% resource utilization, i.e., all the DL PDSCH slots are occupied. This allows the model to learn the interference pattern across all the slots.
- Number of interferers: Up to 4 interferers for a given scenario, each occupying the same 100 MHz DL resources.

- Interferer strength: Scenario A is used for training in which Interferers 0, 1, 2, and 3 introduce approximately 15 dB, 9 dB, 12 dB, and 6dB SINR degradation to the worker UE, respectively. We consider these levels as the reference state.
- Interference periodicity: Each interferer has a semi-persistent bursty traffic scheduled at different periodicities. In our scenario, Interferer 0 and Interferer 1 have similar traffic profiles such that it occupies 4 or 5 consecutive PDSCH slots every 33 msec. Interferer 2 and Interferer 3 have similar traffic profiles such that it occupies 4 or 5 consecutive PDSCH slots every 16 msec.
- Traffic pattern: UDP traffic on both the worker and interferer UEs.
- Training data: The following training data is collected from the worker UE under various scenarios listed in the later sections: Periodic wideband CSI-RS reports and PDSCH decoding SINR measurements with timestamps. This training data is input to the inference prediction block which predicts the adjusted MCS for a given DL slot.
- Number of datasets: Each dataset consists of 2-minute captures on DL. Total of 7 datasets, each set having different start reference points of worker and interference traffic, to obtain diversity, are used for training.
- Mode: Offline trained, online inferred.

A few salient challenges for the LSTM based NN Model predictor to overcome are as follows. First, although each interferer is subjected to a periodic traffic, overall the interference process experienced at the worker UE is *multi-periodic* with more than one periodicities of interferers superposed. Second, the traffic at these multiple interferers is not coordinated, i.e., the relative arrivals of interference bursts from multiple interferers over the air vary run to run, and can differ significantly between runs and training data. Further, the periodicities of the interferer traffic are larger than the lookback used for the model, to mimic real-life conditions for time series prediction where such periodicities cannot be known in advance. The trained NN Model, therefore, must perform an online inference in the OTA network, taking into account various delay components in the network. The prediction then interacts with the link adaptation/MCS selection module as outlined in Fig. 10 to perform MCS selection.

The trained NN model for the aforementioned scenarios is utilized to characterize IAIS OTA gains under the following scenarios. Number of RUs and interferers, interference periodicities, and traffic patterns are alike the training scenarios where the first transmission PDSCH BLER target is set to 10%. To characterize the IAIS performance gain, two scenarios are chosen. The first is Scenario A, where, either or all Cells 2, 3, 5, and 8 have DL traffic causing interference on the worker UE. The second is Scenario B, where either or all Cells 2, 3, 6, and 7 have DL traffic causing interference on the worker UE. Cells 6 and 7 in Scenario B have larger interference strength compared to cells 5 and 8 in Scenario A as the stationary worker UE is closer.

C. TEST RESULTS

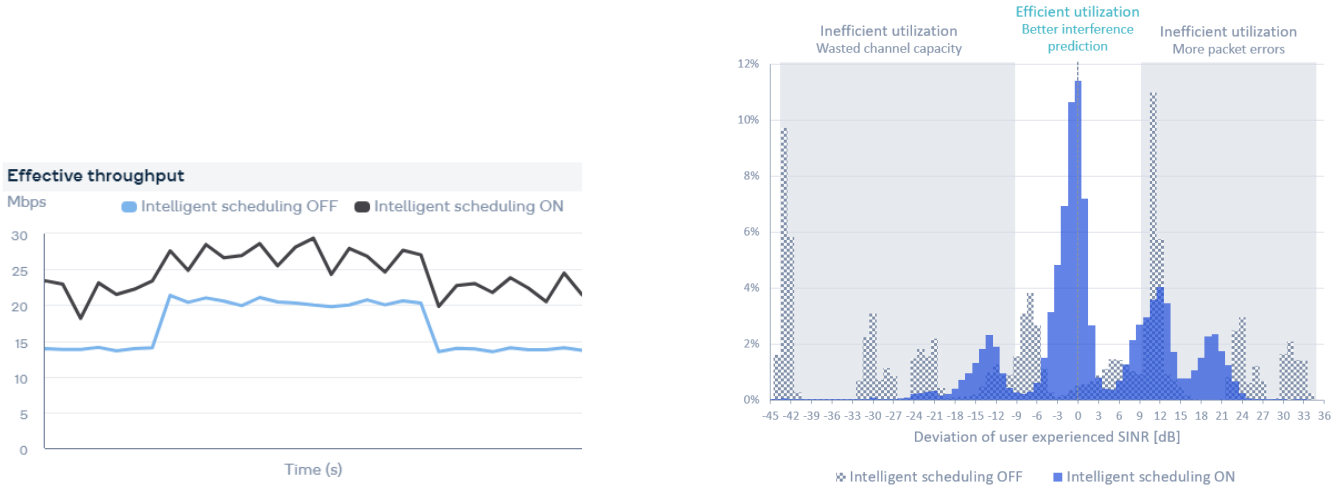
Fig. 14a shows the impact of Interferer 0 on PDSCH SINR at the worker UE. The SINR degrades at the presence of interference traffic occasions. This translates to block errors causing throughput degradation. When the interferers are not present, the PDSCH SINR dips are not seen. Therefore, the DU schedules DL transmissions with an MCS that achieves maximum spectral efficiency (SE) depending on the channel. Fig. 14b shows the impact when Interferer 0 is stronger by 7 dB. Higher interference strength leads to larger SINR degradation on the worker UE. Similarly, when all the four interferers have traffic running, Fig. 14c shows the impact on the worker UE's SINR. As the number of interfering cells increase, we observe more frequent SINR dips on the worker UE. The dips also have varying magnitude because of different distances of stationary worker UE from the interfering cells. We compare the performance of worker UE under baseline and IAIS in two scenarios: Scenario A and Scenario B.

The results in Figs. 15a and 15b show the comparison of DL throughput performance between baseline DL LA and IAIS. IAIS performs better in terms of DL throughput for a given reliability criteria in the presence of semi-persistent bursty interfering traffic compared to baseline DL LA. The gain is largely because the NN model is capable of efficiently predicting the interference occasion and adjusting the spectral efficiency accordingly. In baseline DL LA, the MCS adjustments become either more conservative or aggressive leading to inefficient resource utilization either due to wasted channel capacity or higher BLER, respectively. In Scenario A, there is 39% gain between the performance of DL LA and IAIS for the target of 10% first transmission PDSCH BLER. This is because in IAIS, the NN model is able to estimate SINR and utilize channel capacity. When we compare the baseline performance of Scenario A and Scenario B, we observe the effective DL throughput reduces from 20.3 Mbps to 14 Mbps. This is because of higher interference strength where the corresponding CQI reports map to lower SE. Under such conditions, IAIS provides about 70% gain. This also indicates that the IAIS gain varies and depends on multiple factors. Note that the IAIS approach tends to show increasingly better gains in conditions where the baseline DL LA performance tends to degrade.

The SINR deviation results in Fig. 16b show that the difference between the CSI-RS based SINR calculation and the UE-measured PDSCH SINR is large in both negative and positive directions with baseline DL LA, i.e., when IAIS is disabled. Along the negative axis, when the UE-measured PDSCH SINR is higher, it leads to wastage in channel capacity as the UE is capable of decoding with higher SE on such occasions. Along the positive axis, when the UE-measured PDSCH SINR is lower, the DU schedules PDSCH at larger SE than what the channel can support in the

(a) (b)

FIGURE 15. (a) Scenario A: Average effective DL throughput comparison between IAIS OFF (baseline CSI-RS based DL LA) and IAIS ON for the 5G UE in the OTA deployment, (b) Scenario B: Average effective DL throughput comparison between IAIS OFF (baseline CSI-RS based DL LA) and IAIS ON for the 5G UE in the OTA deployment. Note that the interferer cells in Scenario B have larger interference strength than the ones in Scenario A.



(a) (b)

FIGURE 16. (a) Effective DL throughput vs. time comparison between IAIS OFF (baseline CSI-RS based DL LA) and IAIS ON for the 5G UE in the OTA deployment for both Scenario A and Scenario B, (b) Distribution of deviation of user experienced SINR for IAIS OFF (baseline CSI-RS based DL LA) and IAIS ON cases for the 5G UE in the OTA deployment.

presence of interference, causing BLER and loss of packets. Both cases of inefficient utilization can be mitigated with accurate interference prediction that can be provided by IAIS. The results show that the SINR deviation is comparatively much lower with IAIS. Effective throughput figure in Fig. 16a also provides another perspective of IAIS gain with both high and low interference scenarios. The plot shows Worker UE's DL effective throughput calculated every 0.5 sec. Under both Scenarios A and B, we see IAIS gains.

VII. CONCLUSION

We demonstrated performance gains and efficient use of wireless channel resources that can be realized with Interference-aware Intelligent Scheduling, IAIS, technique in our indoor OTA test-bed private network environment under a realistic factory automation use-case. We also demonstrated the realization of such techniques as part of Real-Time RIC and achieve AI/ML optimized time-stringent closed loop control with the DU scheduler in the fully disaggregated, virtualized, and in-house developed Next Generation RAN research platform. Data-driven AI/ML interference-prediction techniques such as IAIS enable network deployments to extract higher radio link capacity while operating with the required latency and reliability constraints, allowing larger number of devices to be serviced and making 5G NR radio access network deployments cost-effective.

ACKNOWLEDGMENT

The authors would like to thank Chandresh Tiwari, James Wilson, Hai Hong, Sumanth Govindappa, Hemanshu Tyagi, Aaron Bailey, Rahul Ankushrao Kawadgave, Divya Ravichandran, Hetal Pathak, Victor Mai, Yun Lin, Tao Liu, Alisa Moshkovych, Siva Jayaraman, Ronen Yizhaq-Gilad, Vince Baglin, Rajat Prakash, and associated technical teams for their technical expertise and contributions to the work presented in this article. They would also like to thank Microsoft Azure Technical Support Team for their help in provisioning Microsoft Azure Private 5G Core

and interoperability testing during integration with the Next Generation RAN research platform, Vijay Shirsathe, VP of Engineering at Qualcomm Technologies Inc., for his technical guidance and constant support toward development of the Next Generation RAN research platform, and John Smee, SVP of Engineering at Qualcomm Technologies Inc., for his unwavering support and overall direction without which this work would not have been possible.

Snapdragon and Qualcomm branded products are products of Qualcomm Technologies Inc. and/or its subsidiaries. Any opinions, findings, conclusions, or recommendations expressed in this article are those of the author(s) and do not necessarily reflect the views of Qualcomm Technologies Inc.

REFERENCES

[1] *Transforming Enterprise and Industry With 5G Private Networks*, Qualcomm Technologies Inc., San Diego, CA, USA, Oct. 2020. [Online]. Available: https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/transforming_enterprise_and_industry_with_5g_private_networks.pdf

[2] G. Brown. (Jul. 2019). *Private 5G Mobile Networks for Industrial IoT*. A Heavy Reading White Paper Produced for Qualcomm. [Online]. Available: https://www.qualcomm.com/content/dam/qcomm-martech/dm-assets/documents/private_5g_networks_for_industrial_iot.pdf

[3] H. Yang, J. Zhang, Y. Ji, and Y. Lee, "C-RoFN: Multi-stratum resources optimization for cloud-based radio over optical fiber networks," *IEEE Commun. Mag.*, vol. 54, no. 8, pp. 118–125, Aug. 2016.

[4] H. Yang, J. Zhang, Y. Ji, Y. He, and Y. Lee, "Experimental demonstration of multi-dimensional resources integration for service provisioning in cloud radio over fiber network," *Sci. Rep.*, vol. 6, p. 30678, no. 1, Jul. 2016.

[5] *O-RAN Architecture Description 8.0*, document TS O-RAN.WG1.OAD-R003-v08.00, O-RAN Alliance, Mar. 2023.

[6] *O-RAN Control, User and Synchronization Plane Specification 11.0*, document TS O-RAN.WG4.CUS.0-R003-v11.00, O-RAN Alliance, Mar. 2023.

[7] *O-RAN NR C-Plane Profile 9.0*, *document TS O-RAN.WG5.C.1-R003-v09.00*, O-RAN Alliance, Mar. 2023.

[8] *NG-RAN; Architecture Description (Release 17)*, 3GPP, document TS 38.401 V17.4.0, Mar. 2023.

[9] *Management and Orchestration; Artificial Intelligence/Machine Learning (AI/ML) Management (Release 18)*, 3GPP, document TS 28.105 V18.0.0, Jun. 2023.

[10] *Study on Artificial Intelligence/Machine Learning (AI/ML) Management (Release 18)*, 3GPP, document TR 28.908 V1.2.0, Apr. 2023.

[11] *NR Physical Layer Procedures for Data (Release 17)*, 3GPP, document TS 38.214 V17.5.0, Mar. 2023.

[12] *Robot. RB5 Platform*, Qualcomm Technologies Inc., San Diego, CA, USA. https://www.qualcomm.com/products/internet-of-things/industrial/industrial-automation/robotics-rb5-platform

[13] Thundercomm. *Qualcomm Robotics RB5 Development Kit*. San Diego, CA, USA. Accessed: Jul. 2023. [Online]. Available: https://www.thundercomm.com/product/qualcomm-robotics-rb5-development-kit/

[14] *Study on New Radio Access Technology: Radio Access Architecture and Interfaces (Release 14)*, 3GPP, document TS 38.801 v14.0.0, Mar. 2017.

[15] *NG-RAN; F1 Application Protocol (F1AP)*, 3GPP, document TS 38.473 V17.4.0, Mar. 2023.

[16] *NG-RAN; NR User Plane Protocol*, 3GPP, document TS 38.425 V17.3.0, Mar. 2023.

[17] *NR; Radio Resource Control (RRC); Protocol Specification*, 3GPP, document TS 38.331 v17.4.0, Mar. 2023.

[18] *NR; Packet Data Convergence Protocol (PDCP) Specification*, 3GPP, document TS 38.323 V17.4.0, Mar. 2023.

[19] *Evolved Universal Terrestrial Radio Access (E-UTRA) and NR; Service Data Adaptation Protocol (SDAP) Specification*, 3GPP, document TS 37.324 V17.0.0, Apr. 2022.

[20] *NR; Radio Link Control (RLC) Protocol Specification*, 3GPP, document TS 38.322 V17.2.0, Jan. 2023.

[21] *NR; Medium Access Control (MAC) Protocol Specification*, 3GPP, document TS 38.321 v17.4.0, Mar. 2023.

[22] *O-RAN Alliance*. Accessed: Jul. 2023. [Online]. Available: https://www.o-ran.org/

[23] *Small Cell Forum (SCF)*. Accessed: Jul. 2023. [Online]. Available: https://www.smallcellforum.org/

[24] *5G FAPI: PHY API Specifications*, document SCF222.10.06, Small Cell Forum (SCF), 5G FAPI Specifications, Dec. 2022.

[25] *5G nFAPI Specifications*, document SCF225.3.0, Small Cell Forum (SCF), 5G FAPI Specifications, Jul. 2022.

[26] *O-RAN Non-RT RIC Architecture 2.01*, document TS O-RAN.WG2.Non-RT-RIC-ARCH-TS-v02.01, O-RAN Alliance, Oct. 2022.

[27] *O-RAN Near-RT RIC Architecture 4.0*, document TS O-RAN.WG3.RICARCH-R003-v04.00, O-RAN Alliance, Mar. 2023.

[28] *NG-RAN; E1 Application Protocol (E1AP)*, 3GPP, document TS 38.463 V16.14.0, Jun. 2023.

[29] *NG-RAN; NG Application Protocol (NGAP)*, 3GPP, document TS 38.413 V17.4.0, Apr. 2023.

[30] *NG-RAN; NG Data Transport*, 3GPP, document TS 38.414 V17.0.0, Apr. 2022.

[31] *FSM100 5G RAN Platform for Small Cells*, Qualcomm Technologies Inc., San Diego, CA, USA. https://www.qualcomm.com/products/application/wireless-networks/small-cells/fsm100xx

[32] Qualcomm Technologies Inc. *Cloud AI: The Industry Leader in Performance Per Watt and Performance Density*. San Diego, CA, USA. [Online]. Available: https://www.qualcomm.com/products/technology/processors/cloud-artificial-intelligence?&cmpid=pdsrc-U9CczG3IQB&utm_medium=pdsrc&utm_source=AW&utm_campaign=FS-AI&gclid=EAIaIQobChMInYKSg92c_wIV_x6tBh2HrgJeEAAYASAAEgKUhfD_BwE

[33] *XR Traffic Model and KPI*, 3GPP, document TSG RAN WG1 104-e, Feb. 2021. [Online]. Available: https://www.3gpp.org/ftp/tsg_ran/WG1_RL1/TSGR1_104-e/Inbox/drafts/8.14.1/XR%20Traffic%20Model%20and%20KPI%20-%20Round%201/XR-01-XR%20Traffic%20Model%20and%20KPI.docx

[34] H. Nguyen-An, T. Silverston, T. Yamazaki, and T. Miyoshi, "IoT traffic: Modeling and measurement experiments," *IoT*, vol. 2, pp. 140–162, Feb. 2021, doi: 10.3390/iot2010008.

[35] 5G Alliance for Connected Industries and Automation. (Nov. 2019). *A 5G Traffic Model for Industrial Use Cases*. [Online]. Available: https://5g-acia.org/wp-content/uploads/2021/04/WP_5G_5G_Traffic_Model_for_Industrial_Use_Cases_22.10.19.pdf

[36] *NR and NG-RAN Overall Description Stage-2 (Release 17)*, 3GPP, document TS 38.300 v17.4.0, Mar. 2023.

[37] G. E. P. Box, G. Jenkins, G. C. Reinsel, and G. M. Ljung, *Time Series Analysis: Forecasting and Control*. Hoboken, NJ, USA: Wiley, 2015.

[38] Qualcomm Technologies Inc. *Snapdragon X60 5G Modem-RF System*. San Diego, CA, USA. Accessed: Jul. 2023. [Online]. Available: https://www.qualcomm.com/products/technology/modems/snapdragon-x60-5g-modem

BERK AKGUN received the B.S. and M.S. degrees in electrical and electronics engineering from Middle East Technical University, Ankara, Turkey, in 2012 and 2014, respectively, and the Ph.D. degree in ECE from The University of Arizona, Tucson, AZ, USA, in 2019. From 2012 to 2014, he was a Software Design Engineer with the Communication and Information Technologies Division, Aselsan, Ankara. He is currently a Senior Engineer with Qualcomm Technologies Inc. His research interests include mmWave channel characterization, robust mmWave system design, secure multiuser MIMO systems, and wireless communications and networking, with an emphasis on designing and integrating next-generation private networks.

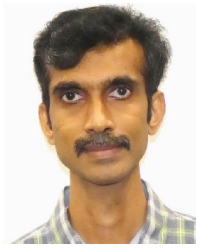
DEEPAK SINGH MAHENDAR SINGH received the bachelor's degree in electronics and communication engineering from the M. S. Ramaiah Institute of Technology, Bengaluru, India, and the M.S. degree in computer engineering from Texas A&M University, College Station, TX, USA, with a focus on communication, networks, and software. Since 2018, he has been with Qualcomm Technologies Inc., where he is currently a Senior Integration and Test Engineer on various 4G, 5G, and satellite communication research and development projects. His current contributions are in the field of wireless cellular technologies, precise positioning, machine learning, and private networks.

SAMATHA KOTLA received the B.E. degree from Osmania University, Hyderabad, India, and the M.S. degree from Arizona State University, Arizona, AZ, USA, with a focus on digital signal processing and wireless communications. In 2009, she joined Qualcomm Technologies Inc., San Diego, CA, USA, where she is currently a Senior Staff Engineer and the Manager. She has been involved in the systems test and integration of multiple advanced wireless systems, including 4G LTE and 5G NR. Her research interests include the Industrial IoT, private networks, ML CoMP, and XR.

VIKAS JAIN received the B.E. degree in electronics and communication engineering from the Thapar Institute of Engineering and Technology, India, in 1995, and the M.S. degree in computer science from the University of California at San Diego, San Diego CA, USA, in 2003. He is currently the Senior Director of Technology with the Wireless Research and Development Group, Qualcomm Technologies Inc., where he works on the software design and development for the 5G next-generation RAN research platform focusing on RAN disaggregation, virtualization, ML-enabled optimizations, and network automation. Since joining Qualcomm Technologies Inc., in 1998, he has contributed to the design, and prototyping of various advanced features as part of RAN software development for 3G CDMA EV-DO, 4G LTE, and 5G NR wireless technologies. Prior to Qualcomm Technologies Inc., he contributed to the software development for digital telephony switching systems at C-DOT, India. He holds 24 granted U.S. patents and more than 120 granted international patents.

SAKSHI NAMDEO received the master's degree in electrical engineering from Stanford University, Stanford, CA, USA, with a focus on communication systems, signal processing, and machine learning. She is currently a Senior Engineer with Qualcomm Technologies Inc. She has been a part of the System Integration and Test Team for more than three years. She works on enabling 5G technologies and solutions on end-to-end systems in the Industrial IoT sphere.

RUPESH ACHARYA received the B.E. degree in electronics and communication from the Institute of Engineering, Pulchowk Campus, Nepal, and the M.S. degree in electrical engineering from The University of Texas at Dallas. In 2009, he joined Qualcomm Technologies Inc., where he is currently a Senior Staff Engineer with the System Integration and Test Team. He has contributed to 3G, 4G, 5G, and Wi-Fi-related end-to-end system integration and testing.



MURUGANANDAM JAYABALAN received the B.E. degree in electronics and communication engineering from the Sathyabama Engineering College, India, in 2000, and the M.S. degree in electrical engineering from Wichita State University, Wichita, KS, USA, in 2004. In 2006, he joined Qualcomm Technologies Inc., where he is currently the Director of Engineering and leading the System Integration and Test Team focusing on RAN Disaggregation on 5G. He is also involved in ultra-mobile broadband (UMB), 4G LTE, and 5G NR: coordinated multipoint, precise positioning, the Industrial IoT, and private networks.

ABHISHEK KUMAR received the master's degree in computer applications from Delhi University, India, in 2006. After completing the master's degree, he joined Hughes Systique Corporation, where he worked on various projects in developing 3G/4G-based UE and RAN side protocol stacks for the LEO and GEO satellite communications. In 2019, he joined Qualcomm Technologies Inc. He has been contributing to the design and development of 5G and 6G cellular wireless technologies of industrial IoT, private networks, and RAN disaggregation.

VINAY CHANDE received the degree in engineering from the Indian Institute of Technology, Mumbai, and the Ph.D. degree in electrical engineering from the University of Maryland. As a Systems Engineer with the Wireless Research and Development Group, Qualcomm Technologies Inc. His current work allows him to participate in and witness the advances in millimeter-wave radio bands, unlicensed spectrum access, and machine learning for the Industrial IoT.

ARUMUGAM KANNAN received the B.E. degree in electronics and communication engineering from Anna University, India, in 2002, and the M.S. and Ph.D. degrees in electrical engineering from the Georgia Institute of Technology, in 2004 and 2007, respectively. He is currently a Principal Engineer with Qualcomm Technologies Inc., where he has contributed to the design, prototyping, and standardization of multiple LTE and 5G cellular wireless technologies ranging from advanced receivers, shared unlicensed spectrum operation of cellular systems, millimeter wave communications, the Industrial IoT, and wireless AI/ML research. He holds more than 200 granted U.S. patents and more than 600 granted international patents.

JALAJ SWAMI received the B.Tech. degree in computer science and engineering from IIT Kanpur, India, and the M.S. degree in information and computer science from the University of Hawai'i at Mānoa. He initially worked on telecom software for C-DOT India. He is currently a Software Engineer with Qualcomm Technologies Inc., where he has worked on device software development, as well as infrastructure software for wireless, broadcast, and satellite communications. He is contributing to 5G and 6G development for the Industrial IoT and RAN disaggregation.

YITAO CHEN received the B.S. degree in ECE from Shanghai Jiao Tong University, in 2014, and the Ph.D. degree in ECE from The University of Texas at Austin, in 2020. He is currently a Senior Engineer with Qualcomm Technologies Inc., where he has contributed to the design, prototyping, and standardization of 5G cellular wireless technologies of MIMO, Industrial IoT, and wireless ML research. He holds 13 U.S. patents. His research interests include wireless communication, information theory, coding theory, signal processing, and machine learning.

JOHN BOYD received the B.S. degree in computer science and engineering from the University of California at San Diego, San Diego, CA, USA, in 1994. He is currently the Director of Engineering with Qualcomm Technologies Inc., where he is also a Software Engineer and the Manager of the Wireless Research and Development Group. Having the opportunity to contribute to a diverse range of projects throughout the company during his 27 years of service. His current focus is a machine learning framework that facilitates the development of RAN optimizations.

XIAOXIA ZHANG received the B.S. and M.S. degrees in ECE from the University of Science and Technology of China and the Ph.D. degree in ECE from The Ohio State University, in 2002. After graduation, she joined Qualcomm Technologies Inc., where she is currently the Senior Director. She is involved in 3G, 4G, and 5G, including PHY/MAC system design, evaluation, prototyping, and standardization. She is leading the project focusing on Industrial IoT, shared/unlicensed spectrum, and RAN disaggregation on 5G/6G technology. She is a prolific inventor with more than 300 granted U.S. patents.