Scalable and Cost-Effective RAN: A Dynamic Resource Management Framework for Optimizing Hardware Utilization

Umar Toseef
Nokia Bell Labs
umar.toseef@nokia-bell-labs.com

Bartek Kozicki
Nokia Bell Labs
bartek.kozicki@nokia-bell-labs.com

Dragan Samardzija
Nokia Bell Labs
dragan.samardzija@nokia-bell-labs.com

***Abstract*—Escalating mobile service demands require Mobile Network Operators (MNOs) to optimize Radio Access Network (RAN) infrastructure for cost, efficiency, and scalability. While Cloud RAN virtualization improves resource pooling, hardware remains underutilized during low-traffic periods. This study introduces a dynamic resource management framework for 5G Cloud RAN that optimizes L1-Hi processing on hardware accelerators like GPUs and SmartNICs. By combining predictive workload modeling with adaptive allocation, the framework scales resources based on real-time traffic. This approach mitigates over-provisioning, thereby increasing hardware utilization, lowering energy consumption, and reducing operational costs. Simulations confirm the framework enhances hardware efficiency while maintaining user Quality of Experience (QoE), enabling more scalable, cost-effective, and sustainable telecom networks.**

***Keywords*—Cloud RAN, Resource Management, Hardware Accelerators, virtual Distributed Unit, Layer 1 Processing**

I. Introduction

The surge in mobile network usage, driven by increased data consumption and a growing subscriber base, pressures Mobile Network Operators (MNOs) to expand Radio Access Network (RAN) infrastructure. RAN accounts for 65–70% of telecom costs, making optimization critical as traffic and user expectations rise [1]. Traditional peak load provisioning leads to significant underutilization, with average RAN usage at just 25–50%—a figure that drops to 20% of peak levels during nighttime hours.

To combat this inefficiency, solutions like Centralized RAN (C-RAN) and Cloud RAN introduce virtualization to consolidate baseband processing and improve resource utilization [2]. In 5G, this architecture splits functions into virtualized Central Units (vCUs) and Distributed Units (vDUs), where vDUs rely on specialized accelerators (e.g., GPUs, SmartNICs) for real-time Layer 1 (L1) processing.

However, these powerful accelerators are also provisioned for peak loads and thus remain idle much of the time, creating a critical need for dynamic management. Addressing this gap, our study proposes a framework to dynamically scale accelerator resources for L1-Hi processing based on real-time traffic demands. This approach enhances hardware utilization, reduces capital (CAPEX) and operational (OPEX) costs, and improves energy efficiency, paving the way for sustainable telecom networks.

The paper is structured as follows: Section II reviews related work, Section III discusses hardware acceleration, Section IV presents our workload modeling, Section V details the proposed framework, Section VI evaluates its performance, and Section VII concludes with future work.

II. Related Work

Significant research has explored balancing computational loads in Baseband Unit (BBU) pools, often migrating workloads from overloaded to underutilized BBUs to improve resource utilization and energy efficiency [3]. Clustering techniques, such as location-aware, load-aware, and QoS-aware clustering, dynamically associate BBUs with Remote Radio Heads (RRHs) to meet user demand [4]. These methods typically involve switching BBUs on or off during low-traffic periods but address only specific optimization aspects and lack rapid adaptability under real-world constraints.

Frameworks like Pompili et al. [5] propose elastic, on-demand BBU resource allocation using virtualization, implementing baseband functions as virtual network functions on general-purpose servers. These models enhance utilization but assume sufficient peak-load capacity, overlooking resource-constrained scenarios [6]. Optimization algorithms, including game-theory and linear programming, distribute computing resources among BBUs to maximize pool utilization [7]. Emerging studies leverage machine learning to predict traffic loads and scale resources dynamically [8]. However, these models' computational complexity limits practicality in environments with frequent short-term fluctuations [9]. Our framework diverges from these methods by employing a lightweight allocation algorithm built on empirically derived workload models, which enhances both adaptability and efficiency. This approach yields significant performance gains, with simulations showing up to a 50% increase in serving capacity under moderate traffic while maintaining high QoE.

The primary novelty of our work lies in its granular focus. Unlike prior research that treats BBUs as monolithic entities, our framework manages resources for vDUs by decoupling L1-Hi processing on specialized accelerators (e.g., GPUs, FPGAs, SmartNICs) from L2 tasks on CPUs. This separation is critical because L1-Hi and L2 workloads scale differently; L1-Hi processing demands are driven by Physical Resource Blocks (PRBs), MIMO layers, and Modulation and Coding Scheme (MCS), whereas L2 scales mainly with user count. By targeting the unique requirements of L1-Hi, our approach provides tailored resource management that directly addresses hardware underutilization and operational inefficiencies, thereby improving the scalability and sustainability of Cloud RAN infrastructures

III. Hardware Acceleration for L1-Hi Processing

As 5G networks evolve to support higher bandwidths and advanced antenna systems, pure CPU-based platforms struggle to meet the intense computational demands of L1-Hi processing, necessitating hardware acceleration. SmartNICs

978-3-903176-75-1 ©2025 IFIP

are an excellent example of acceleration for L1-Hi processing, utilizing specialized System-on-Chip (SoC) architectures. These SoCs contain dedicated components like ARM or RISC-V cores for control plane tasks, Digital Signal Processors (DSPs), and hardware accelerators to handle computationally-intensive tasks. They also feature high-bandwidth memory to facilitate rapid data transfer and ensure low latency. Within the SmartNIC, a high-speed job scheduler dynamically assigns baseband processing tasks to the DSPs and hardware accelerators, enabling efficient parallel processing for multiple cells.

IV. Workload Modelling

SmartNICs perform baseband processing for multiple cells, with their capacity typically defined by full-load Layer 1 (L1) configurations. However, under moderate traffic, the same hardware can support additional cells, significantly improving resource utilization. To quantify this potential, we developed empirical workload models that mathematically map L1 configuration parameters and cell load to their corresponding processing times.

Due to a lack of vendor-provided empirical data, we conducted controlled experiments on a commercial SmartNIC using its development kit. We created custom unit tests to simulate various cell loads and L1 configurations by manipulating parameters such as PRBs, MIMO layers, MCS index, and user count. These automated tests measured key performance indicators, including symbol processing time and resource utilization. Capturing a comprehensive dataset of hardware behavior required running hundreds of unique experiments, one for each distinct L1 configuration.

The models were developed for a SmartNIC advertised to support 16 cells at full load. As shown in Fig. 1, we analyzed processing times for two critical L1 tasks—PDSCH symbol processing and Frequency Offset Compensation (FoC)—as a function of PRB count. The results, reported in Abstract Units (AU) for confidentiality, revealed that PDSCH processing time scales proportionally with PRB count, though its variance widens at higher values, suggesting increased complexity. Conversely, FoC processing time shows a predictable, near-linear scaling. To model these trends, we applied linear regression against the 95th percentile of measured times to establish a robust performance boundary. The models proved highly accurate, achieving coefficients of determination (R^2) of 0.9974 for PDSCH and 0.9994 for FoC. Since processing times also vary with MIMO and modulation, we developed separate models for each configuration to maintain accuracy.

We applied this methodology to other L1-Hi tasks like PUCCH and RACH , using linear, piecewise, or polynomial regression as appropriate. For validation, the dataset was split into training (80%) and validation (20%) subsets. Models were evaluated using Mean Squared Error (MSE) and R^2, and residual analysis confirmed the suitability of our regression techniques. Controlled experiments further verified the models' accuracy. We chose regression over complex methods like deep neural networks for its practicality in a real-time vDU, where low computational overhead and small storage footprints are critical.

While the resulting models are specific to the tested SmartNIC, the methodology is broadly generalizable. The core approach of empirical data collection, regression analysis, and workload-to-processing time mapping can be adapted for other accelerators like FPGAs or GPUs. These workload models are the foundation of our dynamic resource management framework, enabling predictive resource allocation based on real-time network demands. By pre-calculating estimates, the system can proactively optimize resources without imposing test loads. Through this combination of empirical testing and validated modeling, our work establishes a robust method for understanding and optimizing computational workloads in vDUs.

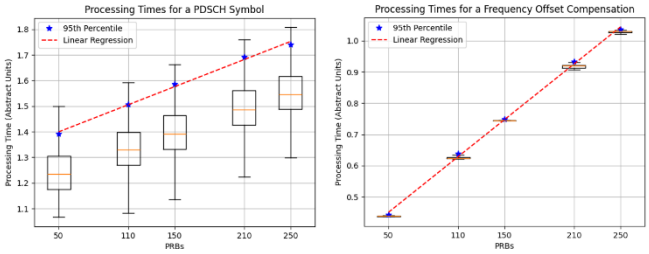

Fig. 1: Processing times for PDSCH and FoC tasks, shown as a function of PRBs. The box plots show the data distribution, blue stars indicate the 95th percentile, and the red dashed line represents the linear regression fit.

V. Proposed Resource Management Framework

This section outlines the requirements, design, and operational mechanisms of the proposed resource management framework, addressing the challenges of computational load distribution, communication overhead, and dynamic traffic conditions.

1. *Requirements for a Resource Management*

The proposed framework is designed to maximize accelerator utilization and ensure reliable operation under fluctuating traffic conditions. It is built on four core requirements:

- Efficiency: A lightweight design to allow frequent updates without computational burden.
- Low Overhead: Minimized control-plane communications.
- Fairness: Dynamic and equitable resource distribution based on traffic loads.
- Reliability: A robust mechanism to prevent accelerator overload and L1-Hi processing failures.

2. *Framework Design*

Building on these requirements, this study proposes a centralized resource management framework where a single manager allocates computational resources from an accelerator, like a SmartNIC, to multiple cells. This system optimizes resource distribution and improves efficiency by using workload models that accurately estimate processing times based on cell loads and Layer 1 (L1) configurations.

Fig. 2 illustrates a vDU host equipped with SmartNIC designed to support multiple cells. The framework operates through a communication cycle. Each cell has a resource management (RM) client that monitors its load and sends its

PRB demands to the central resource manager. The manager uses this information, along with workload models, to determine the necessary resource allocations. These allocations are then sent back to the RM client, which informs the cell's L2 processing subsystem.

The central manager's allocation imposes a new constraint on the L2 scheduler. Traditionally, L2 scheduling is based on factors like Channel Quality Indicators (CQI) and Quality of Service (QoS). With this framework, the L2 subsystem must now ensure its scheduling decisions remain within the allocated resource limits, which are periodically updated to reflect changing traffic conditions. This dynamic process ensures efficient resource utilization. If the total demand from all cells exceeds the accelerator's capacity, the manager adjusts allocations to keep the total distribution within the hardware limits. The vDU host is assumed to have sufficient capacity for L2 processing, allowing the manager to focus exclusively on distributing L1-Hi resources. This approach is further discussed in next subsection.

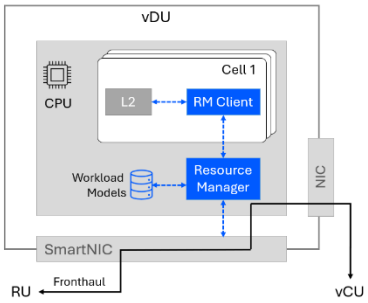

Fig. 2: A vDU with a SmartNIC for L1-Hi processing, showing resource management components in blue.

3. *Resource Distribution*

When the total resource demand from all cells exceeds an accelerator's capacity, the framework manages resource allocation through adaptable distribution strategies. A priority-based distribution gives precedence to high-priority cells, ensuring critical demands are met, which is useful for emergency scenarios. A fair distribution allocates resources proportionally to cell demands using weighted round-robin techniques, ensuring equitable sharing. A hybrid approach combines these methods, prioritizing important cells while fairly distributing any remaining resources.

These strategies are designed to be simple and fast, allowing for real-time execution. They assume cells report their demands accurately to avoid unnecessary complexity and ensure the resource distribution is both effective and scalable for dynamic 5G network traffic.

4. *Traffic Load Prediction*

For managing resources in wireless networks, predicting cell resource demands is critical. This is possible because cellular traffic often follows predictable patterns based on the time of day and the short duration of grant periods (under 100 milliseconds), which causes gradual load changes.

Various methods can be used for forecasting these demands [10]. Classical time-series forecasting techniques like ARIMA and Exponential Smoothing are effective at capturing temporal patterns. On the other hand, machine learning approaches, such as LSTM networks and XGBoost, offer data-driven adaptability for dynamic traffic.

Accurate load predictions allow the resource manager to allocate the optimal amount of resources to each cell, thereby avoiding both service degradation from under-allocation and inefficiency from over-allocation. The task of forecasting is delegated to individual cells to ensure scalability and accuracy. Each cell predicts its own demands and sends the information to the central resource manager, which then focuses on optimizing resource allocation across all cells.

5. *Grant Period Length*

Efficient resource management in cellular networks depends on the grant period length, which determines how often resources are reallocated. A fixed grant period is simple but can be inefficient, as short periods create overhead and long periods can't adapt to rapid traffic changes. Dynamic grant periods are more flexible, adjusting their length to traffic conditions to balance responsiveness and overhead. An alternative is asynchronous updates, where cells report changes only when significant shifts occur, reducing control-plane overhead. The optimal strategy balances overhead, and system complexity to enhance efficiency and scalability.

6. *Avoiding Overloading*

To prevent an overloaded accelerator from causing L1-Hi processing failures, a resource manager uses real-time load Key Performance Indicators (KPIs) from the hardware alongside predictions from a workload model. A dynamic allocation system is employed to keep resource utilization within a safe range, typically targeting 90–95% to leave a 5–10% buffer for unexpected demand spikes. The system adjusts allocations adaptively, scaling up when utilization drops below 85% and scaling down when it exceeds 95%. Shortening the duration of grant periods could improve responsiveness, allowing for more frequent reallocations to match sudden changes in traffic. This ensures that resource distribution aligns with real-time demands, optimizing performance while minimizing the risk of overloading.

7. *RM Client Feedback*

RM clients are crucial for refining resource distribution by providing feedback on both resource utilization and the adequacy of the current grant period. This feedback allows the resource manager to make informed, dynamic decisions. For instance, a cell with high resource utilization may signal an increased load, making it a strong candidate for more resources. Conversely, low utilization could prompt resource adjustments to prevent over-allocation.

Feedback on the grant period is also vital. RM clients can suggest shorter grant periods when accurate load predictions are difficult, or recommend longer ones when predictions are reliable, which helps reduce system overhead. This continuous feedback loop ensures that the resource distribution process is continuously optimized to meet real-world demands and improve overall system performance.

8. *Information Flow*

In each grant period, the resource management framework dynamically allocates resources for L1-Hi processing. As

illustrated in Fig. 3, the process starts when RM clients retrieve load information, including PRB usage, from the L2 processing unit. The RM clients then process this data to estimate PRB needs for the next grant and evaluate the adequacy of the current grant period.

This information is sent to a central resource manager, which consolidates demands from all clients. The resource manager uses workload models and hardware usage data to determine the new PRB allocations for all cells. These allocations are then sent back to the RM clients, which update the L2 subsystem's scheduling parameters. This cyclical process repeats, ensuring an efficient and dynamic allocation of resources. This framework effectively addresses challenges like overloading and scalability by combining centralized management with client feedback and load predictions.

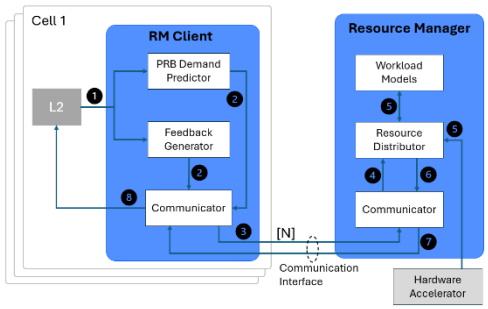

Fig. 3: Information flow within resource management framework

VI. Results and Discussion

A. Workload Models

Workload models, developed for a commercial SmartNIC that supports 16 cells under full load, form the basis of a new resource management framework. These models enable precise estimation of processing times and efficient resource allocation by providing granular insight into the L1-Hi processing chain.

For downlink operations, the SmartNIC processes tasks such as PDSCH and PDCCH sequentially. A test with 16 cells under full load validated the models, showing that they provide accurate, yet slightly conservative, estimates. This conservative approach is intentional, creating a safety buffer to prevent hardware overloading (see Fig. 4).

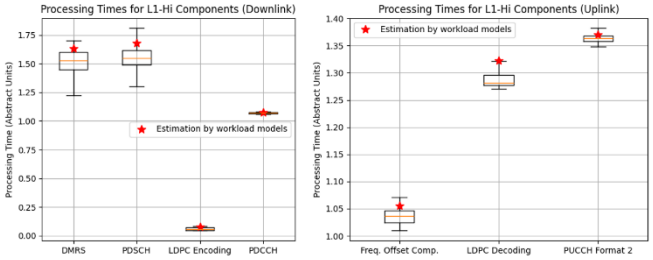

Fig. 4: Boxplots of L1-Hi processing times on SmartNIC. Workload model estimates are marked with stars, demonstrating alignment with actual measurements.

For uplink operations, the SmartNIC uses slot-level processing with parallel hardware accelerators. Similar to the downlink, workload models for uplink tasks also provided precise and slightly conservative predictions that closely matched empirical data. The findings confirm that the methodology for developing these workload models is effective, as they accurately predict processing times based on L1 cell configurations and traffic loads, a crucial function for the resource management framework.

B. Resource Management

To evaluate the proposed resource management framework, this study used the OMNeT++ network simulator with its Simu5G library [11]. The framework was implemented within Simu5G with a central resource manager and an RM client for each cell. The RM client communicates with Layer 2 (L2-PS) to enforce PRB limits given by the resource manager, ensuring resource usage doesn't exceed allocated capacity. Workload models, based on real hardware measurements of a SmartNIC, were integrated into the resource manager to enable accurate and dynamic allocation decisions. The SmartNIC was modeled as a high-level statistical representation, with processing delays calibrated using real hardware measurements.

The simulation scenarios involved multiple gNBs serving users with VoIP and web browsing traffic. User counts and movement were adjusted to simulate varying cell loads. A seeding technique was used to ensure reproducible yet dynamic load predictions, achieving around 90% accuracy. A fixed grant period of 50 ms was chosen for periodic resource allocations based on this prediction accuracy.

Simulations were run with configurations exceeding the SmartNIC's advertised capacity of 16 cells, but with reduced average loads. PRB demands fluctuated over time. When aggregated demands were within the SmartNIC's capacity, grants equaled demands. When demands exceeded capacity, a round-robin approach was used to distribute resources.

The study examined how often demands were met when more than 16 cells were served. Results, as shown in the cumulative distribution function (CDF) of demand-grant differences, demonstrated that cell demands were met most of the time (see Fig. 5). Reductions in grants occurred in less than **6%** of cases and had no noticeable impact on QoE. Key metrics like throughput and latency were monitored to confirm this. For example, the average RLC PDU delay remained at 2.0ms (±0.3ms), and throughput reached 95% of the baseline capacity. These findings suggest that minor reductions in resource allocation are effectively managed by higher-layer scheduling, preserving the end-user experience.

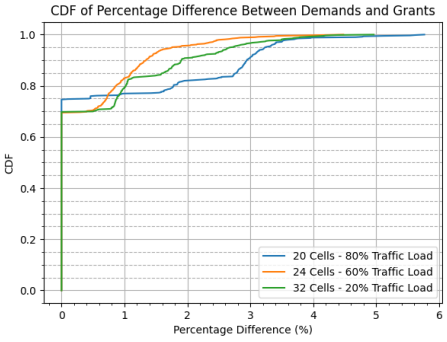

Fig. 5: CDF of the differences between cell demands and grants

While the SmartNIC is advertised to support 16 fully loaded cells, this study found it could handle a significantly higher number under lighter traffic conditions. For instance, it

could support up to 20 cells at an 80% mean load (a 25% capacity increase) and up to 24 cells with a mean load below 60% (a 50% increase). In low-traffic scenarios, such as nighttime, the SmartNIC successfully served 32 cells—double its advertised capacity.

This enhanced capacity allows network operators to consolidate workloads onto fewer vDU hosts during off-peak hours, enabling the powering down of surplus hosts. This strategy leads to significant energy savings, reduced cooling needs, and lower hardware wear. For example, serving 24 cells at 60% load reduces energy usage per cell by about 20% compared to a baseline of 16 fully loaded cells.

The study also observed an overhead when the SmartNIC served more cells than its advertised capacity, meaning the theoretical efficiency of a 50% load was not fully achieved when serving 32 cells. The exact cause of this overhead could not be determined due to the SmartNIC's proprietary, "black box" nature. However, these findings suggest opportunities for future optimization to further extend the practical capacity of these accelerators under reduced load conditions.

In scenarios with increasing cell counts, the SmartNIC's computational resource utilization approaches 90% in the 32-cell scenario, as shown in Fig. 6. This suggests the resource management framework efficiently uses idle resources. However, scaling beyond 32 cells may not be possible without further optimization.

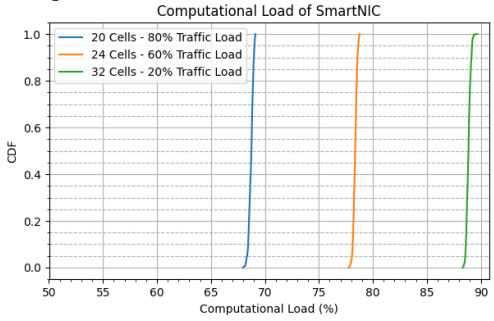

Fig. 5: Computational resource utilization of the SmartNIC across different cell count scenarios

Without the framework, high load peaks caused the SmartNIC to enter an overload state, triggering failure events (where cells aren't processed within the required time). As Fig. 7 shows, the number of cumulative failures increases over time without resource management, emphasizing its importance in preventing these disruptions.

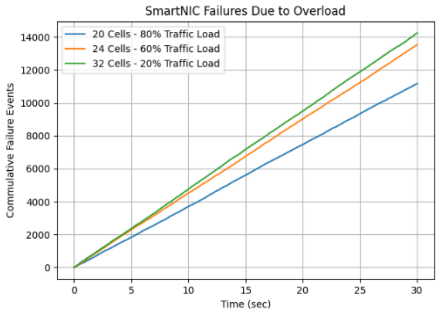

Fig. 6: Cumulative failure events over time without resource management

The proposed framework prevents failures by dynamically controlling load peaks. It assesses computational demands, redistributes resources, and trims unserviceable peaks to keep operations within safe limits. While trimming is vital for stability, prolonged excessive trimming can reduce the QoE. To prevent this, the framework must also scale resources up during long periods of high load.

VII. Conclusions and Future Work

This study presents a dynamic resource management framework aimed at optimizing L1-Hi processing in 5G Cloud RAN. By leveraging predictive workload models and adaptive allocation strategies, the framework addresses inefficiencies associated with static allocation methods. Simulation results demonstrate its effectiveness in extending the capacity of hardware accelerators like SmartNICs, supporting up to 24 cells at 60% load—a 50% improvement over their advertised capacity—without compromising user QoE. In extreme scenarios, such as nighttime traffic conditions with mean cell loads of 20%, the SmartNIC successfully served 32 cells, achieving double its advertised capacity.

The findings highlight the potential of dynamic resource management to optimize operational costs, reduce energy consumption, and enhance the scalability of 5G deployments. Future work will focus on expanding the approach to alternative accelerators, such as GPUs, and exploring integration with radio resource management strategies to create a holistic optimization model for 5G RANs. This research lays a robust foundation for cost-effective, scalable, and energy-efficient next-generation telecommunications, supporting the transition to greener 5G networks and enabling the future of ubiquitous, high-performance connectivity.

VIII. References

[1] O-RAN Alliance, “O-RAN: Towards an Open and Smart RAN,” White Paper, Oct. 2018.

[2] R. Gadiyar and S. Chowdhury, "RAN-in-the-Cloud: Delivering Cloud Economics to 5G RAN," NVIDIA Technical Blog, Feb. 13, 2023. [Online]. Available: https://developer.nvidia.com/blog/ran-in-the-cloud-delivering-cloud-economics-to-5g-ran/.

[3] Ismail, S.F.; Kadhim, D.J. Adaptive BBU Migration Based on Deep Q-Learning for Cloud Radio Access Network. Appl. Sci. 2025, 15, 3494.

[4] A. Ibrahim, A. Elsheikh, B. Mokhtar and J. Prat, "Clustering-Driven Optimization of RRH-BBU Assignment for Green Communication Networks With Big Data Analytics," in IEEE Access, vol. 12, pp. 177080-177092, 2024.

[5] D. Pompili, A. Hajisami and T. X. Tran, "Elastic resource utilization framework for high capacity and energy efficiency in cloud RAN", IEEE Commun. Mag., vol. 54, no. 1, pp. 26-32, Jan. 2016.

[6] S. K. Sharma, S. K. S. Gupta, "Beneficial and Efficient Secure Network Function Virtualization in 5G Wireless Networks," International Journal of Computer Networks and Systems, vol. 7, no. 11, Nov. 2022.

[7] M. Barahman, L. M. Correia and L. S. Ferreira, "A QoS-Demand-Aware Computing Resource Management Scheme in Cloud-RAN," in IEEE Open Journal of the Communications Society, vol. 1, 2020

[8] R. Rodoshi, T. Kim and W. Choi, "Deep Reinforcement Learning Based Dynamic Resource Allocation in Cloud Radio Access Networks," 2020 International Conference on Information and Communication Technology Convergence (ICTC), Jeju, Korea, 2020

[9] MAMMOET—Massive MIMO for Efficient Transmission, Mar. 2020, [online] Available: https://mammoet-project.eu.

[10] Wu, X., & Wu, C. (2024). CLPREM: A Real-Time Traffic Prediction Method for 5G Mobile Network. PLOS ONE, 19(4).

[11] Nardini, G., Sabella, D., Stea, G., Thakkar, P., & Virdis, A. (2020). Simu5G – An OMNeT++ Library for End-to-End Performance Evaluation of 5G Networks. IEEE Acce