[WORKSHOP]Towards Greener Networks: RApp-Based Cell Control over O-RAN Deployments

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

Abstract—**Efficient spectrum management is crucial for the CitizenBroadband Radio Service (CBRS), promoting shared radiofrequency spectrum use. The US Federal CommunicationsCommission (FCC) designated the frequency range of 3550MHz to 3700 MHz for three types of users: incumbent users,Priority Access License (PAL) users, and General AuthorizedAccess (GAA) users. The Spectrum Access System (SAS)coordinates spectrum sharing among user tiers by usingEnvironment Sensing Capabilities (ESC) sensors to detectincumbent users and prioritize their access while operatingin uninformed incumbent detection mode. Machine Learning(ML) based techniques can be used for incumbent detectionand enforce incumbent protection through SAS. VirginiaTech researchers have developed an open-source SAS forspectrum-sharing experimentation. In this work, we leverage the open-source SAS by enabling and testing WirelessInnovation Forum (WInnForum) standards (.i.e CBRS BaseStation Device (CBSD) interface), implementing an ML-basedincumbent radar detection mechanism using a feed-forwardneural network in an Software-defined Radio (SDR)-basedexperimental CBRS network deployed at CCI xG testbedand naming it "OpenSAS". If the presence of the incumbentis determined with at least a certainty of 85**

Index Terms—Keywords.

I. Introduction

What is the problem you are trying to solve? Give some background to it

In the 6G era, the digital, physical and human world seamlessly fuse to trigger extrasensory experiences. Intelligent knowledge systems will be combined with robust computation capabilities to make humans endlessly more efficient and redefine how we live, work and take care of the planet. These applications necessitate support for new capabilities for Enhanced Mobile Broadband (eMBB), Ultra Reliable Low Latency Communications (URLLC), and massive Machine Type Communications (mMTC) [3,4]. While a mobile network consists of multiple parts, the radio access network (RAN) is responsible for most of the energy consumption in a mobile network [Green Future Networks – Sustainability Challenges and Initiatives in Mobile Networks by NGMN Alliance, December 2021,]. [https://arxiv.org/abs/2301.06713]
- While on the one hand, network slicing provides an op-portunity for efficient radio resource management (RRM)through fine-grained control and customization of radio ac-cess networks (RANs), initiatives such as O-RAN [18] haveintroduced the notion of RAN Intelligent Controllers (RICs)that enable enhanced observability and programmabilitywithin the RAN domain, thereby providing a flexible plat-form for robust RAN slice assurance and control. Conse-quently, this has piqued the interest of researchers fromacademia and industry to explore O-RAN-centric RRM tech-niques for network slicing [11, 20]. However, the underlyingRRM strategies prevalent in the prior art have largely re-volved around the common theme of using quality of service(QoS) indicators such as network-level throughput and la-tency metrics to estimate slice resource needs.
- Switching off under-loaded cells during network operation without affecting the user experience (call drops, QoS degradation, etc.) is one way to achieve RAN energy efficiency. A typical energy savings scenario is realized when capacity booster cells are deployed under the umbrella of cells providing basic coverage and the capacity booster cells are switched off to enter dormant mode when its capacity is no longer needed and reactivated on a need basis. - First, the E2 Nodes are configured by the Service Management and Orchestration (SMO) to report the data necessary for energy-saving algorithms via the O1 Interface to the Collection and Control unit. Assuming that the Non-RT RIC and SMO are tightly coupled the NonRT RIC retrieves the collected data through internal SMO communication how???. The O-RUs are involved in this use case. The E2 Nodes need to configure them to report data through the Open RAN Fronthaul Management Plane (Open FH M-Plane) interface.
- Before switching off/on carrier(s) and/or cell(s), the E2 Node may need to perform some preparation actions for off switching (e.g. check ongoing emergency calls and warning messages, to enable, disable, modify Carrier Aggregation and/or Dual Connectivity, to trigger HO traffic and UEs from cells/carriers to other cells or carriers, informing neighbour nodes via X2/Xn interface etc.) as well as for on switching (e.g., cell probing, informing neighbour nodes via X2/Xn interface etc.).

- A radio access network comprises of several cell sites with site infrastructure equipment and base station equipment. O-RAN can contribute to this effort immensely. Its disaggregated and virtualized architecture adds complexity; however, energy is the next major challenge O-RAN must overcome.

- How is this different from all the other implementations? Is this just a simple implementation of a pre-existing idea in an O-RAN specification?
- What are the insights from our experiments? - Results Part The key insights (denoted as "I") resulting from our analysis can be summarized as follows:I 1: We confirm with a recent study an old finding that datesback to the year 2004 [16]: data-plane traffic exhibitsself-similarity properties at both individual uplink anddownlink components and as a whole. This is differentfrom control-plane mobile network traffic [29]I 2: We find that the number of Radio Resource Control (RRC)connected users follows a bi-modal distribution that in-dicates the presence of circadian cycles resulting in twoclusters of different sizes, i.e., users connected during theday and users connected during the night.
- Paper Structure
Edit with Abstract The remainder of this paper is organized into five sections, as follows. Section 2 provides the considered assumptions and formulatesthe void problem in the context of spatial query processing. Section 3 presents the non-functional requirements and the overallbehavior of the proposed distributed serial approach (composed of three main phases: window reaching, query processing, and result reporting). Section 4 presents the underlying mechanisms utilized by the proposed approach (mainly, the curved-stick [7] androlling-ball [8]) and then details the query processing phase. Theobtained simulation results are depicted and discussed in Section 5. Finally, Section 6 concludes the paper and suggests future research directions.

II. BACKGROUND AND RELATED WORK

Recent works have focused on... - Rimedo Labs recently released Energy Saving rApp (ES-rApp) for Non-RT RIC focusing on Massive MIMO use cases.

- The RF carrier shutdown feature (typically hosted by the SMO and non-RT RIC in O-RAN architecture) periodically checks the service load of multiple carriers and if the service load is below a specified threshold, the capacity-layers are shut off (see Figure 12). The UEs served by those carriers can camp on or access the services from the carrier providing basic coverage. When the load of the carrier providing basic coverage is higher than a specified threshold, the base station turns on the carriers that have been shut down for service provisioning.
- RF carriers can be shut down by non-RT RIC rAPPs more intelligently using information from telemetry data from E2 interface with O-RAN.
- When the booster gNB with CU/DU split decides to switch off cell(s) to the dormant state, the decision is typically made by the gNB-DU based on cell load information or by the OAM entity (non-RT RIC in O-RAN architecture). Before the cell in the gNB-DU enters into the dormant mode, the gNB-DU will send the gNB-DU configuration update message to the gNB-CU to indicate that the gNB-DU will switch off the cell subsequently sometime later.
- During the switch-off period, the gNB-CU offloads the UE to a neighboring cell and simultaneously will not admit any incoming UE to this cell being switched-off. Is this load balancing performed?
- https://networkbuilders.intel.com/solutionslibrary/a-holistic-study-of-power-consumption-and-energy-savings-strategies-for-open-vran-systems

- RF channel switch off/on
- However, the switch off/on decisions are need a lot of KPIs reporting and efficient actions so that guarantee the overall user experience. Also, there are conflicting targets between system performance and energy savings.

- Offline learning is normally preferred (including reinforcement learning which is usually performed online) due to the nature of the network environment, which is prone to misconfiguration and errors leading to outages. The proposed approach is for the model to first be trained with offline data and that the trained model then be deployed in the network for inference.
- This implementation is akin to a traffic steering xApp, as it involves offloading traffic from low-load cells to other cells, ensuring that as many RUs as possible can enter a low-power sleep mode. The handovers and traffic redistribution are predicted and managed in real-time by the xApp, enhancing the overall energy efficiency of the network. ==¿ https://www.diva-portal.org/smash/get/diva2:1765998/FULLTEXT01.pdf

- Ericcson, Rimodo Labs, Juniper Networks, VMWare have preexisiting rApps on the market, but have not described them very well.

- During the ES-rApp operation, it gets information from the O1 interface about cells available in the network, their type (macro cell or small cell), and the PRB usage of each cell. ES-rApp collects such data during predefined times and calculates average O-DU PRB usage in the time domain. Periodically, the rApp can make decisions about enabling or disabling one of the cells. ES-rApp will enable a cell in case of congestion (high PRB usage) observed in at least one cell that is currently enabled. ES-rApp will disable a cell in case of average PRB usage below some threshold. If none of the situations occurs, the ES-rApp continues observation. ==¿ https://rimedolabs.com/blog/es-rapp-control-over-ts-xapp-in-oran/

- VMWare and VIAVI ==¿ https://www.virtualexhibition.o-ran.org/classic/generation/2024/category/intelligent-ran-

control-demonstrations/sub/intelligent-control/394

- Juniper's implementation ==¿ https://blogs.juniper.net/en-us/service-provider-transformation/delivering-on-the-o-ran-promise-with-juniper-networks-ran-intelligent-controller-ric

last 1 para for the novelty of our paper compared to existing work. should i write anything here? *something like:* To the best of our knowledge, comparisons between ARIMA, Prophet, LSTM trained on a single dataset but applied in various conditions have not been well studied in the literature

A. Key Questions

Key questions to answer to implement above:
- Key question 1 - When to turn off and on the cells?
- Key question 2 - Which cells to turn on and off?
- Key question 3 - what is the goodness measure of particular energy saving decision

III. RELATED WORK

- Theme 1

- Theme 2

- Theme 3

IV. ES RAPP ARCHITECTURE OVERVIEW

Good overall overview of how the rApp works is required here. The rApp is data driven in the sense that it does not incorporate a rules-based logic but determines the rules which meet the target objective based on the input data and network configuration. A Dashboard for visualization of the Radio Mapping Database is also used.
- What is the output of the algorithm? Does it send a reccomendation to the Near-RT RIC over what to do? - sends a policy, sent as a declarative statement. sent across A1 interface.
- - Is the decision made periodically? - 1hr prediction, 15 minutes slots (every 15 minutes)

- Why rApp over an xApp? How does this link to the TS xApp? (receive a policy from the rApp)

The rApp has the following components:

V. ARCHITECTURE OVERVIEW

1) Digital Twin: Uses CloudRF to determine the coverage. Simulation which takes our parameters if need into context. More information required.

2) Radio Database: The **Radio Database** is a geospatial database that indexes data using latitude, longitude, and altitude, including clutter information. Currently cloudRF has internal clutter database and we rely on it (it is not exposed outside cloudRF). This database is initialized with network inventory and predicted RF power (downlink) for each pixel from sectors exceeding a predefined threshold (Pth). **How is threshold decided? - Sensitivity of mobile devices in use. Defined by user.**. The predictions are generated using a Radio Link budget simulator, using CloudRF. Additionally, the database stores timestamped measurement reports from gNBs in a sliding window with a preconfigured depth (td). Notably, this Radio Database can be external to the rApp, allowing it to be shared across multiple rApps. It also has the capability to import data from RF link simulators and drive tests through an external interface.

3) Traffic Predictor: The **Traffic Predictor** estimates the net traffic volume, percentage PRB utilization, and the number of active UEs for each sector as a function of time. This prediction is based on historical data and previous measurements. The Traffic Predictor employs ARIMA/SERIMA algorithms to forecast these values for the future. Input information for this component is expected in 15-minute intervals.

Are LSTMs used here? Refer to Doc2.

A. Coverage Predictor

The Coverage Predictor is responsible for predicting the coverage overlap between adjacent sectors. It also updates the link level prediction model based on actual measured values to enhance accuracy. The input to the system is the simulated received power level (obtained from cloudRF) for each pixel from all the sectors with contributions higher than Pth. How is Pth decided?

Convert to Algo Following steps are followed:
Step-C1: For each pixel, find the sector which has the highest power to that pixel. This is marked as the serving sector. This step determines the sector boundaries in the network. The other contributors are marked as interfering power.
Step-C2: For each sector iden+fied in step C1, find the overlap in coverage with adjacent sectors. This is done using the following steps:
- Step-C2.1: Two sectors Si and Sj are deemed adjacent if there exists a pixel in Si where Sj is the strongest interferer and also if there exists a pixel in Sj where Si is the strongest interferer.
- Step-C2.2: For all pairs of sectors Si and Sj iden+fied in Step-C2.1, find the count of pixels in Si, where Sj is the strongest interferer and all the pixels in Sj where Si is the strongest interferer. Find the value Eij, which is the sum of all the pixels thus iden+fied. - Step-C2.3: Generate the Overlap Graph of the network where each sector Si is the vertex while the Eij computed in Step-C2.2 is the weighted edge

B. Energy Saving Decision Entity

The **Energy Saving Decision Entity** utilizes the results from both the Traffic Predictor and the Coverage Predictor to identify sectors in the network that can be shut down with minimal service impact. This entity performs simulations to evaluate the impact of energy-saving decisions before configuring the network through the SMO or EMS. The functioning of this entity is defined in Algo. 1

1. For each sector Si, iden+fy sectors Sj, Sk, . . . , Su, which has an edge with Si as found from the Coverage Predictor

2. From the historical results of the Traffic Predictor, for each sector Si, find a set of vectors Ri(n) ¡rj, rk, . . . , ru¿, where each component of the vector is the PRB u+liza+on ra+o of the corresponding sector, such that it occurs concurrently and the magnitude of the distance between any two such vectors is greater than defined threshold. What does this mean? What is the point of this being taken?

3. SINR calculation. the strength of the wanted signal compared to the unwanted interference and noise. Mobile network operators seek to maximize SINR at all sites to deliver the best possible customer experience, either by transmitting at a higher power, or by minimizing the interference and noise.

4. CQI calculation - Table published by 3GPP.

Where Pi is the es+mated received power in the pixel from sector i (serving sector) and Pj, Pk, . . . Pu are the es+mated received interference power from all the adjacent sectors. PB is the background noise associated with Sector Si. All the above power values are in Wafs. If the input values are in dBm, they need to be converted to Wafs. The SINR value is a ra+o. If the value needs to be in dB, it must be converted accordingly.

C. Anomaly Detection Entity

Finally, the **Anomaly Detection Entity** compares real-time measurements with the results from the Traffic and Coverage Predictors to identify anomalies. If significant deviations between the measured and predicted values are detected, the energy-saving decision-making process is suspended to prevent potential issues.

Questions To Ask:
- More information is required for the Shutdown part - go through Doc1 again Pulak

VI. TRAFFIC PREDICTION MODEL: DESIGN RATIONALE

The standalone application for an ESC node described in 2.2 is connected to the OpenSAS. This application independently senses the CBRS spectrum for any activity. If activityis detected, it sends IQ data to the model running insidethe OpenSAS for incumbent detection. The current imple-mentation is to detect incumbent (radar) in a 5G New Radio(NR) based CBRS network deployment. Additionally, the re-searchers could use this platform to experiment with theirown models for detecting signals of their interest throughthe ESC node in testbed environments.

Network traffic prediction has always been a largely explored subject in networking, with a flurry of recent proposals ushered in by the recent development of machine and deep learning tools. Such deep learning-based algorithms have recently been explored to find potential representations of network traffic flows for all types of networks, including Internet, cellular, etc. We first categorize cellular traffic problems into two main types – temporal prediction problems and spatiotemporal prediction problems. Modelling the traffic flow through a node exclusively as a time series is an example of the temporal approach towards network traffic prediction [11]. High traffic on a given node in a cellular network often implies a high load on the other nearby nodes. Taking the traffic flow of nearby nodes and other external factors into consideration when modelling is known as the spatiotemporal approach to network traffic prediction. Spatiotemporal approaches are found to give slightly more accurate forecasts [12].

Both types of problems can be formulated as supervised learning problems with a difference being in the form of feature representation. In the temporal approach, the collected traffic data can be represented as a univariate time series and the prediction for the values in the future time steps is based on the historical data of the past time steps. In [13], Clemente et Al used Naive Bayes classification and the Holt-Winters method to perform the temporal network forecasting in real time Clemenete et Al first performed systematic preprocessing to reduce bias by selecting the cells with less missing data occurrences, which was then selected to train the classifies to allocate the cells between predictable and non- predictable, taking into account previous traffic forecast error.

Building upon the temporal approach, Zhang et al. [14] presented a new technique for traffic forecasting that takes advantage of the tremendous capabilities of a deep convolutional neural network by treating traffic data as images. The spatial and temporal variability of cell traffic is well captured within the dimensions of the images. The experiments show that our proposed model is applicable and effective. Even with the ease of machine learning implementations, regression based models have been found to be fairly accurate, as proven by Yu et Al in [15]. In [15], Yu et Al applied a switching ARIMA model to learn the patterns present in traffic flow series, where the variability of duration is introduced and the sigmoid function describes the relation between the duration of the time series and the transition probability of the patterns. The MGCN-LSTM model, presented in [16] by Len et Al, was a spatial-temporal traffic prediction model which implemented a multi-graph convolutional network (MGCN) to capture spatial features, and a multi-channel long short-term memory (LSTM) to recognise the temporal patterns among short-term, daily, and weekly periodic data. The proposed model was found to greatly outperform commonly implemented algorithms such as ARIMA, LSTM and ConvLSTM.

Hybrid models can handle a variety of data types and structures, making them ideal for diverse applications along

with combining the best features of different methodologies. This very principle is proven by Kuber et Al in [17] which proposes a linear ensemble model composed of three separate sub-models. Each sub-model is used to predict the traffic load in terms of time, space and historical pattern respectively, handling one dimension particularly. Different methodologies such as time series analysis, linear regression and regression tree are applied to the sub-models, which is aggregated and found to perform comparable to a ResNet-based CNN model. Another approach for the same is highlighted in [18] Tian et Al. The approach involves analysing the chaotic property of network traffic by analyzing the chaos characteristics of the network data. [18] proposes a neural network optimization method based on efficient global search capability of quantum genetic algorithm and based on the study of artificial neural networks, wavelet transform theory and quantum genetic algorithm. The proposed quantum genetic artificial neural network model can predict the network traffic more accurately compared to a similarly implemented ARMA model.

A. Model Selection

Model Selection. Why did we select the model we did?

B. Data Selection

Data Selection. How did we select data to train our model on? How did we know the model will work with less data?

C. Performance Metrics

Performance Metrics. Might not be needed if explained well in previous section.

VII. Performance Evaluation

- While QoS metrics are generic indicators of end-user performance, they do not necessarily reflect the quality of experience (QoE) associated with the specific application in use. Therefore, network performance optimization should also take into account QoE-related considerations in addition to existing QoS-driven primitives. In order to achieve this, it is imperative to first better understand the interplay be-tween QoS, QoE, network slicing, and RAN performance. To that end, this paper introduces AweRAN, an in-depth com-prehensive study on the impact of network slicing on RAN performance and QoS, along with the resulting QoE implica-tions for a variety of conventional and emerging end-user applications. why we consider what we do?

CQI + total throughput + energy consumption.

- This should otherwise describe the overall system architecture and where does the rApp reside and with which other components it interacts with.

- Setup
- Explain why we are taking these results in particular
- two/three seperate setups, show CQI and throughput are still good, while energy consumption reduces

VIII. Conclusions

Conclusions.

This work aimed to develop and evaluate an Energy Saving rApp for the O-RAN architecture using decision making algorithms and LSTMs for prediction. A LSTM neural network model was trained on various time-series datasets generated using the NIST radar (incum-bent) waveform generator. Also, 5G NR data was captured using the ESC sensor node with the SDR based CBSD as the signal source. The collected 5G NR data was mixed with the NIST generated datasets to train the ML model for 5GNR non-incumbent detection. The model training results show better performance at higher SNR values as expected. The highest prediction accuracy of 95.83% was achieved for [PULAK - Edit in the end] dataset with signals in the SNR range 40-50 dB. OTA incum-bent detection is achieved by transmitting the signals in the dataset OTA and results are presented. In the OTA results for incumbent detection, the highest accuracy achieved is 85.35%. Additionally, the results for the OTA non-incumbent(5G-NR) signal is presented. The model achieves an accuracy of 91.3

The energy saving results via ML-enabled rApp control in the the simulated NS-3 environment are encouraging and provide a basis for further enhancement in the ML model as well as the decision-making entity to incorporate other decision variables as the future scope of the work. Also, other prediction models can be implemented to analyze different model performances in the end-to-end experimental deployment. Furthermore, the enhanced rApp version provides an overall energy-saving solution to be used for efficient RAN control/management, not only in experimental simulations but also in any real-world environment.

- Summary

- Limitations and Future Work

IX. Acknowledgements

Acknowledgements.

X. Algorithms