Towards Greener Networks: RApp-Based Cell Control Over O-RAN Deployments

Anonymous Authors

ABSTRACT

Energy consumption is one of the most pressing challenges in the use of radio access networks (RANs), making it a critical area for research. Recent studies show that RANs contribute significantly to the total energy usage of mobile networks. The Open Radio Access Network (O-RAN) architecture, particularly with the use of xApps and rApps interfaced with RAN Intelligent Controllers (RIC), offers promising solutions for effective energy management in RANs.

In this work, we have developed a platform-agnostic Energy-Saving (ES) rApp which will interface with any O-RAN compliant networks. The rApp comprises multiple components, including a simple LSTM model for predicting the network's short-term energy needs, and a separate decision-making entity that decides which cells to switch off to save energy. [PULAK] more infomation of our testing environment The obtained simulation results demonstrate the energy efficiency and performance approach of the proposed control in terms of a [PULAK] After results.% reduction in power consumed as well as an increase in coverage [PULAK] After results.. This improvement is achieved without affecting the overall throughput and CQI of the connected UEs.

ACM Reference Format:
Anonymous Authors. 2024. Towards Greener Networks: RApp-Based Cell Control Over O-RAN Deployments. In *Proceedings of ACM Conference (Conference'17).* ACM, New York, NY, USA, 6 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With the rise in Internet literacy network providers face the challenge of creating a robust infrastructure that supports a growing number of users while also accommodating increasingly complex and data-intensive applications. This expansion, along with the advent of next-generation networks, results in larger and more intricate networks. As the number of network nodes increases, the energy required to maintain such dense and complex cellular networks also escalates, making their energy consumption a significant concern [might cite something here].

Radio Access Networks (RAN) have been found to be one the most significant users of a mobile network's total power supplied [Green Future Networks – Sustainability Challenges and Initiatives in Mobile Networks by NGMN Alliance, December 2021,]. [https://arxiv.org/abs/2301.06713]] In today's date, a RAN comprises of several cell sites with site infrastructure equipment and base station equipment. Existing RAN infrastructures are often built on older technologies that are not optimized for energy efficiency, making retrofitting for energy savings complex and costly. [might cite something here] To address the challenge of energy consumption in cellular networks, especially with the widespread usage of technologies like ultra-dense networks (UDNs) and network slicing [might cite something here], it is crucial to understand the current energy usage patterns. Studies have shown that 5G base stations consume approximately three times more power than their 4G counterparts, primarily due to the need for denser deployments and advanced technologies like massive MIMO antennas. [https://spectrum.ieee.org/5gs-waveform-is-a-battery-vampire]

On the other hand, virtualization makes it possible for network functions and resources to be performed and allocated to different parts of the networks in a dynamic matter thus making the RAN-as-a-service rather than as dedicated hardware as obtained in the previous generations of RANs [P. K. Thiruvasagam, V. Venkataram, V. R. Ilangovan, M. Perapalla, R. Payyanur, V. Kumar et al., Open RAN: Evolution of Architecture Deployment Aspects and Future Directions, 2023.]

Recent initiatives such as Open RAN (O-RAN) [ORAN Alliance, 2023, [online] Available: https://www.o-ran.org/.] have introduced the concept of RAN Intelligent Controllers (RICs) as a flexible platform for robust RAN control. O-RAN control is enbaled using applications called xApps (for Near-Real-Time RIC) and rApps (for Non-Real-Time RIC), with the choice made depending on the timeframe of the control. [might cite something here] O-RAN's open interfaces and standardized architecture allow for advanced algorithms that dynamically allocate resources based on real-time traffic demands, thereby optimizing energy usage.

The programmability of ORAN facilitates the deployment of AI-driven solutions that can predict traffic patterns and adjust energy consumption proactively, further enhancing efficiency. [https://www.tatacommunic
and-open-ran-how-they-are-transforming-the-telecom-industry/] The non-RT RIC, in particular, is designed to handle tasks that do not require immediate response, making it ideal for applications focused on long-term optimization and strategic planning. 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

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

Conference'17, July 2017, Washington, DC, USA

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

Unpublished working draft. Not for distribution.

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. The Non-RT RIC is responsible for the energy-saving algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is responsible for the algorithm execution. The algorithm is triggered by the SMO and the Non-RT RIC is responsible for the algorithm execution. The Non-RT RIC is

The expert reader may argue why we focus on implemeting a rApp instead of an xApp considering it's shorter timeframe of operation, and therefore control. There are several reasons.

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.).

- 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?
To the extent of our knowledge, none of the state-of-the-art approaches has tackled this issue. To fill this gap, we present in this paper a novel serial algorithm that has been designed to handle all network topologies compliant with the O-RAN architecture. - To bypass holes and ensure query completeness, the pro-posed algorithm leverages boundary traversal.
- To eliminate collisions, the proposed algorithm utilizes asingle packet to visit nodes, query them, and collect theiranswers.
- To ensure scalability, the proposed algorithm has been de-signed to rely only on the local one-hop information avail-able at the level of each visited node (no extra informationis needed).
- To be robust against failures and topology changes, the pro-posed approach does not require any structure building ormaintenance. The traversal path is gradually constructed byeach visited node.

- 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.

The remainder of this paper is organized into six sections, as follows. Section 2 provides some more background on the topic and a discussion of current approaches to energy saving with the RAN stack. Section 3 contains an overall overview of the rApp and it's functioning. Section 4 presents the architecture and the overall flow of the proposed energy saving algorithm. Section 5 presents the underlying rationale utilized by the proposed approach and then details the model selection and training. The results obtained from the rApp evaluation in the software-defined O-RAN simulation are depicted and discussed in Section 6. Finally, Section 7 concludes the paper and suggests future research directions.

2 ENERGY-SAVING RAPP OVERVIEW

Good overall overview of how the rApp works is required here. O-RAN defines rApps [O-RAN Alliance, "O-RAN Architecture Description," O-RAN.WG1.O-RAN-Architecture-Description-v03.00, July 2020.] as modular applications designed to consume and/or produces non real time management and automation services for the orchestration and optimization of network resources and operations. The non-RT RIC, in particular, is designed to handle tasks that do not require immediate response. This makes it ideal for applications focused on long-term optimization and strategic planning, such as energy control. The rApp is a data-driven application that uses machine learning algorithms to predict traffic patterns and optimize energy consumption in the RAN. The rApp receives input data from the Radio Database, Traffic Predictor, and Coverage Predictor, and sends a policy to the Near-RT RIC over what to do. The decision is made periodically, with a 1-hour prediction window

Unpublished working draft.
Not for distribution.

and 15-minute slots. The rApp is designed to be shared across multiple rApps and can import data from RF link simulators and drive tests through an external interface. A Dashboard for visualization of the Radio Mapping Database is also used.

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.

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

Space-time partitioning: This is a technique used to divide data based on both spatial (location) and temporal (time) dimensions. In the context of the rApp, this could involve organizing Key Performance Indicators (KPIs) by specific geographical areas (cells, sectors, etc.) and time periods to better manage and analyze the data.

Continuous time-based aggregation: This refers to the process of continuously collecting and summarizing data over time. Instead of analyzing data at discrete intervals, it is aggregated in a continuous manner, which allows for more fluid and accurate monitoring of KPIs.

Group KPIs by time: This involves organizing the Key Performance Indicators (KPIs) into groups based on the time they were recorded. This helps in analyzing trends and patterns over specific time periods.

3 RAPP ARCHITECTURE AND COMPONENTS

3.1 Digital Twin

Uses CloudRF to determine the coverage. Simulation which takes our parameters if need into context. More information required.

3.2 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.

3.3 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.

How are LSTMs used here? - every 1 hr, four predictions are made (+15, +30, +45, +60) - made on initial trained data (initial 300 entries from NS3 simulator) - inputs to the LSTM model? throughput, cell to which throughput belongs, timestamp

3.4 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 identified 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 identified 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 identified. - 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

3.5 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, identify 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 utilization ratio 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.

Unpublished working draft.
Not for distribution.

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.

For each sector S_i, for all the unique vectors $R_i(n)$, find the SINR for all the pixels in S_i using the relation:

$$\text{SINR}_{in} = \frac{P_i}{P_j + P_k + \ldots + P_u + P_B}$$

Where P_i is the estimated received power in the pixel from sector i (serving sector), and P_j, P_k , . . . , P_u are the estimated received interference power from all the adjacent sectors. P_B is the background noise associated with Sector S_i. All the above power values are in Watts (W). If the input values are in dBm, they need to be converted to Watts. The SINR value is a ratio. If the value needs to be in dB, it must be converted accordingly.

4. CQI calculation For each sector Si, find the average SINR and CQI values for all the pixels in the sector.
The CQI values are found by refering to this table published by 3GPP [[PRAMIT, MAKARAND] - Please provide this]

5. Energy Saving Decision Making
How is threshold defined?
Based on CQI value and threshold defined earlier, the sectors are classified into two categories:
- High CQI: Sectors with CQI values above the threshold. These sectors are not considered for shutdown.
- Medium CQI: Sectors with CQI values below the threshold but above a certain value. These sectors are considered for shutdown.
- Low CQI: Sectors with CQI values below a certain value. These sectors are considered for shutdown.

4 DESIGN RATIONALE

The standalone application for an ESC node described in 2.2 is connected to the OpenSAS. This application indepen-dently 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

Unpublished work in draft
Not for distribution

accurately compared to a similarly implemented ARMA model.

4.1 Model Selection

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

Model will handle missing data points during training and inferencing and raise appropriate flags

4.2 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?

4.3 Performance Metrics

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

5 CONCLUSION AND FUTURE WORK

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 capturedusing the ESC sensor node with the SDR based CBSD asthe signal source. The collected 5G NR data was mixed withthe NIST generated datasets to train the ML model for 5GNR non-incumbent detection. The model training resultsshow 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 thedataset OTA and results are presented. In the OTA resultsfor 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 accuracyof 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.

Unpublished working draft.
Not for distribution.

REFERENCES

Unpublished working draft.
Not for distribution.