



2022 Final Project Report
for
Academic Consortium for the 21st Century (AC21)
Special Project Fund

**Future Programmable Converged Wireless-
Optical Infrastructure for Beyond 5G/6G Networks**

Project Group Leader: Dr. Shih-Chun Lin, Assistant Professor

North Carolina State University

Date of Activities: March 2022 – February 2023



Contents

| | |
|---------------------------------|----|
| Project Abstract..... | 2 |
| Acknowledgment..... | 2 |
| Project Description | 3 |
| Activities and Reports..... | 7 |
| Achievement of Activities | 8 |
| Conclusion..... | 9 |
| Appendix | 10 |

Project Abstract

The steep traffic growth on the Internet (+20-30%/year) continues and even accelerates due to the COVID-19 pandemic. The widespread teleconferencing and working at remote sites boost global traffic by 50% in a year. We must bridge these end-to-end with minimized latency and broad bandwidth to fully take advantage of the 5G and future 6G mobile communication infrastructure and the data centers' considerable computation and storage resources. Led by North Carolina State University, aka NC State, this AC21 SPF 2022 project aims at UN's SDGs #9 (Industry, Innovation, and Infrastructure) and initiates collaborative research discussion and external grant planning for developing future programmable and resilient converged wireless-optical networks. Dr. Shih-Chun Lin (NC State), Dr. Hiroshi Hasegawa (Nagoya University), Dr. Peng Shi (University of Adelaide), Dr. Ta-Sung Lee (National Chiao Tung University), and Dr. Shao-Yu Lien (Institute for Information Industry) joined their efforts in this project with outstanding accomplishments in 2022-2023. These achievements include several joint proposal submissions, a seminar talk and teleconferences, technical paper publications, a technical workshop, and research awards.

Acknowledgment

The project team greatly acknowledges AC21 General Secretariat and the AC21 Special Project Fund (SPF) for their kind help and generous financial support.

Project Description

This project aims to initiate joint research discussion and external proposal development on building programmable converged wireless-optical infrastructure for future networks. Team members investigated zero-touch edge clouds and resource management, optically powered passive fronthaul networks and their architectural designs, programmable multi-petabit optical networks, and end-to-end network orchestration. We discussed working tasks to establish an experimental testbed and evaluate the testbed for beyond 5G/6G applications. The project empowers the realization of future large-scale communications infrastructure with the promised performance of high reliability, low latency, resilience, and cost-efficiency. The conducted activities promote creating a community interested in this new opportunity.

Based on the initial discussion, we mainly sought external grants during this project, reinforcing institutions' strategic research partnerships. We set several research discussions and implemented programmable wireless-optical systems and their possible enabling technologies. As a result, we got the following funding support for comprehensive research.

- Non-Terrestrial Integrated Access and Backhaul for 6G LEO Satellite MVNO, Cisco Systems, Inc., February 2023.
- Enabling Zero-Touch Prioritized Traffic Steering for Self-Healing Satellite Swarms, Lockheed Martin Corp., November 2022.
- DRL-ORAN Platform for Large-Scale Networking Resource Management, Meta, October 2022.
- Collaborative Research: NeTS: JUNO3: End-to-End Network Slicing and Orchestration in Future Programmable Converged Wireless-Optical Networks, NSF, September 2022.
- O-RAN A1 Interface Policy Management, Institute for Information Industry (III), September 2022.
- Towards Eigen-Spatial Filtering and Spreading for Anti-Jammed MIMO p-LEO Satellites, Lockheed Martin Corp., June 2022.
- 6G Serverless Computing Architecture with SLA Assurance for Cross-Constellation C3, NASA: North Carolina Space Grant, June 2022.
- Towards 6G SmartFab with SLA Assurance and Reconfigurable Multi-Robot Task Assignment, NC State 2022 FRPD, May 2022.
- Future Scalable and Resilient Converged Wireless-Optical Infrastructure for Beyond 5G/6G Networks: International Collaborative Research Planning, the Harry C. Kelly Memorial Fund, March 2022.

Notably, stemming from this AC 21 project, we submitted the comprehensive research and got a joint awarded project, funded by NSF Japan-US Network Opportunity (JUNO), to facilitate high-quality and long-term research collaboration. NC Japan Center at NC State also provided a supplemental fund to help with research engagement and international travel. Moreover, based on our designed platform, we worked with III to detail the development of intelligent controllers concerning different time granularities in decision and management. The results empowered beyond 5G/6G services and new end-


to-end applications. We also expanded the programmable platform to 6G satellite communications for a new project supported by the NC pace grant. The primary objective is to realize application programming interfaces (APIs) for cross-constellation control, command, and communications.

At the end of 2022, Dr. Lin conducted two and a half months of research and education exchange with several intense discussions and meetings in Taiwan to generate this project’s outcomes. A seminar talk on “AI-native federated networks,” shown in Figure 1, was delivered at National Taiwan University in December 2022.

- “AI-Native Federated Networks for 6G and Edge Intelligence,” *the GICE, National Taiwan University*, Taipei, Taiwan, December 19, 2022.



NC STATE UNIVERSITY



AI-Native Networks

NG intelligent networked edges

- **Networks for AI** (AI-aware networks)
 - Network edges leverage on-device & edge AI in coalition to solve large networking problems
 - E.g., multi-timescale AI algorithms
- **AI for networks**
 - New **foundational distributed AI** emerges to consider mobile network constraints and dynamics (domain knowledge); more than simply applying known AI techniques
 - Unique design opportunity: efficient, resilient, secure, adaptive intelligence

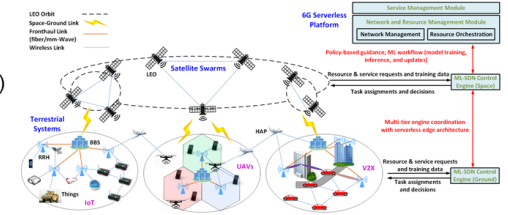


Figure 1. Dr. Lin gave a “AI-native federated networks” seminar at National Taiwan University.

This invited talk was one of the annual seminars for students in the school’s graduate institute of communications engineering. Dr. Lin presented the research and development insights of 6G networks and edge intelligence based on this AC 21 project’s results. This presentation shows our substantive engagement in ultra-low latency connected vehicle infrastructure, distributed intelligence over wireless edge networks, and 6G intelligent edge practices with non-terrestrial networks and end-to-end slicing orchestration. The details can be found in the attached flyer of this report.

In addition, we published three international conference papers and submitted a few journal articles, under review, in the research scope of the AC21 project. A completed list is provided below.

- K. V. S. Rohit, S.-C. Lin, and L. C. Chu, “SPELS: Scalable and Programmable Testbed for Evaluating LEO Satellite Swarm Communications,” in *Proc. of IEEE INFOCOM Workshop*, New York area, USA, May 2023.
- D. Haro-Mendoza, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, L. Marrone, S.-C. Lin, “On the Resource Allocation for Radio Access Network Slicing in Cellular IoT with Massive Traffic,” in *CSCI*, Las Vegas, USA, December 2022.
- D. Haro-Mendoza, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, L. Marrone, S.-C. Lin, “Modeling the Resource Allocation in 5G Radio Access Networks with Network Slicing,” in *CSCI*, Las Vegas, USA, December 2022.
- S.-C. Lin, C.-H. Lin, L. C. Chu, and S.-Y. Lien, “Enabling Resilient Access Equality for 6G LEO Satellite Swarm Networks,” under review, 2022.
- C.-H. Lin, K. V. S. Rohit, S.-C. Lin, and L. C. Chu, “6G-AUTOR: Autonomic CSI-Free Transceiver via Realtime On-Device Signal Analytics,” under review, 2022.
- S.-C. Lin, C.-H. Lin, and M. Lee, “Privacy-Preserving Serverless Edge Learning with Decentralized Small Data,” under review, 2022.
- S.-Y. Lien, Y.-C. Huang, C.-C. Tseng, S.-C. Lin, C.-L. I, Xiaofei Xu, and D.-J. Deng, “Universal Vertical Application Adaptation for Open RAN: Sustainable RIC and Autonomous Intelligent xAPP Generation,” under review, 2022.
- M. F. Pervej, R. Jin, S.-C. Lin, H. D, “Efficient Content Delivery in Cache-Enabled VEN with Deadline-Constrained Heterogeneous Demands: A User-Centric Approach,” under review, 2022.

Particularly, the SPELS work was published by the top-tier communications conference, the IEEE International Conference on Computer Communications (INFOCOM). It introduces a scalable and programmable OTA (over-the-air) testbed to provide a real-time architectural implementation of satellite swarm systems and demonstrate the testbed’s effectiveness in online swarm communications. Experimental evaluations validate the superiority of our swarm combiner with learning-enabled channel coding for online frontend operations, thus facilitating LEO swarm readiness. Besides, we studied the radio resource allocation problem of a 5G gNB on an uplink random access channel concerning different user traffic types. Network slicing solutions were proposed to assign each slice’s preambles on the service priority; the corresponding random access procedure can maximize users’ successful access probabilities in each slice. The results were published in Computational Science & Computational Intelligence (CSCI’22).

Moreover, after a few rounds of discussions and thorough preparation, Dr. Lin and several faculties organized a workshop on “NG-OPERA: Next-Generation Open and Programmable Radio Access Networks” in the IEEE INFOCOM 2023. As shown in Figure 2, this technical workshop focused on state-of-the-art and practice solutions for addressing critical challenges in developing and operationalizing open radio access networks. The workshop will be held with the leading conference in the New York area on May 20, 2023.



Figure 2. IEEE INFOCOM 2023 NG-OPERA workshop, May 20, 2023.

We also received research and demo awards closely related to this AC21 project from an industrial company and government agency, as shown below.

- **2022 AI4AI Research Award**, “DRL-ORAN Platform for Large-Scale Networking Resource Management,” Meta, 2022.
- **Finalist**, “Data-Driven Modulation and Coding for Beyond 5G Communications,” 4th Annual Beyond 5G SDR University Challenge, Air Force Research Laboratory (AFRL), 2022.

Activities and Reports

| | |
|------------------------------|--|
| Mar. 2022 | <p>Have several discussions for the project roadmap and proposal preparation for potential external grants.</p> <ul style="list-style-type: none"> Participating members initiated teleconferencing meetings to discuss a detailed timeline and tasks for the research activities. |
| Apr. 2022 | <p>Conducted beyond 5G communications demo at the 4th Annual Beyond 5G SDR University Challenge at Air Force Research Laboratory (AFRL). https://www.wbi-innovates.com/blogs/post/5G-awards</p> |
| May – June 2022 | <p>Conducted research on programmable converged wireless-optical networks and their extended use cases.</p> <ul style="list-style-type: none"> Prepared several related journal manuscripts and external proposals. |
| July 2022 | <p>Engaged with Institute for Information Industry (III) to expand this project with ORAN interface designs.</p> |
| Aug. 2022 | <p>Engaged with Meta Research to extend this project in large-scale networking scenarios. https://research.facebook.com/research-awards/2022-ai4ai-research-request-for-proposals/</p> |
| Sep. – Oct. 2022 | <p>Prepare conference papers to be submitted to IEEE INFOCOM and CSCI https://www.american-cse.org/csci2022/</p> <ul style="list-style-type: none"> Summarized the results for SPELS implementations and new 5G gNB network slicing designs. |
| Nov. 2022 | <p>Prepared a technical workshop.</p> <ul style="list-style-type: none"> Organized a workshop on Open and Programmable Radio Access Networks in IEEE INFOCOM 2023 https://infocom2023.ieee-infocom.org/next-generation-open-and-programmable-radio-access-networks-ng-opera |
| Dec. 2022 – Feb. 2023 | <p>Two and a half months visit for research and education engagements.</p> <ul style="list-style-type: none"> Visited a non-AC21 institution to deliver a seminar talk about the AC21 project outcomes. Discussed potential student exchange opportunities and the project’s sustainability for the planned scope and impact beyond the AC21 grant period. Reinforced the complementary strengths of the participating institutions’ strategic partnership. |
| Feb. 2023 | <p>Wrapped up the project.</p> <ul style="list-style-type: none"> Provided a final report, newsletter, and technical papers. This collaborative project successfully empowered the realization of future programmable converged infrastructure for new beyond 5G/6G services with the promised performance of high reliability, low latency, resilience, and cost-efficiency |

Achievement of Activities

The longstanding relationship and strategic partnership among NC State, Nagoya University (NU), and the University of Adelaide (AU) leverage complementary strengths and transdisciplinary scholarship to advance research collaboration and academic exchanges. By jointly planning collaborative research activities, the project reinforces three university linkages through communications between PIs' teams. The leading PI Lin has already established a very strong collaboration with PI Shi and NU's faculties during the last two AC21 projects. The PIs used essential conference forums, e.g., IEEE INFOCOM and CSCI, to advance industrial partnerships and achieve funding success. This project stimulated international collaboration with the following contributions and broader impacts on research, education, and international exchange.

Research: This project outcomes systematically accomplished programmable converged wireless-optical networks for beyond 5G/6G services and new end-to-end applications. It led researchers' and students' substantive engagement in optical networks and communications, power fiber designs, system programming, telecommunications, and wireless networking. The discussion involved the multidisciplinary knowledge of edge clouds, power over fibers, fronthaul architecture design, multi-petabit optical networks, and end-to-end network orchestration.

Education: The project results were incorporated into Optimizations and Algorithms and Introduction to Computer Networking courses at undergraduate and graduate levels in the ECE Department at NC State, the school of EEE at AU, and the IMaSS at NU. The PIs also trained their Ph.D. students to become experts in this fast-evolving field and involved M.S. and undergraduate students in the proposed research by assigning them sub-problems to solve.

International Exchange: The PIs also plan their institutions' visiting/exchange Ph.D. student programs. Exchanging students will enhance the cross-linkage among team members' research lines and further strengthen the AC21 network.

Conclusion

This project has successfully initiated joint research discussions, seminars, external grant preparation, and workshop organization for beyond 5G/6G networks. Several academic and research activities were conducted to place participating AC21 members in a unique leading position. We foresee more engineers and researchers will engage in developing our proposed programmable converged wireless-optical infrastructure and continue to foster our strategic partnerships based on this project's outcomes.

AI-Native Federated Networks for 6G and Edge Intelligence

Speaker : Dr. Shih-Chun Lin

Assistant Professor of North Carolina State University

Time : 2022/12/19 2:20-3:10p.m.

Location : BL112



Abstract

Along with unprecedented growths of AI and its applications, networking technologies have become the other most transformative technological paradigm for future wireless systems and distributed intelligence. An AI-native network aims at enabling the synergy of AI and networking to facilitate intelligent network management and support emerging AI services. It indicates that AI exists as a built-in architectural component in the software-defined networking controller for managing network resource slices and in network slices as services for end users. Hence, on the one hand, intelligent network edges can leverage on-device/edge AI in coalition to resolve large-scale networking problems (i.e., networks for AI). On the other hand, new foundational distributed AI emerges by investigating mobile network constraints and dynamics (i.e., AI for networks).

In this seminar, the speaker will give key aspects of designing AI-native federated networks and identify a set of exciting research directions through two case studies involving connected vehicle transportation and federated multi-task learning. Further, the speaker will also describe the latest 6G edge network practice and how AI techniques can be developed to empower system performance.

Biography

Dr. Shih-Chun Lin received his Ph.D. from Georgia Institute of Technology, Atlanta, USA, in 2017 and immediately joined North Carolina State University as an Assistant Professor. At NC State, he leads the Intelligent Wireless Networking (iWN) Laboratory. His lab was a finalist in 2022 Beyond 5G SDR University Challenge, hosted by the Air Force Research Laboratory. His team also won the Creativity Technology Award in 5G system software in 2021 5G Craft by the Taiwan Ministry of Economic Affairs.

Dr. Lin's research interests span the areas of communication, networking, machine learning, mathematical optimization, and hardware/software implementation. He is particularly interested in wireless software-defined infrastructure, federated edge learning, and performance evaluation. He has published more than 50 peer-reviewed papers and 12 U.S. patents. He received the Distinguished TPC Member Award in IEEE INFOCOM 2020.



SPELS: Scalable and Programmable Testbed for Evaluating LEO Satellite Swarm Communications

K V S Rohit
iWN Lab, ECE Department
North Carolina State University
Raleigh, NC, USA
vkanthe@ncsu.edu

Shih-Chun Lin
iWN Lab, ECE Department
North Carolina State University
Raleigh, NC, USA
sclin23@ncsu.edu

Liang C. Chu
Lockheed Martin Space Systems
Company (LMSSC)
Sunnyvale, CA, USA
liang.c.chu@lmco.com

Abstract—Low earth orbit (LEO) satellite communications promise next-generation mobile networks with seamless connectivity to rural, remote, and inaccessible areas. Notably, due to low-cost deployment and quick turn-around times in production, proliferated LEOs deployed and orchestrated as a swarm of satellites can support ultra-broad transmissions for the ever-evolving communications and aid current wireless network infrastructure. This paper introduces a scalable and programmable OTA (over-the-air) testbed, called SPELS, to provide a real-time architectural implementation of satellite swarm systems and demonstrate the testbed's effectiveness in online swarm communications. First, the in-lab SPELS testbed is established with COTS (commercial off-the-shelf) software-defined radios, a high-performance host computer, and wireless softwarization. Accordingly, the latest AI-enabled wireless communications and real-time signal processing constraints can be easily realized upon various frontends by decoupling radio swarm networks' control and data planes. Furthermore, based on the designed infrastructure, an end-to-end module is proposed for timely and resilient satellite swarm communications. This module consists of swarm-MRC, an optimal swarm combining technique, and a 5G-compliant deep learning-based LDPC scheme. Experimental evaluations validate the superiority of our swarm combiner with learning-enabled channel coding for online frontend operations, thus facilitating LEO swarm readiness.

Index Terms—Proliferated LEOs, satellite swarm communication, programmable testbed, swarm-maximal ratio combining, low-density parity-check coding, data-driven communications.

I. INTRODUCTION

LEO satellite communications development and their integration with the existing terrestrial radio access networks is a highly-researched area, with a motivation to provide cost-effective and high-capacity connectivity to rural, remote and other inaccessible areas [1]–[3]. In the context of 6G and beyond infrastructure, such a system exploits the mobility of satellite nodes, and inter-satellite communication to provide enhanced coverage and high performance in terms of data rates and spectral efficiency, thereby adhering to the massive machine type communication (mMTC) and ultra-reliable low latency communication (URLLC) requirements easily compared to the GEO (geosynchronous-earth orbit) and MEO (medium-earth orbit) deployments [4]–[6]. Incorporating LEO

satellites as an extension to terrestrial communication has been proposed through a *transparent* satellite deployment between the disaggregated next-generation radio access network structures in [5], which can be complemented by additional roles suitable for the Internet-of-things (IoT) ecosystems [6] and edge computing frameworks [7].

However, some seminal hindrances have been identified here, which need to be addressed to cater to the aforementioned needs [8]–[10]. One of the most predominant factors is the lack of diversity in wireless channels owing to a strong line-of-sight (LOS) property. As a means to overcome this issue, a swarm structure consisting of proliferated single-antenna LEO satellites is being proposed, to work as a distributed antenna transceiver system. Owing to the joint operation of several satellites in the swarm structure, ultra-wide area coverage is enabled, all the while providing the necessary receiver and spatial diversities to improve the data rates. The ground terminals communicating with the satellite nodes can constitute devices of varied performance levels, from a single-antenna transceiver with limited processing capability to high-end communication systems embedded with a large array of antennas enabling massive MIMO implementation.

With an aim to demonstrate the effectiveness of our proposed architecture, we have developed **SPELS**: a Scalable and Programmable testbed for Evaluating LEO satellite Swarm communications. We use USRP B210s as the ground terminal and satellite swarm nodes, communicating with each other over-the-air (OTA), in an indoor laboratory environment. The reconfigurability of USRPs in terms of transmission/reception gains and local oscillator (LO) frequency tuning, and an external host PC-aided signal processing help create a semi-controlled indoor environment to emulate LEO swarm communications. Owing to high computational capacity of the host PC, it is also plausible to incorporate intelligence by utilizing machine learning-based models in the data transmission mechanisms, and also expanding its usage to higher layer applications such as software defined networking and edge computing. Based on the designed infrastructure, we have developed an end-to-end communication module to illustrate the effectiveness of swarm structure, for realizing robust satellite communications in real time. This module constitutes swarm maximal ratio combining (swarm-MRC) as an optimal receiver

This work was supported in part by LMSSC, the NC Space Grant, the AC21 Special Project Fund (SPF), the National Science Foundation (NSF) under Grant CNS-221034, and Meta 2022 AI4AI Research.

diversity technique, and a 5G-compliant deep learning-aided LDPC channel coding scheme for building error correction capability in individual nodes. The swarm-MRC employs data aggregation and channel estimation at swarm nodes to improve the received data rate, based on the maximal ratio combining (MRC) diversity technique in terrestrial communications. A deep learning model has been utilized to implement LDPC decoding, to incorporate intelligence into our architecture, thereby realizing a softwarizeable radios and access network.

The remainder of the paper is categorized systematically to explore the testbed development and implementation of the aforementioned schemes in a real-time OTA environment. Section II provides the designed SPELS testbed development. Section III presents our proposed end-to-end satellite swarm communications. Section IV demonstrates the results from swarm-MRC and LDPC channel coding on receiver data rate improvement. Section V concludes the paper by presenting several opportunities of exploration for enhancing our existing testbed, to both improve our current testbed performance and to incorporate new additions for catering the needs of the 6G ecosystem.

II. SPELS TESTBED DEVELOPMENT

This sections details the realization of the swarm architecture using SPELS. It also discusses a theoretical LOS channel model for LEO satellite communications and the practical channel estimate mechanism used for our experiments.

A. Swarm System Design and Theoretical Channel Model

Proliferated LEO satellite swarms provide unprecedented opportunities for enabling ubiquitous wireless coverage owing to a large population of satellite nodes leading to a global footprint [11], while also exploiting the inherent diversity to improve the communication link performance. Fig. 1 shows a satellite swarm architecture constituting a ground-to-space communication where the ground terminal is equipped with N_T transmitting antennas and the satellite swarm consists of N_S single-antenna satellite nodes. D_S defines the inter-satellite distance, which is a fixed entity decided by the motion trajectory of the satellite swarm. We also define d_l^i as the distance between the i^{th} transmission antenna and the l^{th} satellite node, which is generalized as d_l to represent all distances from the ground terminal to the said node.

LEO satellites serving non-urban areas have a large LOS probability [12], and thus, we consider a pure LOS channel model in this paper. We begin with a ground-to-space SISO link, where the data transfer can be represented as [13], [14]:

$$y(t) = h \cdot e^{j(2\pi f_{\text{off}} t)} \sum_{n=-\infty}^{\infty} x[n]g_s(t - nT_0 - \epsilon T_0) + \eta(t), \quad (1)$$

where $x[n]$ is the encoded transmitted symbol, $y(t)$ is the baseband data captured at the receiver, h is the complex wireless channel response, f_{off} is the frequency offset observed at reception, $g_s(\cdot)$ is a pulse shaping filter, ϵ is the timing offset, T_0 is the symbol period and $\eta(t)$ is the additive gaussian

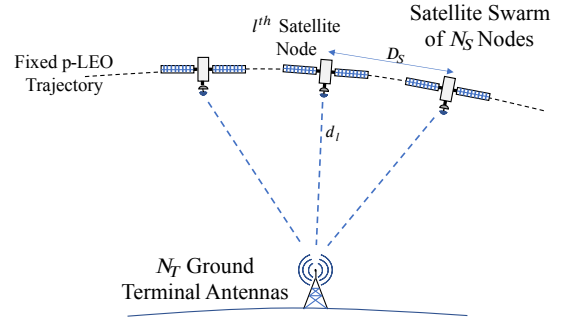


Fig. 1. Proliferated low earth orbit swarm architecture

noise. Specifically for a LEO communication scenario, the complex channel component h can be further described using [8]

$$h = \frac{1}{\sqrt{L(d, f_c)}} e^{-j(kd + \phi_{\text{atm}})}, \quad (2)$$

where $k = 2\pi f_c / c_0$ is the wave number of the transmitted signal, f_c is the central frequency of operation (in GHz), c_0 is the speed of light, d is the distance between the ground station and the LEO satellite node, ϕ_{atm} is the uniformly distributed phase shift due to the atmosphere and $L(d, f_c)$ (in dB) is the path loss, which, as per [12], can be dissected into free space path loss $PL_{\text{fs}}(d, f_c) = 32.45 + 20 \log_{10}(f_c d)$, shadow fading loss (PL_{sf}), clutter loss (PL_{cl}), attenuation from atmospheric gases (PL_{g}) and ionospheric or tropospheric scintillation losses (PL_{sc}) [15] depending on the operating central frequency:

$$L(d, f_c) = PL_{\text{fs}}(d, f_c) + PL_{\text{sf}} + PL_{\text{cl}} + PL_{\text{g}} + PL_{\text{sc}}. \quad (3)$$

For the proposed satellite swarm architecture in fig. 1, each of the N_S satellite nodes are sufficiently distant from each other, and if it is ensured that the N_T transmitting antennas are spaced at least $f_c / 2c_0$ farther in succession, then all the communication channels can be considered uncorrelated and a channel matrix $\mathbf{H} \in \mathbb{C}^{N_S \times N_T}$ can be obtained as a collation of channel vectors at each satellite node (\mathbf{h}_l):

$$\mathbf{H} = [\mathbf{h}_1 \ \mathbf{h}_2 \ \dots \ \mathbf{h}_l \ \dots \ \mathbf{h}_{N_S}]^T \quad (4)$$

where, $[\cdot]^T$ is the transpose operation and channel vector at l^{th} satellite node is defined as an array of channel values from all transmitting antennas $\mathbf{h}_l = [h_l^1 \ h_l^2 \ \dots \ h_l^{N_T}]$.

And communication between the i^{th} transmitting antenna and l^{th} satellite node would then be represented, using (2) and (3), by the channel h_l^i as

$$h_l^i = \frac{1}{\sqrt{L(d_l^i, f_c)}} e^{-j(kd_l^i + \phi_{\text{atm}, l})}, \quad (5)$$

where $i \in \{1, 2, \dots, N_T\}$, $l \in \{1, 2, \dots, N_S\}$ and $\phi_{\text{atm}, l}$ is the uniform phase shift due to atmosphere at satellite l .

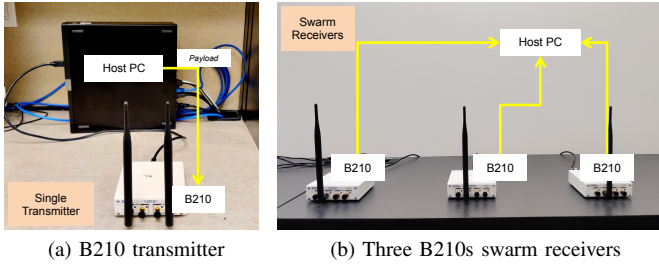


Fig. 2. Testbed Hardware Setup

B. SPELS Hardware Setup

In order to evaluate the proposed swarm structure and ground-to-space communication, we have developed an easily scalable and programmable indoor testbed setup by employing the USRP B210s to act as both the ground terminal transmitter and satellite swarm receivers. To stay as close to real-life deployments as possible, we built the testbed without an anechoic chamber, while physically maintaining a LOS connection at all times. Since the transmitter and receivers do not share a common LO, a random phase shift is introduced into every new transmission and reception [16], which emulates the effect of different channel conditions at different swarm nodes in a practical deployment. Also, a USRP B210 supports two wireless channels simultaneously, therefore, it can be used either as a single-antenna radio for SISO/SIMO applications or as a dual-antenna radio, allowing its usage for MIMO applications. For implementing the swarm-MRC, the ground terminal was emulated using a single-antenna radio transmitter and the swarm was emulated using three single-antenna receivers. The testbed currently has stationary nodes acting as swarm transceivers, which will be enhanced in the future with UAVs to enable mobility among the swarm receiver nodes. The testbed hardware setup is as depicted in the fig. 2.

All the USRPs are connected to a host PC, in order to enable an end-to-end data flow from the transmitter to the receiver. The PC interacts with the radios using a software tool, which interfaces with the USRP Hardware Driver (UHD) for configuring them and sending out/capturing symbols. MATLAB has been chosen for this purpose in our experiments, owing to its extensive documentation, hassle-free integration with radio-independent signal processing packages and ease of documentation.

C. Preamble-Based Channel Estimation

For the purposes of over-the-air experimentation, one needs to perform channel estimation based on the real data transferred between the transceivers. We have used a single-tap time-domain filter to represent the wireless channel, which was then estimated by performing a correlation between the transmitted preamble sequence at the ground terminal and the captured preamble sequence at the receiving satellite nodes [17].

If $x[n]$ is the transmitted preamble sequence and $y_l[m]$ represents the sequence received at l^{th} satellite node in the

swarm, then the channel estimate at this node in time domain is obtained using \hat{h}_l :

$$\begin{aligned}\hat{h}_l &= \text{corr}(y_l[m], x[n]) \\ &= y_l[m] * x'[-n] \\ &= h_l[m] * x[m] * x'[-n] + \eta_l[m] * x'[-n]\end{aligned}\quad (6)$$

where $\text{corr}(\cdot)$ is a correlation operator, x' represents a complex conjugate of x , $h_l[m]$ is the actual channel value at l^{th} satellite, for which we are finding an estimate using \hat{h}_l , and $\eta_l[m]$ is the i.i.d noise at satellite l .

From (6), the term $x[m] * x'[-n]$ represents the autocorrelation of preamble sequence at the receiver end, with respect to the transmitted sequence. Since preamble sequences with good autocorrelation properties are chosen for communications, $m = n$ represents a time when the captured preamble coincides the most with the transmitted sequence, resulting in close-to-impulse response. That is, $x[m] * x'[-n] \approx 1$ when $m = n$ and approximates to 0 otherwise. Then, the correlation in (6) reduces to

$$\hat{h}_l = h_l[m] + \hat{\eta}_l[m],\quad (7)$$

where $\hat{\eta}_l[m] = \eta_l[m] * x'[-m]$ is the distortion in channel estimate due to the presence of noise.

Since the channel is assumed to be a single-tap value, \hat{h}_l in (7) represents the best channel estimate for the satellite node under the assumed conditions.

III. END-TO-END SATELLITE SWARM COMMUNICATIONS

In this section, we develop an optimal swarm receiver combining using the swarm-MRC. We also investigate deep learning-aided 5G-compliant LDPC channel coding.

A. Optimal Receive Diversity via Swarm-MRC Provisioning

By incorporating a swarm structure for realizing a satellite receiver, we have introduced uncorrelated channels at each node represented by (5). This implies a higher possibility of at least some channels having better data transmission conditions compared to others at most times. Thus, if we were to combine individual satellite node performances in proportion to their channel state information estimates, better channels will be given higher weightage and lower performing channels will have a smaller effect on the overall output.

Let us consider a wireless communication system involving a single-antenna ground terminal acting as a transmitter ($N_T = 1$) and multiple single-antenna B210 receivers emulating the satellite swarm nodes with LOS channels. Swarm-MRC combines the outputs from individual satellite nodes such that the overall SNR can be maximized during the operation, thereby resulting in improved effective data rates, compared to any individual receiver performance.

It is significant to note that, while the swarm structure aids in diversity, the distributed nature of receiver nodes implies the existence of varied frequency offsets and timing offsets among them due to uncorrelated LOs and clocks. This is in contrast with a centralized receiver architecture, where the

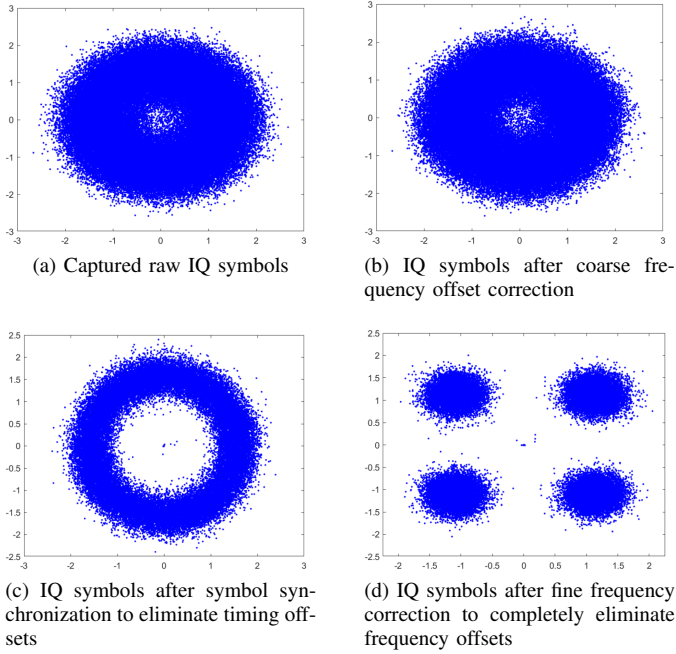


Fig. 3. Postprocessing of captured symbols at individual receiver nodes

offsets obtained at all antennas would be identical, since a single LO and a clock reference would be shared among all receiving circuits. Thus, for a swarm architecture to work well, the offsets at individual receiving antennas should be eliminated within the satellite node, before performing any optimization scheme.

We have achieved this offset compensation in individual nodes by utilizing the *Communications Toolbox* in MATLAB in a modular fashion. A *Coarse Frequency Compensator* functional block was first employed to eliminate large frequency offsets in the received signals. This was followed by eliminating timing offsets by identifying suitable samples from the oversampled captured signal using the *Symbol Synchronizer*. Finally, a PLL-based *Carrier Synchronizer* was utilized to eliminate any residual finer frequency offsets. Fig. 3 shows the offset correction in at one of the receivers, after each correction phase.

Looking back at the architecture, since $N_T = 1$, from (1) and (5), considering the sampled, offset-corrected received data sequence as $y_l[n]$ at l^{th} receiver node,

$$y_l[n] = h_l x[n] + \eta_l \quad (8)$$

where, $l \in \{1, 2, \dots, N_S\}$, $x[n]$ denotes the transmitted data sequence, $h_l^1 = h_l$ is the channel response for the respective receiver and η_l is the i.i.d noise at the satellite node l .

With the aim to combine the data from all receivers, we have utilized a combining weight vector $\mathbf{w} = [w_1 \ w_2 \ \dots \ w_{N_S}]$, which results in a combined received data sequence as $y_{\text{out}} =$

$\mathbf{w}\mathbf{y}$, where $\mathbf{y} = [y_1 \ y_2 \ \dots \ y_{N_S}]^T$,

$$y_{\text{out}} = \mathbf{w}\mathbf{y} = \sum_{l=1}^{N_S} w_l y_l = \sum_{l=1}^{N_S} (w_l h_l x + w_l \eta_l). \quad (9)$$

If p_{tx} is power of the transmitted signal, then SNR for the combined result is expressed as [18]:

$$\text{SNR} = \frac{p_{\text{tx}} \sum_{l=1}^{N_S} \|w_l h_l\|^2}{N_0 \sum_{l=1}^{N_S} \|w_l\|^2}, \quad (10)$$

where N_0 is the AWGN noise power level. Swarm-MRC aims to maximize the signal-to-noise ratio defined in (10) by deriving a suitable weight matrix \mathbf{w}_{MRC} using the Cauchy-Schwarz inequality [13],

$$\mathbf{w}_{\text{MRC}} = \underset{\mathbf{w}}{\text{argmax}} \frac{\sum_{l=1}^{N_S} \|w_l h_l\|^2}{\sum_{l=1}^{N_S} \|w_l\|^2} = \mathbf{h}^H, \quad (11)$$

where $\mathbf{h} = [h_1 h_2 \dots h_{N_S}]$ and $(\cdot)^H$ is a Hermitian operation.

The weighted combination $y_{\text{out}} = \mathbf{w}_{\text{MRC}}\mathbf{y}$ represents the optimized output at the end of swarm-MRC scheme in our ground-to-space uplink communication.

B. Deep Learning Aided Low-Density Parity-Check Codes

LDPC codes are a form of soft-in soft-out linear codes, where the operation adds a certain number of redundant bits to a binary message, resulting in a larger codeword, to provide robustness to errors due to harsh channel environments. LDPC decoders take in soft demodulated bits as inputs (represented by log-likelihood ratios or LLRs) and give out soft outputs, as an estimate of the codeword. The relation between inputs and outputs is represented using a parity check matrix \mathbf{P} . If ‘ c ’ is an encoded codeword, channel coding defines $\mathbf{P} \cdot c^T = 0$.

Traditionally, parity check matrices are represented as bipartite graphs known as Tanner graphs. The first node type of the bipartite graph is a ‘variable node’ that corresponds to the encoded codewords/columns and the second node type is a ‘check node’ that represents the linearly independent matrix rows. Edges of the graphs represent the non-zero inputs of \mathbf{P} . LDPC decoding can be realized as an iterative process consisting of two LLR exchange operations between the check nodes and the variable nodes.

We have adopted a neural network-based LDPC decoder presented in [19], where each layer has two sets of nodes (variable and check) aligned with the Tanner graph representation, and extended its functionality from a simulation environment to real-time OTA experiments. The decoder utilizes a protograph-LDPC BG2 parity check matrix defined by 3GPP for 5G [20], with a lifting factor of $Z = 16$. This corresponds to an input size of $52 \times 16 = 832$ bits. Each layer consists of neurons corresponding to the number of edges in the Tanner graph that represents the lifted BG2 matrix. Here, a min-sum (MS) decoding algorithm was employed which consists of a single-parity-check operation at each check node and a repetition code operation at the variable nodes. The iterative implementation of decoder is realized by unfolding this operation into layers in the neural network structure.

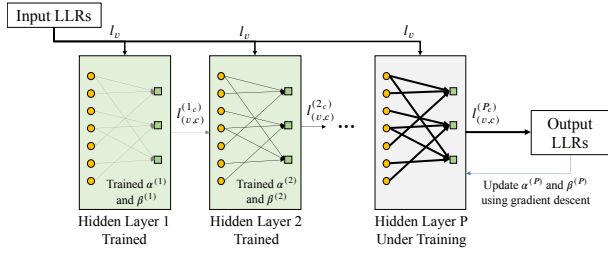


Fig. 4. Neural Network showing ‘ P ’ hidden layers, with ‘ P_{th} ’ layer being trained using gradient descent. Since the training is iterative in nature, first layer gets trained completely to identify α_1, β_1 , then the next layer is added and the process is repeated, and so on

The network was built as a deep neural network structure consisting of 25 hidden layers, in order to accommodate for training large input sequences. Authors in [19] had utilized such a large structure to train and infer inputs from multiple lifting factors simultaneously, while our adaptation uses a single lifting factor and extends its application to OTA implementation.

Considering an edge in the parity check matrix $e = (v, c)$, where v represents a variable node and c represents a check node, let $\mathcal{E} = \{e = (v, c)\}$ be defined as a set of all edges in the neural network. LLR inputs into the neural network are given as l_v . For i^{th} hidden layer, if LLR update for its variable node is represented as l_e^v and that for the check node is given as l_e^c , for edge e , then the update operation is performed as

$$l_{e=(v,c)}^v = l_v + \sum_{e'=(v,c'), c' \neq c} l_{e'}^{(i-1)c}, \quad (12)$$

and check node update l_e^c for edge e is defined as

$$l_{e=(v,c)}^c = \left(\prod_{e'=(v,c'), v' \neq v} \text{sgn}(l_{e'}^v) \right) \times \text{ReLU} \left(\alpha_e^i \times \min_{e'=(v,c'), v' \neq v} |l_{e'}^v| - \beta_e^i \right), \quad (13)$$

where $\text{sgn}(\cdot)$ is the sign operator, α^i , the normalization factor, is represented by weights of i^{th} hidden layer and β^i , the offset factor, is represented by bias of i^{th} hidden layer.

For the purpose of training and obtaining suitable network parameters, random messages of 832 bits were generated as a synthesized training dataset. In order to eliminate the problem of vanishing gradients, the network was trained in an iteration-by-iteration fashion, where each layer was trained independently using gradient descent and cross-correntropy loss. Fig. 4 show a diagrammatic representation of the training process.

IV. PERFORMANCE EVALUATION DEMONSTRATION

This paper emphasizes on the development of a testbed and evaluating an end-to-end communication module based on the proposed LEO satellite swarm architecture, to provide resilient ground-to-space communication. While the inter-satellite link (ISL) communications in the proposed architecture can be realized through prevalent high-speed optical ISL mechanisms

[21]–[23], since the receiver nodes are emulated by software defined radios deployed in an indoor environment, the optical ISL is replaced by operations on a host computer. For our experiment, we have utilized a central frequency of 1.85GHz, operating at a narrowband bandwidth of 1MHz, and VERT900 isometric antennas to provide a 3dBi gain at transmitter and receiver nodes.

A single-carrier mode of data transfer was adopted for our experiments, with data modulated using the quadrature phase shift keying (QPSK) scheme. The transmitter power was varied over a range of values by reconfiguring the low-noise amplifiers in the radios, in order to evaluate the receiver performances at each stage. We have used a 12-core Ubuntu 20.04 machine with 16GB memory installed and MATLAB R2022b for radio interfacing, radio configuration, channel encoding/decoding and signal post-processing after reception.

A. Swarm Maximal Ratio Combining

In our demonstration of swarm-MRC, for the sake of completeness, we have implemented the equal ratio combining (ERC) and selection combining (SC) methods, in addition to MRC, to prove that the swarm-MRC weighted combining scheme delivers the best performance in comparison to other prevalent schemes. If \mathbf{h} is a vector containing channel gains from all receivers, for a particular symbol i , weights for each of the combining methods is given as,

$$\begin{aligned} \mathbf{w}_{\text{MRC}} &= |\mathbf{h}|e^{-j\angle \mathbf{h}}, \\ \mathbf{w}_{\text{ERC}} &= e^{-j\angle \mathbf{h}} \quad \text{and} \\ \mathbf{w}_{\text{SC},i} &= \begin{cases} 1 & \text{if } |h_i|^2 \text{ is the maximum among all receivers} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \quad (14)$$

While both ERC and MRC schemes combine results from all receivers through weighted vector operations, unlike selection combining which chooses only the best performing receiver output, the swarm-MRC scales individual receiver outputs by their channel estimates, which accentuates the performance from a better quality channel while decreasing the distortion due to lower quality channels. This property is described in (14) and is also evident from the practical experiment setup for swarm-MRC operation as shown in the Fig. 5.

B. Data Rate Enhancement through LDPC Channel Coding

In order to evaluate cross-layer implementations and incorporate intelligence into the swarm architecture, we have deployed a deep learning-aided LDPC decoding model into the receivers for providing error correction capability, which can further increase the data rate performance. Since the trained neural network had utilized the 3GPP-defined BG2 parity check matrix, every 10 data bits would be coded with additional 42 codebits, resulting in a 52-bit codeword. By utilizing a lifting factor of $Z = 16$, we deployed an LDPC encoder which takes in data vectors of size 160 bits and generates coded frames of size 832 bits. These chunks of

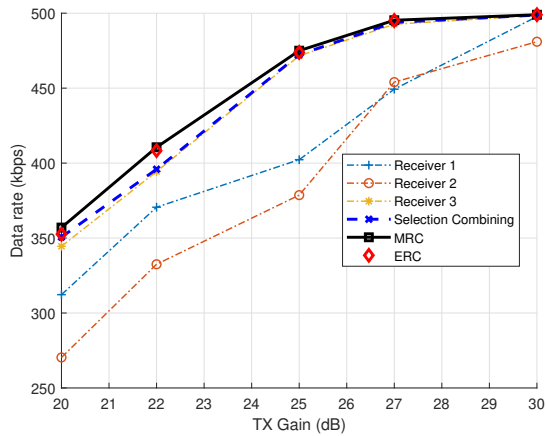


Fig. 5. Data rates of individual routers compared with different combining schemes. MRC is shown to be superior over all combining methods

coded frames are further processed and sent out the RF frontend.

At the receivers, after data capture and postprocessing such as synchronization, demodulation and combining, the obtained soft inputs are provided to the neural LDPC decoder model for inferring the data vector to the best ability. Fig. 6 shows that the swarm-MRC output from the physical layer has either improved or retained the performance, at most places in the operating region, by incorporating LDPC channel coding. It is also noteworthy to observe that the neural network had been trained purely on simulated data, but owing to its large hidden layer architecture, the model performs well even on real OTA data. The data rates can be improved further by incorporating the real-time OTA data into training the neural MS decoder.

V. CONCLUSIONS

In this paper, we have developed the SPELS testbed to successfully evaluate a LEO swarm satellite communication architecture. We have proved the effectiveness of a swarm structure, through the said testbed, by implementing an end-to-end communication module consisting of a receiver diversity-based swarm-MRC technique and a deep learning-aided LDPC decoding algorithm. The testbed was developed to scale well according to the evaluation needs. For instance, owing to the support of 2 RF channels in B210, a second transmission antenna can be attached to the transmitter for evaluating MIMO communication in the uplink. For applications needing higher number of antennas and phase synchronous operations, more sophisticated software defined radios such as USRP X310 and N3xx series, for example, can replace existing devices in a plug-and-play fashion, with minimal changes in the radio configuration setup.

We intend to extend our existing testbed to include MIMO applications ($N_T > 1$), specifically with an aim to develop a jamming-resistant ground-to-space communication. Spatial diversity techniques such as singular value decomposition (SVD)-nulling and also frequency hopping spread spectrum (FHSS)-based anti-jamming functionality will be explored in

the future works. Owing to a high-performance host PC, there is an also an opportunity to incorporate intelligence into several aspects of the communication scheme. While we have demonstrated this notion by introducing a model-driven channel coding scheme, using a pre-trained deep learning model using simulated datasets, the performance of such a setup can be improved even further by incorporating the real time OTA data into the training. Apache Spark, for example, can be utilized for distributing processing of the large repositories of OTA data efficiently within the swarm, and in conjunction with well established machine learning frameworks such as TensorFlow [24], can be used for re-training and updating the existing deployed models for improved performance. More relevant use cases for possible exploration can constitute applications ranging from real-time channel estimation [25] to building capability for serverless edge computing as postulated in [11].

REFERENCES

- [1] M. Giordani and M. Zorzi, "Non-terrestrial networks in the 6g era: challenges and opportunities," *IEEE Network*, vol. 35, no. 2, pp. 244–251, 2021.
- [2] M. M. Azari, S. Solanki, S. Chatzinotas, O. Kodheli, H. Sallouha, A. Colpaert, J. F. Mendoza Montoya, S. Pollin, A. Haqiqatnejad, A. Mostaani, E. Lagunas, and B. Ottersten, "Evolution of non-terrestrial networks from 5g to 6g: a survey," *IEEE Communications Surveys & Tutorials*, vol. 24, no. 4, pp. 2633–2672, 2022.
- [3] X. Lin, S. Rommer, S. Euler, E. A. Yavuz, and R. S. Karlsson, "5g from space: an overview of 3gpp non-terrestrial networks," *IEEE Communications Standards Magazine*, vol. 5, no. 4, pp. 147–153, 2021.
- [4] H. Al-Hraishawi, H. Chougrani, S. Kisseleff, E. Lagunas, and S. Chatzinotas, "A survey on non-geostationary satellite systems: the communication perspective," *IEEE Communications Surveys & Tutorials*, pp. 1–1, 2022.
- [5] *Technical specification group radio access network: solutions for nr to support non-terrestrial networks (nntn) (release 16)*, 3rd Generation Partnership Project Std. 3GPP TR 38.821 V16.1.0, 2021.
- [6] Z. Qu, G. Zhang, H. Cao, and J. Xie, "Leo satellite constellation for internet of things," *IEEE Access*, vol. 5, pp. 18 391–18 401, 2017.
- [7] C. Li, Y. Zhang, R. Xie, X. Hao, and T. Huang, "Integrating edge computing into low earth orbit satellite networks: architecture and prototype," *IEEE Access*, vol. 9, pp. 39 126–39 137, 2021.
- [8] M. Röper, B. Matthiesen, D. Wübben, P. Popovski, and A. Dekorsy, "Beamspace mimo for satellite swarms," in *2022 IEEE Wireless Communications and Networking Conference (WCNC)*, 2022, pp. 1307–1312.
- [9] O. Kodheli, E. Lagunas, N. Maturo, S. K. Sharma, B. Shankar, J. F. M. Montoya, J. C. M. Duncan, D. Spano, S. Chatzinotas, S. Kisseleff, J. Querol, L. Lei, T. X. Vu, and G. Goussetis, "Satellite communications in the new space era: a survey and future challenges," *IEEE Communications Surveys & Tutorials*, vol. 23, no. 1, pp. 70–109, 2021.
- [10] C. A. Hofmann, R. T. Schwarz, and A. Knopp, "Sotm measurements for the characterization of the wideband mobile satellite channel at ku-band," in *SCC 2015; 10th International ITG Conference on Systems, Communications and Coding*, 2015, pp. 1–8.
- [11] L. Shih-Chun, L. Chia-Hung, C. L. C., and L. Shao-Yu, "Towards resilient access equality for 6g serverless p-leo satellite networks," 2022. [Online]. Available: <https://arxiv.org/abs/2205.08430>
- [12] *Technical specification group radio access network; study on new radio (nr) to support non-terrestrial networks (release 15)*, 3rd Generation Partnership Project Std. 3GPP TR 38.811 V15.4.0, 2020.
- [13] R. W. Heath Jr., *Introduction to wireless digital communication: a signal processing perspective*, 1st ed. Pearson Education, 2017.
- [14] J. Ma, C. Peng, S.-C. Lin, and T. Qiu, "Asynchronous blind modulation classification in the presence of non-gaussian noise," in *2019 IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, 2019, pp. 1–10.

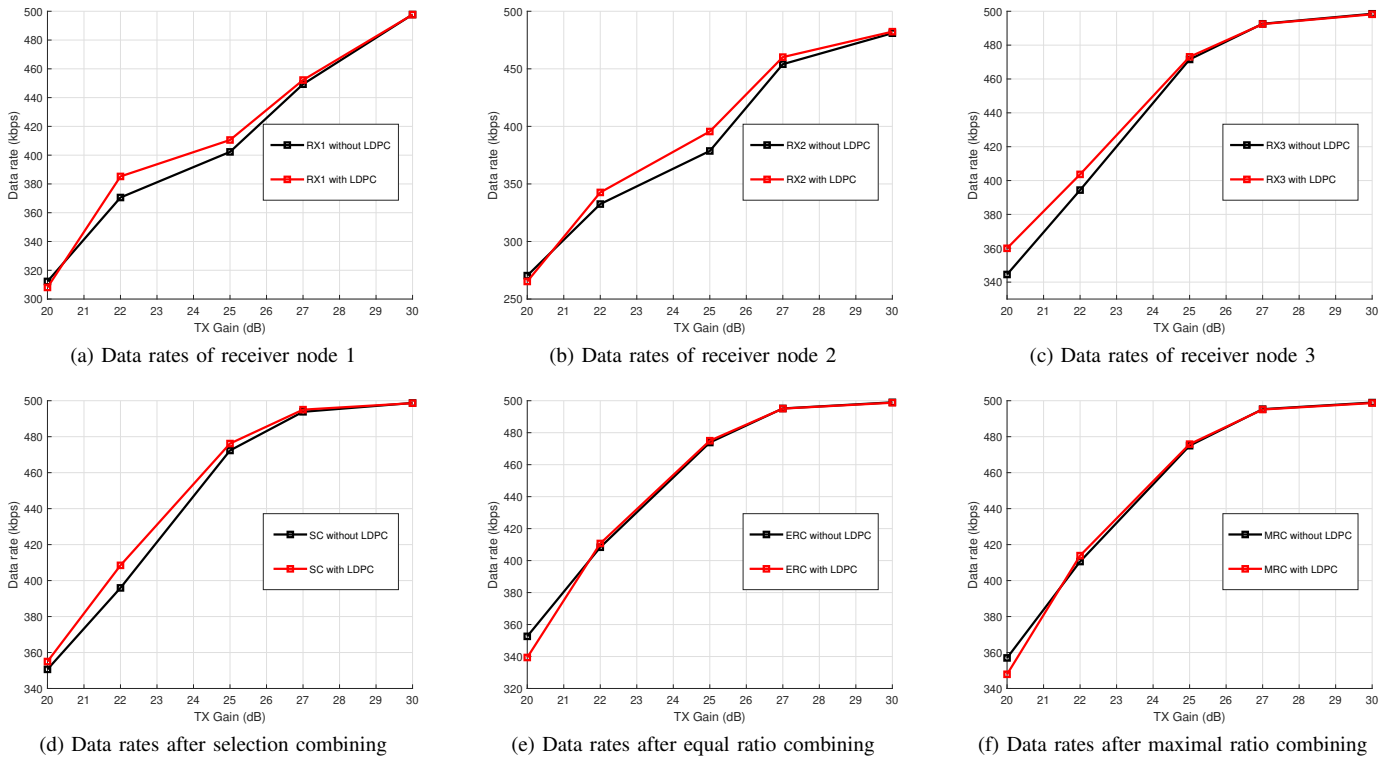


Fig. 6. Comparison of data rates for individual satellite nodes and combining schemes with and without LDPC channel coding enabled

- [15] L. J. Ippolito Jr., *Satellite communications system engineering: atmospheric effects, satellite link design and system performance*. John Wiley and Sons, Ltd, 2008.
- [16] "Usrp hardware driver and usrp manual: device synchronization," 2022. [Online]. Available: https://files.ettus.com/manual/page_sync.html
- [17] Q. Chaudari, *Wireless communications from the ground up: an sdr perspective*, 2018, ch. 8, sec. 2.
- [18] T. David and P. Viswanath, *Fundamentals of wireless communication*. Cambridge University Press, 2005.
- [19] J. Dai, K. Tan, Z. Si, K. Niu, M. Chen, H. V. Poor, and S. Cui, "Learning to decode protograph ldpc codes," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 7, pp. 1983–1999, 2021.
- [20] *Technical specification group radio access network nr; multiplexing and channel coding (release 16)*, 3rd Generation Partnership Project Std. 3GPP TS 38.212 V16.6.0, 2021.
- [21] S. Nie and I. F. Akyildiz, "Channel modeling and analysis of inter-small-satellite links in terahertz band space networks," *IEEE Transactions on Communications*, vol. 69, no. 12, pp. 8585–8599, 2021.
- [22] A. U. Chaudhry and H. Yanikomeroglu, "Laser intersatellite links in a starlink constellation: a classification and analysis," *IEEE Vehicular Technology Magazine*, vol. 16, no. 2, pp. 48–56, 2021.
- [23] "Deep space optical communications (dsoc)," 2022. [Online]. Available: https://www.nasa.gov/mission_pages/tm/dsoc/index.html
- [24] "Open sourcing tensorflowspark: distributed deep learning on big-data clusters," 2017. [Online]. Available: <https://developer.yahoo.com/blogs/157196317141>
- [25] Y. Zhang, Y. Wu, A. Liu, X. Xia, T. Pan, and X. Liu, "Deep learning-based channel prediction for leo satellite massive mimo communication system," *IEEE Wireless Communications Letters*, vol. 10, no. 8, pp. 1835–1839, 2021.

On the resource allocation for radio access network slicing in cellular IoT with massive traffic

Daniel Haro-Mendoza

¹*Universidad Nacional de Chimborazo*
Ecuador

²*Universidad Nacional de La Plata*
Argentina

Luis Tello-Oquendo

¹*Universidad Nacional de Chimborazo*
Ecuador

²*North Carolina State University*
United States

Vicent Pla

Communications Department
Universitat Politècnica de València
Spain

Jorge Martinez-Bauset

Communications Department
Universitat Politècnica de València
Spain

Luis A. Marrone

LINTI
Universidad Nacional de La Plata
Argentina

Shih-Chun Lin

iWN Lab, Dept. of Electrical and Computer Engr.
North Carolina State University
United States

Abstract—The limited capacity of the random access channel (RACH) represents a challenge for adequate resource allocation in 5G radio access networks with network slicing. Furthermore, a fair division of scarce radio resources is required to simultaneously support many users with heterogeneous service requirements. In this work, we look at the problem of uplink radio resource allocation to slices on the radio interface of one cell in a non-stationary regime with mMTC, eMBB, and H2H traffic. We analyze four resource allocation policies for efficient random access to improve each slice’s capacity in terms of successful access probability, the number of preamble transmissions, and access delay. Besides the number of available preambles in the RACH, we also consider the limitation of uplink grants in the radio access network.

Index Terms—cellular systems; machine-type communications; RAN slicing; resource allocation; performance analysis.

I. INTRODUCTION

Unrestricted access to information and services will soon be possible because of a vast number of linked gadgets. Most of these devices, collectively referred to as user equipments (UEs), send data sparsely over time using Internet of Things (IoT) applications. Cellular networks are the greatest option for UE interconnection because of their well-developed infrastructure.

In addition to building on the success of the 4G cellular network, the fifth-generation (5G) wireless technology is anticipated to enable a wide range of network services with various performance needs. One of the foundational technologies for 5G is the Network Slicing (NS) paradigm [1]. It can be viewed as a specially designed logical network made up of virtualized and dedicated resources used to meet the needs of a specific service [2]. It allows serving users from various verticals on the same physical infrastructure. Heterogeneous traffic types, their combined requirements and interactions, and NS in the Radio Access Network (RAN) are being studied from several angles [3], [4]. One of the most important issues to address is resource allocation, and as a result, several proposals are emerging.

Three macro classes have been established to categorize 5G services with distinct traffic patterns and needs: i) *enhanced mobile broadband* (eMBB), which comprises traffic mostly produced by multimedia services; it was common in previous generations, ii) *ultra-reliable and low-latency communications* (URLLC) that must adhere to strict latency and reliability standards, and iii) *massive machine-type communications* (mMTC or mIoT, indistinctly) the most capacity-intensive type of communication.

In this paper, we look at the problem of uplink (UL) radio resource allocation to slices on the radio interface of one cell in a non-stationary scenario of transient mMTC initial access. For this, we focus mainly on the coexistence of H2H, mMTC, and eMBB slices that use two uplink resources, namely preambles and UL grants, during the random access (RA) procedure. Regarding the URLLC slice, we assume it can use only dedicated resources (preambles) that are pre-allocated and fixed in time due to the stringent requirements of such applications. For the evaluation, we obtain the key performance indicators (KPI) defined by the 3GPP [5], namely, access success probability, number of preamble transmissions per access attempt, and access delay.

The rest of the paper is organized as follows. We review studies analyzing NS in Section II. Then, we describe the system model, RAN slicing policies, and the network configuration parameters used in this study in Section III, Section IV, and Section V, respectively. Our most relevant results are presented in Section VI, and finally, we present our conclusions in Section VII.

II. RELATED WORK

Although several papers have focused on resource management and orchestration in 5G networks implementing NS, only a few have addressed resource allocation strategies at the RAN level, particularly in the random access channel (RACH). A significant problem is the coexistence of eMBB, mMTC, and URLLC services and applications in a 5G slice at the RAN

level. While there are already several pieces of research on the performance evaluation of 5G downlink (DL) use cases, there are few results on the UL [6].

In the RAN, the slicing is usually performed using orthogonal resource allocation. In [7], the performance of non-orthogonal slicing of RAN resources in the UL is investigated. The resources are shared by a set of service devices: eMMB, URLLC, and mMTC, with different reliability requirements. The RA procedure for resource allocation is not considered in the study. In an infrastructure with equivalent QoS requirements and different slicing configurations, it is concluded that, with non-orthogonal slicing, the UL presents a higher degradation than the DL in the RA process [8].

In [9] the authors propose prioritizing access to RACH through a segmentation of the preambles available in the system. It consists of a fixed separation of the preambles available for the RA procedure. For this, the preambles are divided into subsets. For example, the authors in [10] propose dividing the preambles into subsets to serve the HTC and MTC services in LTE. In these studies, the preamble allocation remains static regardless of the system load.

In [11], a preamble allocation model is presented based on the estimation of the system load and the priority given to each service class. Three classes of service, URLLC, eMBB, and mMTC, are considered. The load is estimated before each random access opportunity (RAO). Based on the arrival load estimate, the number of preambles allocated for each device class is updated before each RAO.

In [12], the RACH resource allocation in a 5G network implementing NS is studied. Two types of generic 5G services are considered: eMBB and mMTC. Each service can receive dedicated and shared subsets of RAN and RACH resources. The proposed model analyzes the system performance in terms of blocking probability for each slice. It also compares an equal and proportional allocation of resources. An allocation of dedicated and shared preambles is performed. The evaluation is performed only for a network with two slices and includes neither the RA procedure nor the segmentation of UL grants.

The limited capacity of the RACH represents a challenge for adequate resource allocation. Furthermore, a fair division of scarce radio resources is required to simultaneously support many users with heterogeneous service requirements. This work seeks an efficient RA resource allocation policy considering preambles and UL grants.

III. SYSTEM MODEL

A RAN with a set of $S = \{1 \dots S\}$ slices is considered. We concentrate on a cell-level resource allocation issue and study the allocation of UL resources used in the RA procedure. UEs are fully informed of the slice to which they belong. The base station (gNB in 5G) broadcasts system information about the access process and slice configuration. A slice policy (described in Section IV) determines how radio resources are distributed.

The RA can operate in two modes: contention-free and contention-based. The former is used for critical situations

such as handover or positioning. The latter is the standard mode for network access; it is used by UEs to change the RRC state from idle to connected, to recover from radio link failure, to perform UL synchronization, or to send scheduling requests [13].

Random access attempts are allowed in predefined time/frequency resources, called RAOs. The gNB broadcasts the periodicity of the RAOs using a variable referred to as the PRACH Configuration Index. The periodicity varies between a minimum of 1 RAO every two frames (i.e., 1 RAO every 20 ms) and a maximum of 1 RAO per 1 sub-frame (i.e., every 1 ms) [14].

The physical RACH (PRACH) signals a connection request when a UE needs to access the RAN. It carries a preamble for initial access to the network. Up to $R = 64$ orthogonal preambles are available to the UEs per cell [14]. In contention-free mode, there is a coordinated assignment of preambles, so collision is avoided, but gNBs can only assign these preambles during specific slots to specific UEs. Hence, UEs can only use these preambles if assigned by the gNB and during specific slots. In the contention-based mode, preambles are selected randomly by the UEs, so there is a risk of collision; that is, there is a probability that multiple UEs in the cell pick the same preamble; therefore, contention resolution is needed. In the sequel, we focus on the contention-based random access mode.

A. Contention-based Random Access Procedure

A UE initiates its access attempt by sending *Msg1* to the gNB. *Msg1* contains a preamble randomly chosen by the UE from a set of preambles. Due to preamble orthogonality, several UEs can access the gNB in the same RAO using different preambles. However, if two or more UEs transmit the same preamble, the transmitted preamble cannot be decoded by the gNB, i.e., an *Msg1* transmission collision occurs [15]. If *Msg1* has sufficient transmission power, it will be decoded by the gNB [15]–[17]. If it is not decoded, the UE will make a new attempt by increasing the transmission power.

The gNB responds with an *Msg2* to each successfully decoded *Msg1*. The *Msg2* includes identification information for the detected preamble and the granting of reserved resources (UL Grant) for the *Msg3* transmission [15], [17]. The UEs that do not receive the *Msg3* within the W_{RAR} time window will raise their power and perform retransmission by randomly choosing a new preamble. All UEs that receive an UL grant through *Msg2* will be able to transmit *Msg3*. The transmission of *Msg3* is guaranteed through the hybrid automatic repeat request (HARQ) [15], [17].

The gNB transmits *Msg4* in response to *Msg3*. *Msg4* also uses the HARQ process. If the UE does not receive *Msg4* within the contention resolution time, the connection is declared failed, and a new access attempt is planned by increasing the transmission power. If a UE reaches the limit of unsuccessful re-transmissions, the network is declared unreachable, terminating the RA procedure [15]. UEs that complete the RA procedure receive a block of time-frequency

resources for communication. All UEs that fail their transmission must execute a backoff procedure, regardless of the reason for the failure or the slice to which they belong. In this procedure, the UE waits for a random time $\mathcal{U}(0, BI)$ ms before starting a new preamble transmission in a new RAO. BI is the backoff indicator, defined by the gNB and sent to the UEs in the $Msg2$ [17], [18].

IV. RAN SLICING POLICIES

5G networks implementing NS require defining the allocation of RAN resources among the different slices. We analyze the allocation of the preamble and UL grants between the gNB and UEs statically and adaptively. In both cases, we consider i) a full isolation level between slices (Fully-sliced) in which preambles and UL grants are reserved for each slice; and ii) a medium isolation level between slices (Partially-sliced) in which the UL grants are not reserved but shared by all slices.

1) *Fully-sliced Static Policy*: Since the number of preambles assigned to each slice has a high impact on the probability of collision [19], in this proposal, we assume a fixed allocation in which the number of allocated preambles and UL grants are proportional to each other. This number is determined by the priority of the service using the slice. A cell with S slices is considered; the gNB performs a fixed allocation of subsets of different preambles and UL grants to each slice. Doing so allows additional QoS requirements to be handled with isolation between slices.

We consider three services: mIoT, eMBB, and H2H. Each service accesses a slice with different priorities (high, medium, low). For example, the mIoT service serves a hefty load of access requests from applications with machine-type devices, requiring a high-priority slice. On the other hand, eMBB requires a medium priority slice to serve a moderate number of access requests with high bandwidth requirements [9]. Finally, H2H traffic in which few accesses (compared to expected mIoT [15]) requires a low-priority slice.

To calculate the number of preambles assigned to each slice, we define a weight $\{w_i | \sum_{i=1}^S w_i = 1\}$ for high, medium, and low priority slices, respectively. Thus, the slice s (mIoT, eMBB, or H2H) receives a percentage of the total number of preambles available in the system calculated as

$$r_i = \begin{cases} \lceil R * w_i \rceil, & i = 1, \dots, S-1 \\ R - \sum_{j=1}^{S-1} r_j, & i = S. \end{cases} \quad (1)$$

In addition, to ensure the isolation of each slice, an allocation of the available UL grants θ is performed by

$$g_i = \begin{cases} \lceil \theta * w_i \rceil, & i = 1, \dots, S-1 \\ \theta - \sum_{j=1}^{S-1} g_j, & i = S. \end{cases} \quad (2)$$

2) *Fully-sliced Adaptive Policy*: The probability of successful access to a slice depends on the number of devices accessing and competing for system resources. Therefore, a static preamble allocation policy based on priorities alone will not be efficient. Ideally, it should be combined with the number of active requests in the RACH at each RAO [20].

Unfortunately, the number of active requests in the RACH is time-varying, composed of requests for new accesses and those requests that collided and are attempting again. Therefore, we need an algorithm that considers the number of active devices at each RAO to assign preambles to each slice.

We consider a slice with dedicated preambles for each type of traffic mIoT, eMBB, and H2H. In addition, we reserve a set of preambles shared by traffic flows of the dedicated slices that pass the conditions explained below. As indicated in Eq. (3), out of a total of R preambles available in the system, r_i preambles are reserved for the i th slice, and all slices share r_s preambles.

$$R = r_s + \sum_{i=1}^S r_i. \quad (3)$$

A higher number of collisions occur when a slice does not have enough preambles allocated. In addition, the gNB has a limited number of UL grants θ to respond to successfully detected preambles. Therefore, when the number of preambles detected by the gNB in a RAO is greater than θ there will be preambles that do not receive a UL grant. UEs that do not receive a UL grant should perform a new access attempt [19].

We then propose using the r_s subset as an alternative way to serve accesses with a high probability of failure if they use the preambles dedicated to their slice. This way, we mainly prevent these accesses from causing a collision and affecting other UEs. Access attempts using r_s contend for preambles other than those assigned to their slice. In this work, since we are considering collision detection in $Msg1$, having UL grants reserved for the shared preambles is unnecessary. Only detected and non-collided accesses using the r_s subset will require UL grants reserved to their slice.

To determine the percentage of shared preambles and UL grants assigned to each slice, we use the coefficient δ in Eq. (4). In high-traffic scenarios, the higher the level of sharing, the higher the collision probability is [12], [20].

$$r_s = \lceil \delta * R \rceil. \quad (4)$$

We calculate the subset of preambles assigned to each slice from the remaining preambles. The initial configuration of the proposal considers that the gNB will reserve some dedicated preambles for each slice equally; this number is calculated as

$$r_0 = \frac{R - r_s}{S} = (1 - \delta)R/S. \quad (5)$$

The number of preambles and UL grants assigned to each slice will be dynamically updated by the gNB using the $SIB2$ message, which allows the gNB to transmit the configuration parameters to the UEs with a periodicity of 80 ms = 16 RAOs [17]. The number of preambles and UL grants assigned in each period is calculated based on the number of active devices in the i th slice N_i per RAO and is obtained as follows

$$r_i = \frac{\overline{N}_i}{-\ln(w_i) - \ln(x)}, \quad (6)$$

where w_i is the weight assigned to the i th slice, \overline{N}_i is the average number of active devices in the i th slice per *SIB2* update period, and x represents a proportionality factor in ensuring that the available resources (preambles or UL grants) of the RACH are not exceeded; it is tuned to satisfy

$$R = r_s + \sum_{i=1}^S \frac{\overline{N}_i}{-\ln(w_i) - \ln(x)}. \quad (7)$$

The number of active devices for the i th slice N_i that access the RACH and wait for the preamble assignment varies in each RAO. Moreover, the gNB has no way of knowing this information; this value is estimated using the process reported in our previous works [17], [18], [21].

Bearing in mind that the maximum number of successful attempts is obtained when the number of contending UEs at the i th slice is approximately the number of preambles assigned to that slice (i.e., $r_s \approx N_i$) [15], we define thresholds for each service traffic. Requests from active nodes that exceed the corresponding threshold will use the r_s subset of preambles. In this way, we ensure each slice's maximum capacity, avoiding excess of collisions and retransmissions. Requests using r_s will attempt to complete the RA procedure with a lower successful access probability. Those UEs attempts that do not collide in the transmission of *Msg1* and are correctly detected by the gNB will wait for a UL grant to finish the procedure successfully. In contrast, the UEs attempts that used the r_s and failed will be able to make their next attempt once the backoff time has elapsed.

3) *Partially-sliced scheme for Static and Adaptive Policies:* We also analyzed a variation to the fully-sliced scheme in both Static and Adaptive policies where UL grants are not reserved for each slice. Instead, the UL grants are shared and available to access attempts that complete the *Msg1* and are correctly detected by the gNB regardless of the subset of slice preambles they used. It is evident that the access attempts will constantly utilize all UL grants in high traffic. A disadvantage of this variation is the partial loss of isolation using slice resources. That is, this scheme isolates preambles but not UL grants.

V. NETWORK CONFIGURATION PARAMETERS

A discrete-event simulator of the 5G RAN with NS has been developed in C++ to evaluate the proposals. Additionally, these results were corroborated with MATLAB simulations independently. The system accommodates three types of traffic in each simulation: mMTC, eMBB, and H2H, with different access request intensities. The distribution and parameters used by each traffic model are described in Table I. The contention-based RA procedure described in Section III-A is replicated with the parameters listed in Table II. Simulations were run j times until the difference of computing the corresponding metric in the j th simulation run differs from the one computed in the $j - 1$ th simulation run by less than 1%, considering a minimum value for j such as 10^3 . The simulator provides the flexibility of choosing the parameters of interest, including the type of traffic, number of devices, timing, processing and

Table I
TRAFFIC MODELS FOR 5G NS RACH EVALUATION

| Characteristics | Traffic Model mMTC | Traffic Model eMBB | Traffic Model H2H |
|-------------------------|---------------------|--------------------|-------------------|
| Arrival distribution | Beta(3,4) over T | Poisson(5) over T | Uniform over T |
| Number of devices | 2500, . . . , 30000 | 1000 | 33000 |
| Distribution period (T) | 10 seconds | 10 seconds | 60 seconds |

Table II
GENERAL RACH SLICING CONFIGURATION

| Parameter | Setting |
|--|---------------------------|
| Number of slices | 3 |
| PRACH Configuration Index | 6 |
| RA Periodicity (RAO) | 5 ms |
| Subframe length | 1 ms |
| Total number of preambles | 54 |
| Maximum number of preamble transmissions | preambleTransMax = 10 |
| RAR window size | $W_{RAR} = 5$ |
| mac-ContentionResolutionTimer | 48 sub-frames |
| Maximum number of UL grants per subframe | $N_{RAR} = 3$ |
| Backoff Indicator | $BI = 20$ ms |
| Preamble detection probability for kth preamble transmission | $P_d = 1 - \frac{1}{e^k}$ |
| HARQ re-transmission probability for Msg3 and Msg4 (non-adaptive HARQ) | 10% |
| Maximum number of HARQ TX for Msg3 and Msg4 (non-adaptive HARQ) | 5 |
| Periodicity of RAOs | 5 ms |
| Preamble transmission time | 1 ms |

channel parameters such as the number of available preambles, number of slices, priorities, and backoff window size.

A. Performance Metrics

The three KPIs for the purpose of RACH capacity evaluation with each slicing policy are the following [5]:

- 1) Access success probability P_s is the probability of successfully completing the random access procedure within the maximum number of preamble transmissions.
- 2) Statistics of the number of preamble transmissions per access attempt K .
- 3) Statistics of access delay D defined as the time elapsed between the arrival of a UE and the successful completion of its RA procedure.

B. Static-sliced Policies

We define the vector $w = [0.64, 0.32, 0.04]$ for the high, medium, and low priority slices, respectively. We find the number of preambles assigned to each slice r_i using Eq. (1). It remains constant throughout the simulation and is reserved for use by the UEs of each slice. With the same logic, we use Eq. (2) for reserving the UL grants of each slice g_i .

C. Adaptive-sliced Policies

To evaluate these policies, prior to the start of the RA procedure, we find the number of shared preambles r_s . For this, we assume a $\delta = 10$ in Eq. (4) since it is the factor that maximizes performance in a high-traffic scenario, as observed

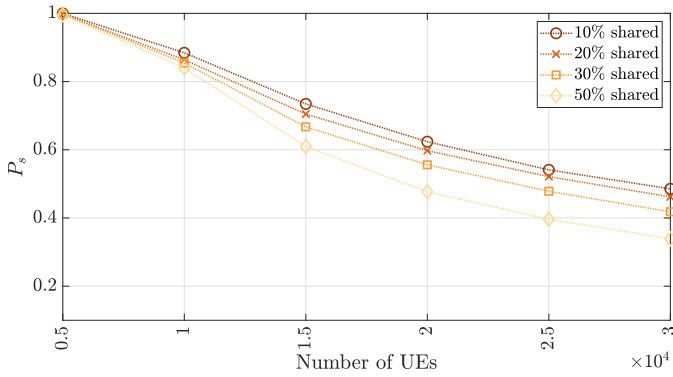


Figure 1. Successful access probability of mIoT traffic for each δ

in Fig. 1. The remaining preambles will be assigned using Eq. (5) to each slice dedicated to mIoT, eMBB, and H2H services.

In the following RAOs, the allocation of preambles and UL grants to each dedicated slice will be performed dynamically using Eqs. (6) and (7) each *SIB2* period. In addition, we define a priority vector $w = [0.57, 0.29, 0.14]$ for mIoT, eMBB, and H2H services.

VI. RESULTS

In the following, we detail the results for each service according to the traffic models detailed in Table I and the network configuration described in Table II. The eMBB and H2H services are evaluated when mIoT traffic varies from light (2500 access requests) to heavy load (30000 access requests). We consider each scenario's eMBB service with medium-load (1000 access requests) and H2H service as background traffic. For the sake of comparison, we also evaluate a scenario without implementing network slicing, called *Un sliced*.

A. mIoT service

Fig. 2 illustrates the P_s as a function of the number of mIoT UEs. As expected, P_s decreases as the number of mIoT UEs increases. The Adaptive-sliced policies maintain a higher value of P_s than the Un sliced and Static-sliced configurations. For light load scenarios (i.e., less than 10000 UEs), all slice policies present a high P_s value; it is evident that as the number of UEs competing for access in the RA procedure increases, the P_s drops drastically. Moreover, it is observed that the fully-sliced adaptive performance is very close to that of the partially-sliced adaptive, where the UL grants are not reserved but available for any service. The advantage of these policies is that the isolation level is improved (i.e., any flow changes in a slice can affect the performance of the remaining slices in a lesser way) since resource allocations are made dynamically with the evolution of active accesses. Fig. 3 depicts the average number of preamble transmissions required for successful access. Un sliced and static-sliced policies require a higher K than the adaptive-sliced ones. Finally, Fig. 4 illustrates the 95th percentile of the access delay D_{95} . We observe that it increases with the number of UEs in all

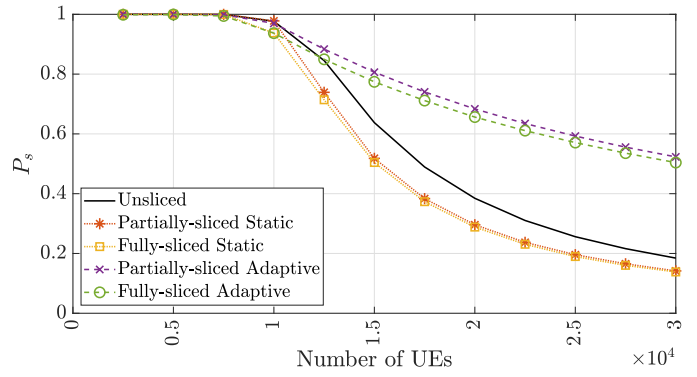


Figure 2. mIoT slice. Successful access probability P_s

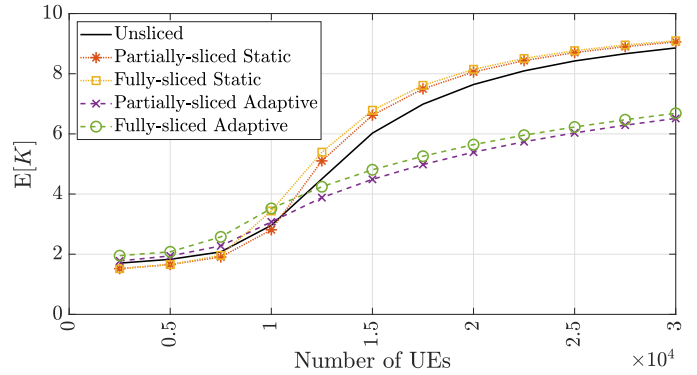


Figure 3. mIoT slice. Average number of preamble transmissions required for successful access $E[K]$

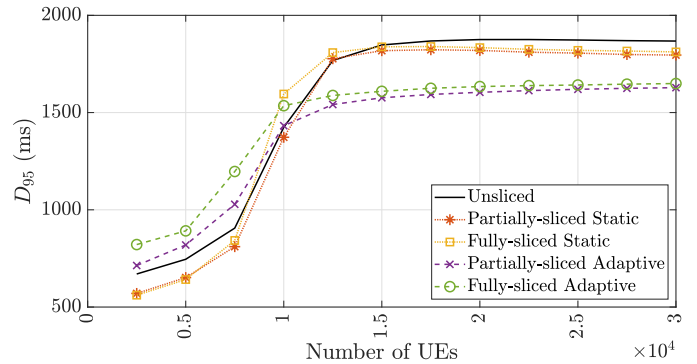


Figure 4. mIoT slice. 95th percentile of access delay D_{95}

cases. The Adaptive-sliced policies achieve a smaller D_{95} in heavy load conditions.

B. eMBB service

Fig. 5 illustrates the behavior of P_s . The Partially-sliced static policy performs better for light and heavy traffic conditions than other policies. From 10000 UEs onwards, both static-sliced policies provide higher P_s . Concerning K , Adaptive-sliced and Un sliced policies perform similarly in light load conditions as observed in Fig. 6. The Static-sliced policies perform uniformly for all network load conditions; in particular, the partially-sliced static requires fewer preamble

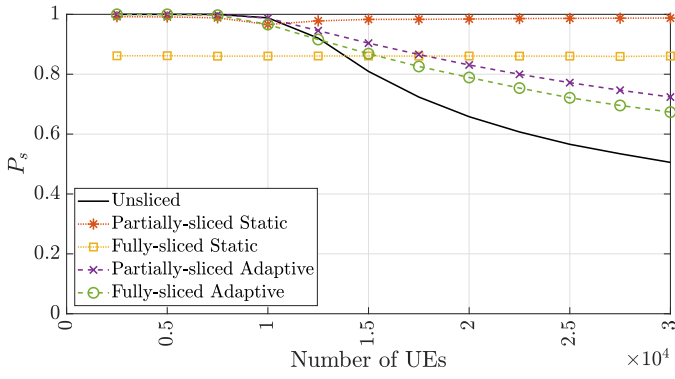


Figure 5. eMBB slice. Successful access probability P_s

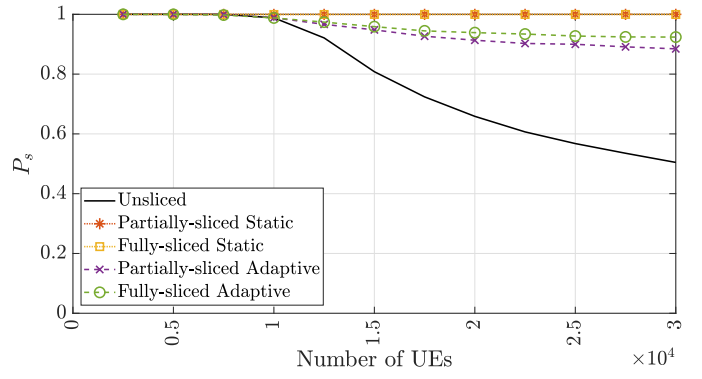


Figure 8. H2H slice. Successful access probability P_s

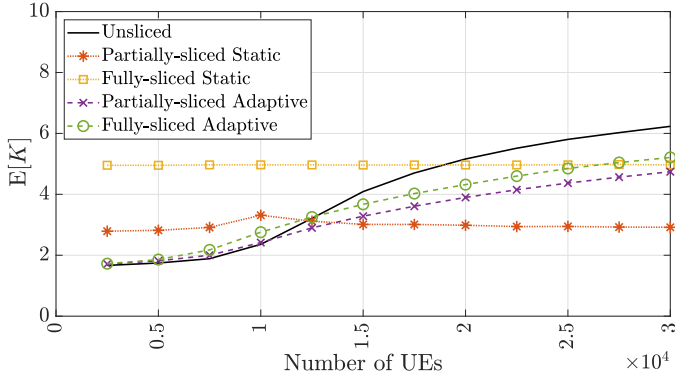


Figure 6. eMBB slice. Average number of preamble transmissions required for successful access $\mathbb{E}[K]$

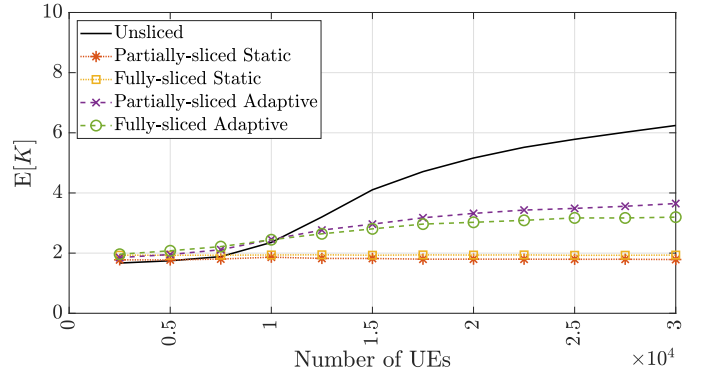


Figure 9. H2H slice. Average number of preamble transmissions required for successful access $\mathbb{E}[K]$

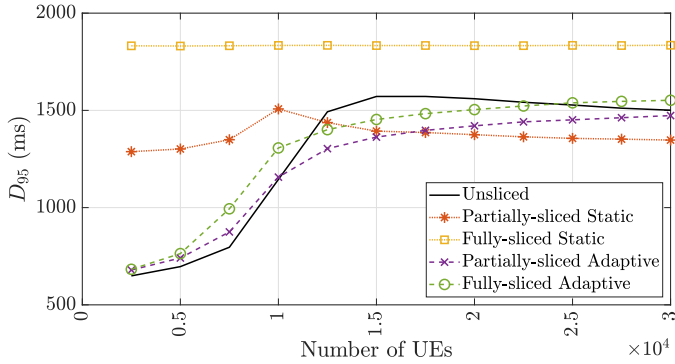


Figure 7. eMBB slice. 95th percentile of access delay D_{95}

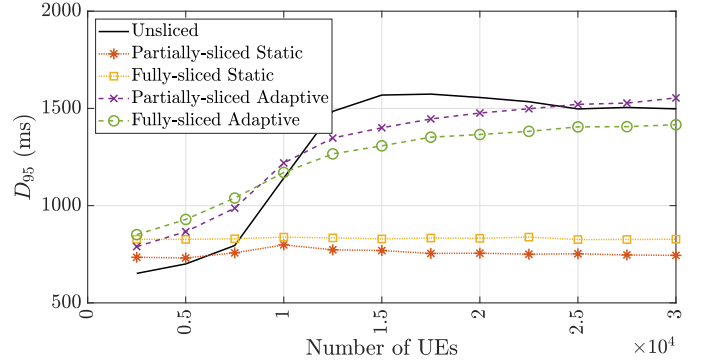


Figure 10. H2H slice. 95th percentile of access delay D_{95}

transmissions in heavy load conditions for successful access. Regarding D , smaller values for light loads (less than 10000 UEs) can be obtained with the adaptive-sliced policies, as observed in Fig. 7. In heavy load scenarios, all policies present a similar behavior, particularly the partially-sliced static policy shows a smaller D .

C. H2H service

Figs. 8, 9, and 10 present the results of the evaluation of the H2H service and the effect it suffers with a variation of the number of mIoT UEs. Fig. 8 indicates that the best performance is obtained with the Static-sliced policies. Both

Adaptive-sliced policies perform considerably better than the Unsliced one. A value of P_s above 90% is guaranteed with all slicing policies. The number of transmitted preambles and the access delay is illustrated in Figs. 9 and 10, respectively. When the number of UEs exceeds 7500, the Static-sliced policies provide the lowest values in the two metrics since they have reserved resources for efficient performance. The H2H slice performance is not affected by increasing the number of mIoT accesses with the Static-sliced policies. Comparing the Adaptive-sliced policies and the Unsliced one, they show similar behavior in terms of D and the Adaptive-sliced outperforms the Unsliced in terms of K .

VII. CONCLUSIONS

Implementing network slicing in 5G radio access networks achieves isolation between different services hosted by different slices. Traffic variations generated in one slice minimally affect the other slices. We verified that limiting the accesses to the maximum capacity of each slice allows for maximizing the utilization of the RACH by increasing the probability of successful access of UEs, isolating each slice from the congestion produced by the different services. A shared preamble subset serves connection requests that exceed capacity.

In the fully-sliced policies, a segmentation of preambles and UL grants is performed, which means that any congestion issue in one slice will not be propagated to the rest. In the partially-sliced policies, a segmentation of preambles and not of UL grants are performed; complete isolation is not reached, but an efficient occupation of the available UL grants is achieved since unused UL grants by slices with a light load can be exploited by access request from other slices.

The partially-sliced static policy can improve the performance of eMBB and H2H slices in heavy-load mIoT scenarios due to a constant allocation of resources. For mIoT services, the adaptive-sliced policies provide better performance.

ACKNOWLEDGMENT

This work was supported in part by Meta 2022 AI4AI Research, the NC State 2022 Faculty Research and Professional Development Program (FRPD), the NC Space Grant, the AC21 Special Project Fund (SPF), and the National Science Foundation (NSF) under Grant CNS-2210344. The work of J. Martinez-Bauset and V. Pla was supported by Grants PGC2018-094151-B-I00 and PID2021-123168NB-I00 funded by MCIN/AEI/10.13039/501100011033 and ERDF *A way of making Europe*.

REFERENCES

- [1] A. Kaloylos, "A survey and an analysis of network slicing in 5G networks," *IEEE Communications Standards Magazine*, vol. 2, no. 1, pp. 60–65, 2018.
- [2] N. Alliance, "Description of network slicing concept," *NGMN 5G P*, vol. 1, no. 1, pp. 1–11, 2016.
- [3] K. Samdanis, S. Wright, A. Banchs, A. Capone, M. Ulema, K. Obana *et al.*, "5G Network Slicing: Part 1—Concepts, Principales, and Architectures [Guest Editorial]," *IEEE Communications Magazine*, vol. 55, no. 5, pp. 70–71, 2017.
- [4] K. Samdanis, S. Wright, A. Banchs, A. Capone, M. Ulema, and K. Obana, "5G network slicing—part 2: Algorithms and practice," *IEEE Communications Magazine*, vol. 55, no. 8, pp. 110–111, 2017.
- [5] 3GPP, *TR 37.868, Study on RAN Improvements for Machine Type Communications*, Sept 2011.
- [6] V. Mancuso, P. Castagno, M. Sereno, and M. A. Marsan, "Serving HTC and Critical MTC in a RAN Slice," in *2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 2021, pp. 189–198.
- [7] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.
- [8] C. Marquez, M. Gramaglia, M. Fiore, A. Banchs, and X. Costa-Perez, "Resource sharing efficiency in network slicing," *IEEE Transactions on Network and Service Management*, vol. 16, no. 3, pp. 909–923, 2019.
- [9] H. Zhang, N. Liu, X. Chu, K. Long, A.-H. Aghvami, and V. C. Leung, "Network slicing based 5G and future mobile networks: mobility, resource management, and challenges," *IEEE communications magazine*, vol. 55, no. 8, pp. 138–145, 2017.

- [10] C. Kalalas, F. Vazquez-Gallego, and J. Alonso-Zarate, "Handling mission-critical communication in smart grid distribution automation services through LTE," in *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, 2016, pp. 399–404.
- [11] H. Althumali, M. Othman, N. K. Noordin, and Z. M. Hanapi, "Priority-based load-adaptive preamble separation random access for QoS-differentiated services in 5G networks," *Journal of Network and Computer Applications*, vol. 203, p. 103396, 2022. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S1084804522000558>
- [12] O. Vikhrova, C. Suraci, A. Tropeano, S. Pizzi, K. Samouylov, and G. Araniti, "Enhanced radio access procedure in sliced 5G networks," in *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. IEEE, 2019, pp. 1–6.
- [13] 3GPP, *TS 36.321, Medium Access Control (MAC) Protocol Specification*, Sept 2012.
- [14] —, *TS 36.211, Physical Channels and Modulation*, Dec. 2014.
- [15] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, 2017.
- [16] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Communications Letters*, vol. 20, no. 2, pp. 372–375, 2015.
- [17] L. Tello-Oquendo, J.-R. Vidal, V. Pla, and L. Guijarro, "Dynamic access class barring parameter tuning in LTE-A networks with massive M2M traffic," in *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, 2018, pp. 1–8.
- [18] J.-R. Vidal, L. Tello-Oquendo, V. Pla, and L. Guijarro, "Performance study and enhancement of access barring for massive machine-type communications," *IEEE Access*, vol. 7, pp. 63 745–63 759, 2019.
- [19] V. Mancuso, P. Castagno, M. Sereno, and M. A. Marsan, "Modeling MTC and HTC radio access in a sliced 5G base station," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 2208–2225, 2020.
- [20] J. Liu, M. Agiwal, M. Qu, and H. Jin, "Online control of preamble groups with priority in massive IoT networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 700–713, 2020.
- [21] L. Tello-Oquendo, V. Pla, I. Leyva-Mayorga, J. Martinez-Bauset, V. Casares-Giner, and L. Guijarro, "Efficient random access channel evaluation and load estimation in lte-a with massive mtc," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1998–2002, 2018.

Modeling the resource allocation in 5G radio access networks with network slicing

Daniel Haro-Mendoza

¹*Universidad Nacional de Chimborazo*
Ecuador

²*Universidad Nacional de La Plata*
Argentina

Luis Tello-Oquendo

¹*Universidad Nacional de Chimborazo*
Ecuador

²*North Carolina State University*
United States

Vicent Pla

Communications Department
Universitat Politècnica de València
Spain

Jorge Martinez-Bauset

Communications Department
Universitat Politècnica de València
Spain

Luis A. Marrone

LINTI
Universidad Nacional de La Plata
Argentina

Shih-Chun Lin

iWN Lab, Dept. of Electrical and Computer Engr.
North Carolina State University
United States

Abstract—Network Slicing (NS) is one of the technologies considered a pillar of 5G networks. It allows the division of the physical infrastructure of a network into several isolated logical networks (slices). The slices can have different sizes and be offered to other use cases. We analyze the radio resource allocation problem through a random access channel model considering the radio access network (RAN) with NS in a steady state. We perform an in-depth study of the random access procedure (RAP) to optimize resource allocation in a 5G RAN with NS. We focus on assigning subsets of preambles for each slice depending on the service’s priority. The main contributions of our work are the following: i) A model for a scenario of n slices; that is, it has no limitation for the number of use cases. ii) An efficient RAP resource allocation policy to maximize the probability of successful access by UEs in each slice.

Index Terms—5G cellular systems; network slicing; analytic model, RAN slicing; resource allocation.

I. INTRODUCTION

Today’s many connected devices allow massive and unrestricted access to information. However, most of these devices, called user equipment (UE), send data sparsely over time, using Internet of Things (IoT) applications. The best interconnection alternative for UEs is cellular networks due to their widely deployed infrastructure. However, cellular technology was conceived to handle human-to-human (H2H) traffic and not many UEs interacting simultaneously, as with machine-to-machine (M2M) communications. This results in many devices trying to connect to the base station of a cellular network with the corresponding congestion problems that this causes.

Fifth-generation (5G) networks emerge as an alternative to satisfy wireless network users’ high service and connectivity requirements. With the implementation of 5G networks, data rates are expected to reach 10 Gbps [1]. It is also estimated that 5G will reach a total of 4.4 billion subscribed devices, which will represent 49% of all mobile subscriptions in 2027 [2]. Besides, the vision of 5G is to provide extremely low latency, higher capacity, and better QoS perceived by users [3].

Unlike 4G, which was conceived to provide mobile broadband communications, the 5G infrastructure is expected to enable the evolution of sectors such as industry 4.0, automotive, e-medicine, and entertainment, among others [4]. Although the vision and benefits of 5G are precise, enabling technologies are

Table I
SLICE TYPES FOR USE CASES

| Slice / Service type | SST value | Characteristics |
|----------------------|-----------|--|
| eMBB | 1 | 5G enhanced mobile broadband |
| URLLC | 2 | Ultra-reliable low latency communications |
| mIoT | 3 | Massive communications IoT |
| V2X | 4 | Vehicle to everything V2X services |
| HMTc | 5 | High-Performance Machine-Type Communications |

an open field of research. One of the technologies considered a pillar of 5G networks is Network Slicing (NS). NS allows the division of the physical infrastructure of a network into several isolated logical networks (slices). The slices can have different features and be offered to other use cases. In [5], a slice is defined as a combination of network functions (NF) and radio access technologies (RAT) for a specific use case. So, NS is allocating a dedicated or shared portion of the network resources for each slice [6].

In the ETSI Technical Specification 123 501 update [7], a slice is identified by a Single Network Slice Selection Assistance Information (S-NSSAI). An S-NSSAI comprises i) a Slice/Service type (SST), which identifies the expected service in the NS, and ii) a differentiator segment that allows distinguishing several NSs belonging to the same type of service. These standardized values in the update allow categorizing five use cases for NS, described in Table I.

In the following, we analyze the problem of radio resource allocation through a random access channel (RACH) model considering the RAN with NS in a steady state. For this, we focus on n traffic flows that, during the Random Access Procedure (RAP), use the uplink resources (preambles and uplink grants). For evaluation purposes, we compute the two Key Performance Indicators (KPIs) defined by the 3GPP [8]: the probability of successful access and the number of preamble transmissions per access attempt.

A. Random access procedure with NS

All UEs needing resources to access service must execute the RAP. It starts when the base station (gNB) offers a random access opportunity (RAO) to the UEs [6]. The RAP execution

uses two physical channels: the PRACH for the transmission of preambles and the PUSCH for the data [9]. A preamble is a specific identifier that UEs transmit to indicate their presence in the cell to the gNB. The preamble signals are orthogonal (i.e., the gNB can distinguish preambles sent simultaneously by multiple UEs). However, the number of preambles in a 5G New Radio (NR) cell is limited to 64. UEs randomly select one of these preambles to start their network access [10].

In the time domain, the access system is divided into slots. Each slot represents a RAO; it occurs periodically, and the *prach-ConfigIndex* parameter determines its periodicity [11]. We consider a subframe length of 1 ms and a RAO periodicity of 5 ms, corresponding to the setting *prach-ConfigIndex* = 6.

The RAP can be performed in two ways: i) contention-free or ii) contention-based. The former allocates reserved preambles during specific intervals, and for specific UEs (collision-free) [9]. In the latter, the UEs choose preambles randomly; two or more UEs in the same cell could choose the same preamble for the same RAO, causing a collision. A high number of collisions will cause a low probability of success and an increased access delay. The 3GPP standard suggests using 54 preambles for contention-based RAP [12].

Before an access attempt, the gNB shares network parameters with UEs through Master Information Block (MIB) and System Information Blocks (SIB) [11] messages. Among the parameters received through the SIB Type 2 is the periodicity in time of the RAOs [13].

1) *Contention-based RAP*: Its operation is based on executing the four-message handshake between the gNB and the UEs. A UE initiates its access attempt by sending *Msg1* to the gNB. *Msg1* contains a preamble randomly chosen by the UE from a set of preambles. Due to preamble orthogonality, several UEs can access the gNB in the same RAO using different preambles. However, if two or more UEs transmit the same preamble, the transmitted preamble cannot be decoded by the gNB, i.e., an *Msg1* transmission collision occurs [13]. If *Msg1* has sufficient transmission power, it will be decoded by the gNB [9], [13], [14]. If it is not decoded, the UE will make a new attempt by increasing the transmission power.

The gNB responds with an *Msg2* to each successfully decoded *Msg1*. The *Msg2* includes identification information for the detected preamble, and the granting of reserved resources (UL grant) for the *Msg3* transmission [9], [13]. The UEs that do not receive the *Msg3* within the W_{RAR} time window will raise their power and perform retransmission by randomly choosing a new preamble. All UEs that receive an UL grant through *Msg2* will be able to transmit *Msg3*. The transmission of *Msg3* is guaranteed through the hybrid automatic repeat request (HARQ) [9], [13].

The gNB transmits *Msg4* in response to *Msg3*. *Msg4* also uses a HARQ scheme. If the UE does not receive *Msg4* within the contention resolution time, the attempt is declared failed, and a new access attempt is planned, and the transmission power is increased. If a UE reaches the maximum number of re-transmissions, the network is declared unreachable, terminating the RA procedure [13]. UEs that complete the RA procedure receive a block of time-frequency resources for communication. All UEs that fail their transmission must execute a backoff procedure, regardless of the reason for the failure or the slice to which they belong. In this procedure, the UE waits for a random time $\mathcal{U}(0, BI)$ before starting a new

preamble transmission. *BI* is the backoff indicator, defined by the gNB and sent to the UEs in *Msg2* [9], [15].

The rest of the paper is organized as follows. We conduct a literature review regarding NS in Section II. Then, we describe the system and analytical models in Section III and Section IV, respectively. Our most relevant results are presented in Section V, and finally, the conclusions are presented in Section VI.

II. RELATED WORK

Most studies have focused on the management and orchestration of resources instead of how to allocate these resources in the 5G radio access network. The limited number of preambles and UL grants available in the RACH represents a resource allocation problem. Another significant issue is the coexistence of eMBB, mMTC, and URLLC services and applications in a 5G segment at the RAN level. While there is much research on performance evaluation of 5G downlink (DL) use cases, there are few results for UL [16].

In [17], an algorithm for optimizing the allocation of radio resources to the slices of the cell of a network that implements NS is proposed. Its performance is evaluated by simulation. The study compares different priority levels assigned to each slice. The priority of each slice is defined through four techniques: i) searches for the order that meets an objective function; ii) performs a random ordering; iii) performs an ordering to maximize the assigned resources; and iv) a prioritization based on the granularity of each slice. Three resource allocation methods that ensure isolation in a RAN with NS are presented in [18]. Notably, a proportional fairness algorithm limits the number of RBs assigned to each slice. The authors show through simulation that the isolation between slices is guaranteed. The results report an improvement in system performance for the three methods: static allocation, allocation to ordered slices, and impartial allocation to slices. In these investigations, the problem of allocating resources and allocating procedures in the RACH access is not considered.

An optimization approach for allocating radio resources in the 5G RAN that implements NS is addressed in [19]. Two types of generic 5G services are considered: eMBB and mMTC. Each service can receive dedicated and shared subsets of RAN and RACH resources. The proposed model analyzes the system's performance in terms of blocking probability for each slice without analyzing the access delay nor the number of retransmissions for successful access. Their model considers that collisions occur in *Msg3* of the RAP and evaluate an equal and proportional allocation of resources for two slices. An optimal resource segmentation alternative based on the number of slices in the system is not presented. This proposal does not achieve complete isolation between slices since segmentation of RAP uplink grants (UL grants) is not performed.

In [20], non-orthogonal random access (NORA) is proposed to reduce the problem of congestion in 5G networks. NORA is based on eliminating collisions caused by accesses from UEs that use the same preamble in *Msg1*. It does this by identifying the difference in arrival time of various UEs with identical preambles. The analysis carried out by simulation shows higher performance in terms of preamble collision probability and access success probability.

This paper considers an in-depth study of the RAP in a 5G network with NS to improve resource allocation. We focus on assigning subsets of preambles for each slice depending on the

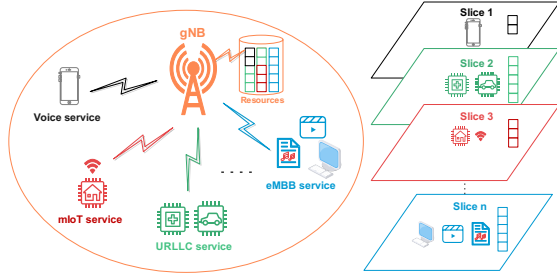


Figure 1. System model, 5G RAN with network slicing.

service's priority. The main contributions of our work are the following: i) a model for a scenario of n slices; that is, it has no limitation for the number of use cases, and ii) an efficient RAP resource allocation policy to maximize the probability of successful access by UEs in each slice.

III. SYSTEM MODEL

We consider resource allocation at the RAP in a cell with S slices as illustrated in Fig. 1. Each slice serves users of a different service: eMBB, voice, mIoT, and URLLC. Each service is assigned a priority level. Each UE in a slice s must complete the 4-step RAP to access the time-frequency resource blocks (RB) for data transfer. The RAP's physical resources (preambles) are allocated by the gNB to each slice, using a resource allocation policy.

We assume that arrivals are generated by a large population of independent users. Therefore, a Poisson process is appropriate to model the arrivals in each slice.

We consider that each of the slices is assigned a block of resources (preambles). Furthermore, we consider that preambles not assigned to any slice are shared and can be used by all slices. That is, we will have S slices and $S + 1$ preamble blocks (number of slices plus the shared block). Finally, it is assumed that UL grants will not be reserved for any slice. In each RAO, all accesses correctly detected by the gNB and that have not collided will compete for the available UL grants regardless of the slice they come from.

A. Collision Model

There are two collision models when two or more UEs simultaneously transmit the same preamble [13]. First, the gNB cannot decode the preambles transmitted by multiple UEs, so all the collisions occur in the transmission of $M_{sg}1$. Second, all $M_{sg}1$ are detected, and collisions occur in $M_{sg}3$. In this work, we intend to study the behavior of the RACH in extreme operation scenarios; therefore, we assume that the collision detection is performed in $M_{sg}1$. That is, only $M_{sg}1$ s that have been correctly decoded and have not collided will have the chance to receive a UL grant.

IV. ANALYTICAL MODEL

Unless otherwise stated or it is evident by the context, variables defined as "number of X" represent the average number of "X" per RAO.

Figure 2 illustrates how the resources are assigned for each slice. We are going to distinguish between slices and blocks

| Resources and system parameters | |
|---|------------|
| Total number of preambles | L |
| Total number of UL grants | G |
| Number of preambles reserved in the block i | L_i |
| Total number of UL grants reserved in the block i | G_i |
| Maximum number of transmission attempts, slice s | k_s^m |
| Power ramping parameter, slice s | Δ_s |
| Traffic | |
| Number of new arrivals, slice s | a_s |
| Number of transmissions that are in the k th attempt, slice s | $a_s(k)$ |
| Number of random access successfully completed, slice s | a_s^* |
| Number of transmissions in the block s | N_s |
| Average number for each random access, slice s | K_s |
| Probabilities | |
| Attempt k detection probability, slice s | $P_s^1(k)$ |
| Probability of receiving a UL grant, slice s | P_s^2 |
| Probability of no collision, slice s | P_s^{nc} |
| Probability of receiving a UL grant in the block s | P_s^3 |
| Probability of no collision in the block s | P_s^{nc} |
| Successful probability of the k th attempt, slice k | $P_s(k)$ |
| Successful probability | P_s |

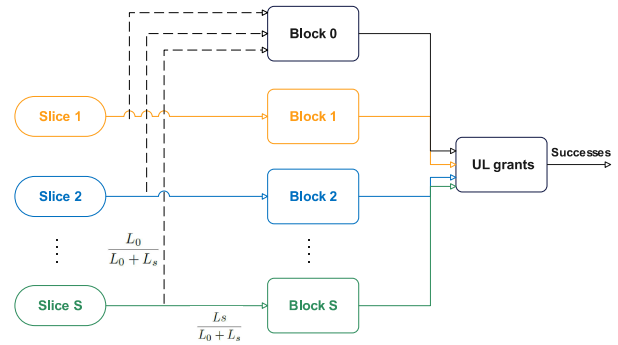


Figure 2. Slices and blocks of resources.

of resources. Each block is assigned many preambles. The distribution of access requests from each slice is proportional to the number of preambles assigned to each block. Thus, the fraction of accesses of slice s that use the shared block is given by

$$\frac{L_0}{L_0 + L_s}, \quad (1)$$

whereas the fraction that used the reserved block is given by

$$\frac{L_s}{L_0 + L_s}, \quad (2)$$

where L_s is the number of preambles reserved for each slice, and L_0 is the number of preambles reserved for block 0.

Let $a_s(k)$ be the number of transmissions of slice s that are in the k th attempt and k_s^m the maximum number of attempts. Taking into account the distribution between the shared common block and the reserved one, the average total number per RAO of preambles transmitted in each block s is obtained as

$$N_s = \frac{L_s}{L_s + L_0} \sum_{k=1}^{k_s^m} a_s(k), \quad s = 1, \dots, S. \quad (3)$$

The average number of preambles that use the shared block per RAO is obtained by adding the contribution of each slice:

$$N_0 = \sum_{s=1}^S \frac{L_0}{L_s + L_0} \sum_{k=1}^{k_s^m} a_s(k) = \sum_{s=1}^S \frac{L_0}{L_s} N_s. \quad (4)$$

For *Msg1* to be successfully transmitted, three conditions must be met: i) *Msg1* is correctly detected by the gNB, ii) *Msg1* does not collide, and iii) detected and not collided preambles get a UL grant. Therefore, to determine the probability of successful accesses, we will calculate the probabilities of success in these situations.

Msg1 detection probability: the probability of success in detecting *Msg1* will depend on the number of previous *Msg1* transmissions in the same access attempt. This is due to the power ramping scheme. To insert an additional level of prioritization between slices, the factor Δ_s is introduced to the 3GPP specification. Therefore, the probability of detecting the k th transmission attempt of a slice s preamble will be given by

$$P_s^1(k) = 1 - e^{-k\Delta_s}. \quad (5)$$

If we multiply the detection probability corresponding to each block by the total number of preambles used, we obtain the total number of detected preambles:

$$N_s^1 = \frac{L_s}{L_s + L_0} \sum_{k=1}^{k_s^m} P_s^1(k) a_s(k), \quad s = 1, \dots, S \quad (6)$$

$$N_0^1 = \sum_{s=1}^S S \frac{L_0}{L_s + L_0} \sum_{k=1}^{k_s^m} P_s^1(k) a_s(k) = \sum_{s=1}^S \frac{L_0}{L_s} N_s^1. \quad (7)$$

Msg1 no collision probability: with the number of preambles of each block and the number of preambles detected by the gNB, we can calculate the *probability of no collision* of the transmitted preambles in the block s as

$$p_s^{\text{nc}} = \left(1 - \frac{1}{L_s}\right)^{N_s^1 - 1}. \quad (8)$$

Probability of getting a UL grant: the probability that a preamble transmitted in block s will get a UL grant (probability of success in *Msg2*) can be estimated as

$$p_s^2 = \min\left(1, \frac{G}{g_s}\right), \quad (9)$$

where G is the number of UL grants available, and g_s is the average number of UL grants needed for the block preambles s , which is calculated as the product of the number of detected preambles and the probability of not having a collision

$$g_s = N_s^1 p_s^{\text{nc}}. \quad (10)$$

From the probabilities of success of *Msg2* and of not colliding, using the proportion of attempts that go through block 0 and the proportion that goes through the block reserved for slice s , it is possible to obtain the probabilities of success in *Msg2* and no collision in slice s as follows:

$$P_s^2 = \frac{L_s p_s^2 + L_0 p_0^2}{L_s + L_0}, \quad s = 1, \dots, S \quad (11)$$

$$P_s^{\text{nc}} = \frac{L_s p_s^{\text{nc}} + L_0 p_0^{\text{nc}}}{L_s + L_0}, \quad s = 1, \dots, S. \quad (12)$$

From (5), (11), and (12), we get the success probability of the k th attempt in slice s :

$$P_s(k) = P_s^1(k) P_s^2 P_s^{\text{nc}}. \quad (13)$$

If the number of new arrivals (first attempt) in slice s is a_s , we have:

$$a_s(1) = a_s \quad (14)$$

$$a_s(k+1) = a_s(k)(1 - P_s(k)), \quad k = 1, \dots, k_s^m - 1. \quad (15)$$

To calculate the *throughput* (average number of successfully completed accesses per RAO) of slice s , we add the product of the number of transmissions a_s by the probability of success P_s of each attempt k

$$a_s^* = \sum_{k=1}^{k_s^m} a_s(k) P_s(k). \quad (16)$$

Finally, the probability of success is calculated as the ratio of successful transmissions to total transmissions:

$$P_s = \frac{a_s^*}{a_s}. \quad (17)$$

In addition, the average number of attempts (preamble transmissions) in slice s is calculated as the sum of the total number of transmissions per RAO divided by the number of new transmissions per RAO:

$$K_s = \frac{1}{a_s} \sum_{k=1}^{k_s^m} a_s(k). \quad (18)$$

V. RESULTS

A. Model validation

The results of the analytical model have been validated with results obtained through computer simulation using MATLAB. For each numerical experiment, we set a basic load vector $a^0 = [a_1, \dots, a_s]$, which establishes the load share of each slice, and then the total load is scaled by a factor f ranging from 0.2 to 2, while the load share of each slice is kept constant $a = f a^0 = f [a_1, \dots, a_s]$. In the following, we detail the results according to the network configuration described in Table III.

Fig. 3 compares the results of the analytical model and the simulation. The horizontal axis represents each slice's initial load variation factor f . The initial load of each slice is the average number of RACH accesses per RAO. The results show a good match between the model and the simulation. The results for a low initial load are shown in Fig. 3. Fig. 4 depicts the results when we vary the initial load in one of the two slices. It is observed that the drop in performance of slice 2 is due to the increase in a load of access requests of slice 2. This is because there is no total isolation by having an assignment different from 0 in the subset of shared resources.

Table III
GENERAL RACH SLICING CONFIGURATION

| Parameter | Setting |
|--|-----------------------|
| PRACH Configuration Index | 6 |
| Subframe length | 1 ms |
| Total number of preambles | 54 |
| Maximum number of preamble transmissions | preambleTransMax = 10 |
| RAR window size | $W_{RAR} = 5$ |
| mac-ContentionResolutionTimer | 48 sub-frames |
| Maximum number of UL grants per subframe | $N_{RAR} = 3$ |
| Backoff Indicator | $BI = 20$ ms |
| HARQ re-transmission probability for <i>Msg3</i> and <i>Msg4</i> (non-adaptive HARQ) | 10% |
| Maximum number of HARQ TX for <i>Msg3</i> and <i>Msg4</i> (non-adaptive HARQ) | 5 |
| Periodicity of RAOs | 5 ms |
| Preamble transmission time | 1 ms |

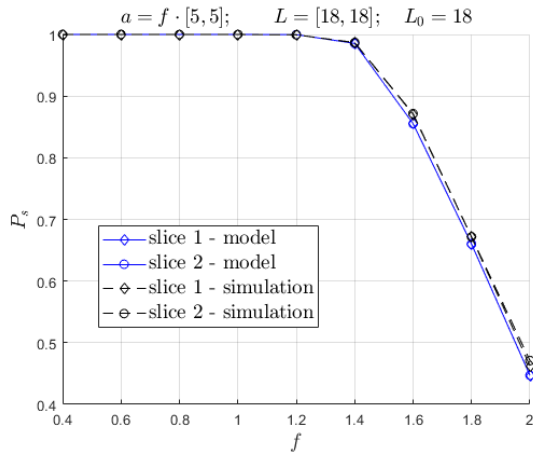


Figure 3. Successful access probability as a function of traffic load. Equitable allocation of resources for two slices in the RAN.

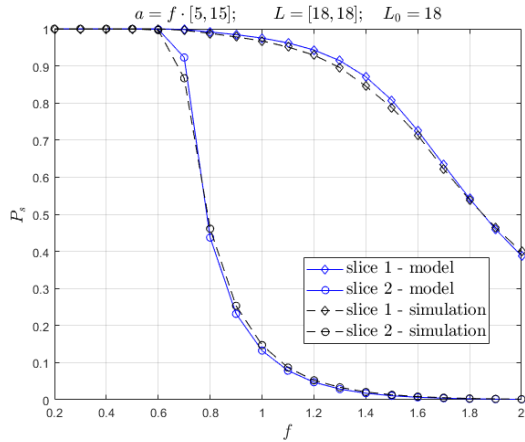


Figure 4. Successful access probability as a function of different traffic load per slice. Equitable allocation of resources for two slices in the RAN.

B. Equal sharing of resources

We analyze the behavior of the analytical model when an equal assignment of preambles is made for 2 slices.

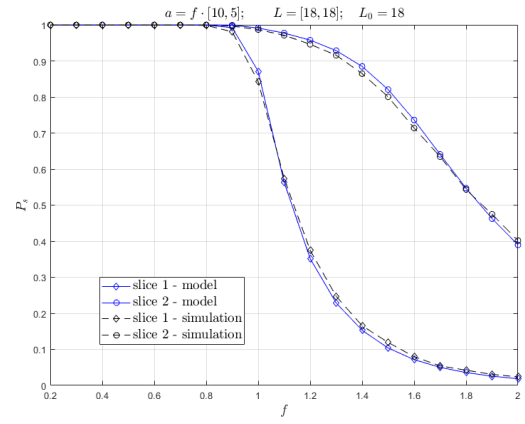


Figure 5. Successful access probability as a function of different traffic load per slice. Equitable allocation of resources for two slices in the RAN.

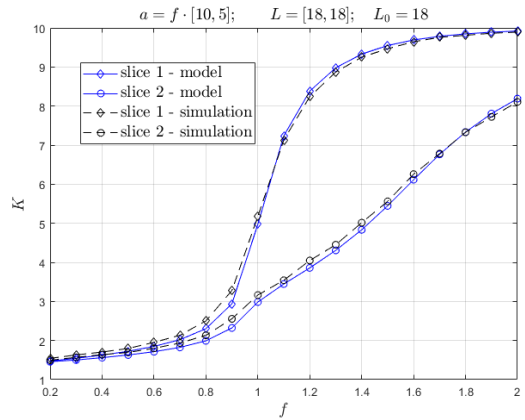


Figure 6. Average number of preamble transmissions as a function of different traffic load per slice. Equitable allocation of resources for two slices in the RAN.

We consider that the gNB will reserve several preambles for each block of resources equitably; this number is computed as

$$r_0 = \left\lceil \frac{R}{S+1} \right\rceil. \quad (19)$$

In Fig. 5, for an initial load $a_s = [10, 5]$, and resource blocks with an allocation $L = [18, 18]$ and $L_0 = 18$, a probability of successful access greater than 90% is obtained up to around $f = 1$ for slice 1 and $f = 1.4$ for slice 2, which represents an average of 10 and 7 access requests per RAO, respectively. Beyond this value, the RACH begins a drop in performance. As far as K is concerned, we can see in Fig. 6 that the number of retransmissions starts to increase significantly when $f > 0.8$ for slices 1 and 2. Reviewing these results, we can say that when $K > 3$, the performance of the RACH starts to drop.

C. Resource allocation proportional to load

To determine the percentage of available preambles allocated to the shared block, we use the coefficient δ as

$$L_0 = \lceil \delta R \rceil. \quad (20)$$

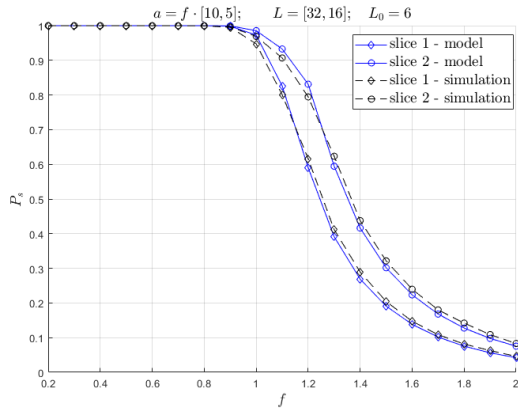


Figure 7. Successful access probability as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 2$.

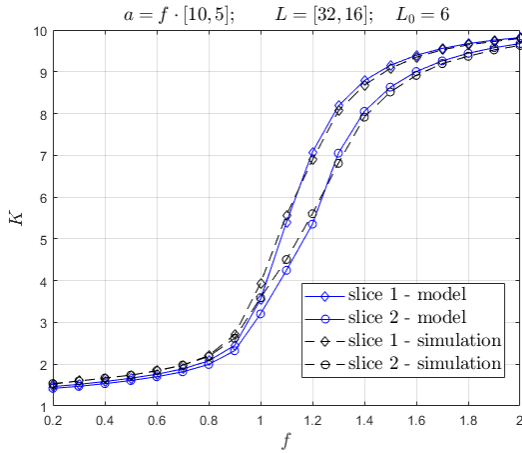


Figure 8. Average number of preamble transmissions as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 2$.

We calculate the number of preambles reserved for each slice from the remaining preambles. To do this, we use the proportion factor β as follows

$$L_1 = \beta L_2. \quad (21)$$

To avoid exceeding the number of available preambles, the maximum value that L_2 can take is 18 (when $\beta = 2$).

For the same initial load as in Fig. 5, in Fig. 7, we can observe the probability of success in the accesses for a scenario of 2 slices in which we set $\delta = 0.10$ and $\beta = 2$. A $P_s \geq 90\%$ is obtained up to around $f = 1.1$ for the 2 slices, averaging 11 and 5.5 access requests per RAO, respectively. Furthermore, we can see that slice 1 has a more pronounced drop in performance from this point on compared to slice 2. In Fig. 8, we can see again that at the inflection points of the curve, the mean number of retransmissions is approximately 3.

Figs. 9 and 10 illustrate the results when we set $\beta = 0.5$ and keep the remaining parameters the same. We can observe that the performance of slice 1 falls drastically due to the decrease in reserved resources, while the opposite occurs for slice 2.

As observed in Figs. 7 to 10, the differences between the model and the simulator for both P_s and K are almost

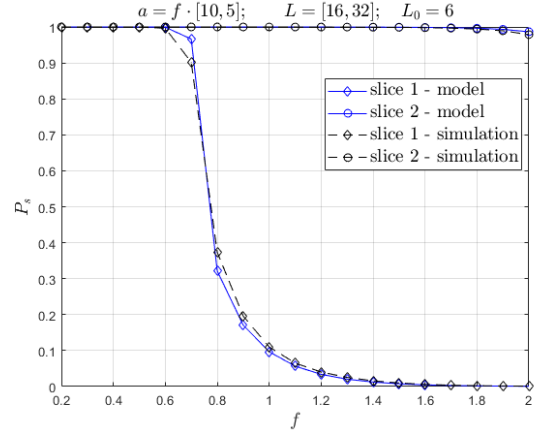


Figure 9. Successful access probability as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 0.5$.

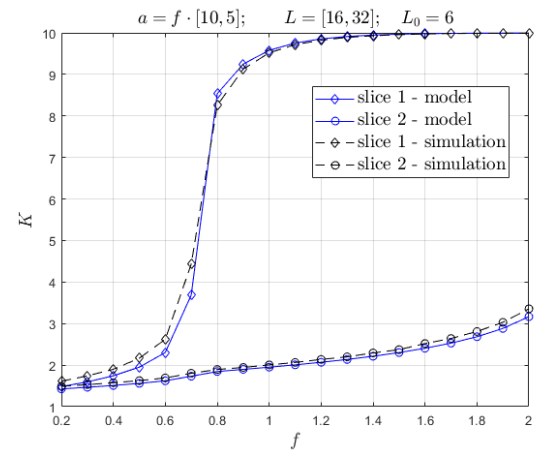


Figure 10. Average number of preamble transmissions as a function of different traffic load per slice. Proportional allocation of resources to the traffic load. $\beta = 0.5$.

indistinguishable. However, slight differences can be seen in certain areas due to multiple reserved blocks going into saturation simultaneously and competing with a higher load for shared resources. After this, the curves even overlap.

D. Increasing the number of slices

Finally, we will evaluate the analytical model when the number of slices exceeds 2. We assume a scenario with 5 slices with equitable allocation of resources serving services with different loads.

This scenario can represent a 5G network with 5 slices, each dedicated for each use case of Table I. As shown in Fig. 11, the performance drop in the P_s is related to the load of the slice. The higher the load, the faster the performance degrades. The same can be seen in Fig. 12, where those slices with a higher load carry out more retransmissions.

The results presented in Figs. 11 and 12 show that our model accurately represents the behavior of a 5G RAN with n slices.

VI. CONCLUSIONS

We have described an analytical model for the RAP of a 5G network implementing NS in detail. Our model can

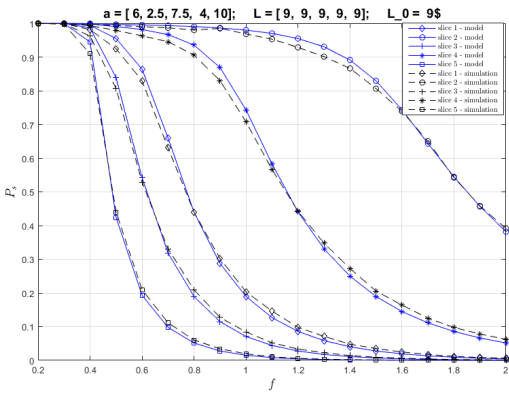


Figure 11. Successful access probability as a function of different traffic load per slice. Equitable allocation of resources for five slices in the RAN.

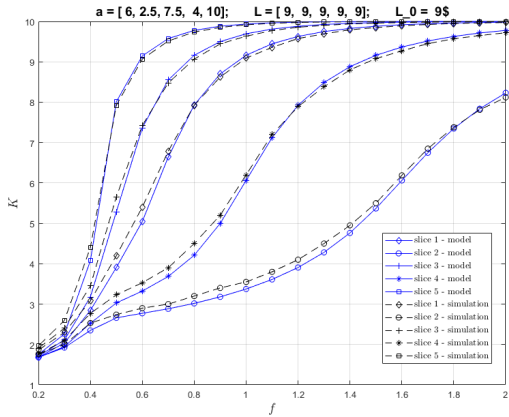


Figure 12. Average number of preamble transmissions as a function of different traffic load per slice. Equitable allocation of resources for five slices in the RAN.

be used to efficiently evaluate the performance of different resource allocation techniques to the 5G RAN slices. Furthermore, through evaluating performance indicators such as the probability of success in access and the number of necessary retransmissions, we have been able to analyze an equitable allocation of resources proportional to a load of each slice, the results of which have allowed us to validate our model. When performing segmentation of preambles and not of UL grants, we have observed partial isolation between slices. Resource isolation is one of the main applications of NS. It has been shown that the probability of success degrades significantly when the number of retransmissions is above three. Beyond this point, the RACH is severely congested. In future work, we plan to extend the model so that the case in which UL grants are reserved for each slice can be studied. This way, different network providers or tenants could use each slice virtually.

ACKNOWLEDGMENT

This work was supported in part by Cisco Systems, Inc., the NC State 2022 Faculty Research and Professional Development Program (FRPD), the NC Space Grant, the AC21 Special Project Fund (SPF), and the National Science Foundation (NSF) under Grant CNS-2210344. The work of

J. Martinez-Bauset and V. Pla was supported by Grants PGC2018-094151-B-I00 and PID2021-123168NB-I00 funded by MCIN/AEI/10.13039/501100011033 and ERDF *A way of making Europe*.

REFERENCES

- [1] J. Zhao, J. Liu, L. Yang, B. Ai, and S. Ni, "Future 5G-oriented system for urban rail transit: Opportunities and challenges," *China Communications*, vol. 18, no. 2, pp. 1–12, 2021.
- [2] Ericsson, "Ericsson mobility report," https://www.ericsson.com/4ae6a5/assets/local/reports-papers/mobility-report/documents/2021/emr_november2021_screen_epsanol.pdf, 2021.
- [3] M. Agiwal, A. Roy, and N. Saxena, "Next Generation 5G Wireless Networks: A Comprehensive Survey," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, 2016.
- [4] J. Navarro-Ortiz, P. Romero-Diaz, S. Sendra, P. Ameigeiras, J. J. Ramos-Munoz, and J. M. Lopez-Soler, "A Survey on 5G Usage Scenarios and Traffic Models," *IEEE Communications Surveys Tutorials*, vol. 22, no. 2, pp. 905–929, 2020.
- [5] N. Alliance, "5G white paper," *Next generation mobile networks, white paper*, vol. 1, no. 2015, 2015.
- [6] V. Mancuso, P. Castagno, M. Sereno, and M. A. Marsan, "Modeling MTC and HTC radio access in a sliced 5G base station," *IEEE Transactions on Network and Service Management*, vol. 18, no. 2, pp. 2208–2225, 2020.
- [7] Etsi.org, "5G System architecture for the 5G System (5GS) (3GPP TS 23.501 version 17.4.0 Release 17)," <https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3144>, 2022.
- [8] 3GPP, *TR 37.868, Study on RAN Improvements for Machine Type Communications*, Apr 2014.
- [9] L. Tello-Oquendo, J.-R. Vidal, V. Pla, and L. Guijarro, "Dynamic access class barring parameter tuning in LTE-A networks with massive M2M traffic," in *2018 17th Annual Mediterranean Ad Hoc Networking Workshop (Med-Hoc-Net)*, 2018, pp. 1–8.
- [10] S. Vural, N. Wang, P. Bucknell, G. Foster, R. Tafazolli, and J. Muller, "Dynamic preamble subset allocation for RAN slicing in 5G networks," *IEEE Access*, vol. 6, pp. 13 015–13 032, 2018.
- [11] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure and the Access Class Barring Scheme," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7785–7799, 2017.
- [12] J. Liu, M. Agiwal, M. Qu, and H. Jin, "Online Control of Preamble Groups With Priority in Massive IoT Networks," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 700–713, 2021.
- [13] L. Tello-Oquendo, I. Leyva-Mayorga, V. Pla, J. Martinez-Bauset, J.-R. Vidal, V. Casares-Giner, and L. Guijarro, "Performance analysis and optimal access class barring parameter configuration in LTE-A networks with massive M2M traffic," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3505–3520, 2017.
- [14] O. Arouk and A. Ksentini, "General model for RACH procedure performance analysis," *IEEE Communications Letters*, vol. 20, no. 2, pp. 372–375, 2015.
- [15] J.-R. Vidal, L. Tello-Oquendo, V. Pla, and L. Guijarro, "Performance study and enhancement of access barring for massive machine-type communications," *IEEE Access*, vol. 7, pp. 63 745–63 759, 2019.
- [16] V. Mancuso, P. Castagno, M. Sereno, and M. A. Marsan, "Serving HTC and Critical MTC in a RAN Slice," in *2021 IEEE 22nd International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM)*. IEEE, 2021, pp. 189–198.
- [17] C.-Y. Chang, N. Nikaein, and T. Spyropoulos, "Radio access network resource slicing for flexible service execution," in *IEEE INFOCOM 2018 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, 2018, pp. 668–673.
- [18] D. Nojima, Y. Katsumata, T. Shimojo, Y. Morihiro, T. Asai, A. Yamada, and S. Iwashina, "Resource isolation in RAN part while utilizing ordinary scheduling algorithm for network slicing," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*. IEEE, 2018, pp. 1–5.
- [19] O. Vikhrova, C. Suraci, A. Tropeano, S. Pizzi, K. Samouylov, and G. Araniti, "Enhanced radio access procedure in sliced 5G networks," in *2019 11th International Congress on Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*. IEEE, 2019, pp. 1–6.
- [20] T. Nomu and R. Aravind, "Non-Orthogonal Random Access scheme in Spatial Group Based Random Access for 5G Networks," *International Journal of Innovative Research in Technology*, vol. 6, 2019.