UBICOMM 2025

The Nineteenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies

ISBN: 978-1-68558-288-3

September 28th - October 2nd, 2025

Lisbon, Portugal

UBICOMM 2025 Editors

Jaime Lloret Mauri, Polytechnic University of Valencia, Spain

UBICOMM 2025

Forward

The Nineteenth International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies (UBICOMM 2025), held between September 28th, 2025, and October 2nd, 2025, in Lisbon, Portugal, continued a series of international events meant to bring together researchers from the academia and practitioners from the industry in order to address fundamentals of ubiquitous systems and the new applications related to them.

The rapid advances in ubiquitous technologies have made fruition of more than 35 years of research in distributed computing systems, and more than two decades of mobile computing. The ubiquity vision is becoming a reality. Hardware and software components evolved to deliver functionality under failure-prone environments with limited resources. The advent of web services and the progress on wearable devices, ambient components, user-generated content, mobile communications, and new business models generated new applications and services. The conference makes a bridge between issues with software and hardware challenges through mobile communications.

Advances in web services technologies along with their integration into mobility, online and new business models provide a technical infrastructure that enables the progress of mobile services and applications. These include dynamic and on-demand service, context-aware services, and mobile web services. While driving new business models and new online services, particular techniques must be developed for web service composition, web service-driven system design methodology, creation of web services, and on-demand web services.

As mobile and ubiquitous computing becomes a reality, more formal and informal learning will take place out of the confines of the traditional classroom. Two trends converge to make this possible: increasingly powerful cell phones and PDAs, and improved access to wireless broadband. At the same time, due to the increasing complexity, modern learners will need tools that operate in an intuitive manner and are flexibly integrated in the surrounding learning environment.

Educational services will become more customized and personalized, and more frequently subjected to changes. Learning and teaching are now becoming less tied to physical locations, co-located members of a group, and co-presence in time. Learning and teaching increasingly take place in fluid combinations of virtual and "real" contexts, and fluid combinations of presence in time, space and participation in community. To the learner full access and abundance in communicative opportunities and information retrieval represents new challenges and affordances. Consequently, the educational challenges are numerous in the intersection of technology development, curriculum development, content development and educational infrastructure.

We take here the opportunity to warmly thank all the members of the UBICOMM 2025 technical program committee, as well as all the reviewers. The creation of such a high-quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to UBICOMM 2025. We truly believe that, thanks to all these efforts, the final conference program consisted of top-quality contributions. We also thank the members of the UBICOMM 2025 organizing committee for their help in handling the logistics of this event.

We hope that UBICOMM 2025 was a successful international forum for the exchange of ideas and results between academia and industry for the promotion of progress related to mobile ubiquitous computing, systems, services, and technologies.

UBICOMM 2025 Chairs

UBICOMM 2025 Steering Committee Chair
Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain

UBICOMM 2025 Steering Committee
Stéphane Galland, Belfort-Montbéliard University of Technology, France
Wladyslaw Homenda, Warsaw University of Technology, Poland
Dmitry Korzun, Petrozavodsk State University, Russia

UBICOMM 2025 Publicity Chairs
Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

UBICOMM 2025
Committee

UBICOMM 2025 Steering Committee Chair

Jaime Lloret Mauri, Universitat Politecnica de Valencia, Spain

UBICOMM 2025 Steering Committee

Stéphane Galland, Belfort-Montbéliard University of Technology, France
Wladyslaw Homenda, Warsaw University of Technology, Poland
Dmitry Korzun, Petrozavodsk State University, Russia

UBICOMM 2025 Publicity Chairs

Laura Garcia, Universidad Politécnica de Cartagena, Spain
Lorena Parra Boronat, Universidad Politécnica de Madrid, Spain

UBICOMM 2025 Technical Program Committee

Afrand Agah, West Chester University of Pennsylvania, USA
Wafaa Ait-Cheik-Bihi, Schneider Electric, France
Mehmet Akşit, TOBB ET University, Ankara, Turkey / University of Twente, The Netherlands
A. B. M. Alim Al Islam, Bangladesh University of Engineering and Technology, Bangladesh
Mrim Alnfiai, Dalhousie University, Canada
Tahssin Altabbaa, Istanbul Gelisim University / Huawei Istanbul, Turkey
Nafisa Anzum, University of Waterloo, Canada
Paramasiven Appavoo, University of Mauritius, Mauritius
Mehran Asadi, Lincoln University, USA
Maryna Averkyna, Lviv Polytechnic National University | Estonian Business School, Estonia
F. Mzee Awuor, Kisii University, Kenya
Muhammed Ali Aydin, Istanbul University - Cerrahpasa, Turkey
Nebojsa Bacanin, Singidunum University, Serbia
Chiara Bachechi, University of Modena and Reggio Emilia, Italy
Matthias Baldauf, Eastern Switzerland University of Applied Sciences, Switzerland
Anoud I Bani-hani, Zayed University, Dubai, UAE
Luca Bedogni, University of Modena and Reggio Emilia, Italy
Oladayo Bello, New Mexico State University, Las Cruces, USA
Imed Ben Dhaou, University of Turku, Finland
Imen Ben Lahmar, ISIM Sfax | ReDCAD laboratory | University of Sfax, Tunisia
Djamal Benslimane, Université Claude Bernard Lyon 1, France
Aurelio Bermúdez, Universidad de Castilla-La Mancha, Spain
Javier Berrocal, University of Extremadura, Spain
Nik Bessis, Edge Hill University, UK
Robert Bestak, Czech Technical University in Prague, Czech Republic
Sourav Kumar Bhoi, Parala Maharaja Engineering College, India

Copyright Information

For your reference, this is the text governing the copyright release for material published by IARIA.

The copyright release is a transfer of publication rights, which allows IARIA and its partners to drive the dissemination of the published material. This allows IARIA to give articles increased visibility via distribution, inclusion in libraries, and arrangements for submission to indexes.

I, the undersigned, declare that the article is original, and that I represent the authors of this article in the copyright release matters. If this work has been done as work-for-hire, I have obtained all necessary clearances to execute a copyright release. I hereby irrevocably transfer exclusive copyright for this material to IARIA. I give IARIA permission or reproduce the work in any media format such as, but not limited to, print, digital, or electronic. I give IARIA permission to distribute the materials without restriction to any institutions or individuals. I give IARIA permission to submit the work for inclusion in article repositories as IARIA sees fit.

I, the undersigned, declare that to the best of my knowledge, the article is does not contain libelous or otherwise unlawful contents or invading the right of privacy or infringing on a proprietary right.

Following the copyright release, any circulated version of the article must bear the copyright notice and any header and footer information that IARIA applies to the published article.

IARIA grants royalty-free permission to the authors to disseminate the work, under the above provisions, for any academic, commercial, or industrial use. IARIA grants royalty-free permission to any individuals or institutions to make the article available electronically, online, or in print.

IARIA acknowledges that rights to any algorithm, process, procedure, apparatus, or articles of manufacture remain with the authors and their employers.

I, the undersigned, understand that IARIA will not be liable, in contract, tort (including, without limitation, negligence), pre-contract or other representations (other than fraudulent misrepresentations) or otherwise in connection with the publication of my work.

Exception to the above is made for work-for-hire performed while employed by the government. In that case, copyright to the material remains with the said government. The rightful owners (authors and government entity) grant unlimited and unrestricted permission to IARIA, IARIA's contractors, and IARIA's partners to further distribute the work.

Table of Contents

Space Data Centers – Future Development and Application Perspectives

Tassa Daniel
Cesi Engineering School
Rue Isabelle Autisser
Lagord, France
Email: daniel.tassa@viacesi.fr

Lilie-Leopold-Kateya
Africa Space Innovation Centre
Cape Peninsula Univ of Technology
Cape Town, South Africa
e-mail: leopoldl@cput.ac.za

Oluwaseyi P. Babalola
Africa Space Innovation Centre
Cape Peninsula Univ of Technology
Cape Town, South Africa
e-mail: babalolao@cput.ac.za

Gunjan Gupta
Africa Space Innovation Centre
Cape Peninsula Univ of Technology
Cape Town, South Africa
email: guptag@cput.ac.za

Ayodele A. Periola
Africa Space Innovation Centre
Cape Peninsula University of Technology
email: periolaa@cput.ac.za

Innocent E. Davidson
Africa Space Innovation Centre
Cape Peninsula University of Technology
email: davidsoni@cput.ac.za

Abstract— **The exponential growth of digital data presents escalating challenges for terrestrial data centers, including energy consumption, ecological impact, and cybersecurity risks. Deploying data centers in space emerges as a compelling alternative, leveraging microgravity, extreme temperatures, and isolation to enhance computational efficiency and environmental sustainability. This study explores the technological foundations and feasibility of emerging Space Data Center concepts. It also highlights the environmental burden imposed by terrestrial data centers, particularly their high water and land usage for hosting web services. In response, a space-based architecture is proposed to support low-cost, eco-friendly web development and content delivery. The research underscores space data centers as a forward-looking strategy for sustainable digital infrastructure.**

Keywords – Data centers; satellite; space Integrated; computing networks; photonics.

I. INTRODUCTION

Data centers are vital for hosting digital content accessed over the Internet, supporting key applications such as social media and data storage. Terrestrial data centers consist of multiple servers requiring substantial power and cooling, often leading to high water and land footprints. Water-intensive cooling systems, such as chillers and indirect evaporative cooling, contribute significantly to environmental strain [1], [2], while power demands escalate with data center scale [3], [4]. These constraints limit deployment in water-scarce, landlocked, or densely populated regions, resulting in content latency and power strain, especially in developing areas. Solutions to mitigate these challenges have been proposed in [5], [6]. The discussion in [5], and [6] also shows that future trends in the continued use of terrestrial data centers aim to address challenges in ensuring that data centers deploy their own power sources and systems. However, it is recognized that these approaches do not eliminate the competition for natural resources between terrestrial data center operators and other entities seeking to deploy power systems for other applications.

Terrestrial centers process space-based data [7], but the growing volume from small satellites and the latency in downlinking such data hinder timely decision-making. To address these limitations, there is a need for alternative infrastructure with lower environmental impact and reduced latency, the Space Data Center (SDC). SDCs eliminate dependence on Earth's land and water resources and are positioned to process satellite data in orbit, enabling rapid access and decision-making for latency-sensitive applications. They also host caches to support low-latency content delivery in landlocked regions, offering advantages over terrestrial centers for satellite-based communications.

The concept of SDCs is gaining global attention, with initiatives such as the European Union's ASCEND (Advanced Space Cloud for European Net Zero Emission and Data Sovereignty) project [8], [9] exploring large-scale orbital data centers powered by solar energy to reduce the carbon footprint of information technology. This paper explores the feasibility, technological underpinnings, and environmental implications of deploying data centers in space. It also investigates the role of SDCs in supporting web development tasks, where the reliance on cloud-based tools typically hosted on terrestrial infrastructure incurs significant environmental costs. By shifting such workloads to SDCs, the potential exists to achieve more sustainable computing.

The research being presented focuses on the design and application of SDCs in the area of web development. The motivation for the consideration of the web development application is that web development utilizes a significant proportion of existing computing resources. It is proposed that web development be migrated to SDCs. This has the benefit of reducing the environmental toll due to the use of existing terrestrial data centers for executing web development. The research presents a network architecture i.e., entity identification and describing entity relations. This is done to achieve the goal of executing web development aboard SDCs. In addition, the research recognizes the SDC

as the new computational workhorse for future execution of web development. It also presents the subsystems and components alongside space related elements for the realization of SDCs to be used in the proposed application. The use of SDCs is expected to be beneficial from an operational perspective. This is because of the abundance of solar power in space, which is accessible by SDC's onboard solar panels during the SDC's lifetime. The SDC does not need to incur high costs in comparison to terrestrial data center operation.

The remainder of this paper is structured as follows: Section II presents background work on data centers; Section III introduces the SDC technology concept; Section IV outlines potential advantages; Section V explores use cases and applications; Section VI discusses enabling hardware and software; and Section VII concludes the study.

II. Limitations and Challenges of Terrestrial Data Centers

Terrestrial data centers are centralized infrastructures that host computing and networking resources, including servers, storage systems, and communication hardware. They form the digital backbone for a broad spectrum of services, such as website hosting, cloud applications, enterprise IT operations, and content delivery. In particular, they are vital for web hosting and online enterprises, providing the essential computational support required for digital content accessibility across the globe. Currently, approximately 4,798 data centers are operating worldwide, with over 500 categorized as hyperscale facilities. These hyperscale data centers, typically operated by global technology firms, such as Amazon, Microsoft, Google, and Meta, are characterized by their immense capacity, architectural scalability, and advanced energy and infrastructure management capabilities [10]. In Africa, South Africa leads in terrestrial data center deployment, with major installations in Cape Town and Johannesburg [15]. These facilities support national and regional digital services but face significant operational constraints.

One of the most pressing challenges is the immense energy demand required to power servers and cooling systems. Data centers globally account for roughly 1–2% of electricity consumption, and this figure is expected to rise as digital services expand [11]. Maintaining an optimal thermal environment for servers further exacerbates energy consumption. Conventional cooling techniques, such as chilled water systems, are especially energy intensive. Even with improved methods like indirect evaporative cooling, the overall environmental toll remains significant, particularly due to high water usage [12].

The spatial requirements of large-scale data centers also contribute to broader urban planning and environmental challenges. These facilities often occupy large tracts of land, including prime real estate in urban areas, thereby competing with residential, agricultural, and infrastructural developments. In regions with high population densities or limited land availability, allocating space for new data centers becomes increasingly problematic. From an environmental standpoint, terrestrial data centers contribute significantly to global carbon emissions. Their high electricity consumption, coupled with reliance on water-intensive cooling systems, has led to increased scrutiny amid growing concerns about climate change and ecological degradation [16]. The environmental cost is especially acute in water-stressed and landlocked regions, where freshwater and land resources are either unavailable or severely constrained. This geographical limitation also leads to latency issues for users located far from major data center hubs.

Security and data privacy represent additional concerns. Data centers manage and store vast quantities of sensitive information, including personal, financial, and corporate data. As a result, they are frequent targets of cyberattacks. Breaches can result in severe financial and reputational damage, prompting constant investment in advanced cybersecurity infrastructure and regulatory compliance mechanisms [17].

Meanwhile, as the Internet of Things (IoT) continues to grow, the volume of data requiring real-time processing has increased significantly. This rise presents new latency and bandwidth challenges for conventional centralized data centers, which often struggle to efficiently support the responsiveness demanded by distributed IoT networks. Although architectural solutions, such as edge computing and fog computing address these issues, they add additional layers of complexity and cost [11].

In addition to supporting digital consumer services, terrestrial data centers process vast volumes of space-derived data, including imagery and telemetry from satellites. The transmission of such data from orbit to ground stations and then to terrestrial data centers introduces latency and relies on limited spectrum availability. As the number of small satellites and Earth observation missions increases, this bottleneck becomes more pronounced, hindering timely data analysis and decision-making [18]. These challenges, ranging from high energy and water consumption to land constraints, environmental impact, and latency, underscore the limitations of traditional terrestrial data center architectures. As demand for data processing escalates and the environmental and infrastructural costs of terrestrial data centers rise, the exploration of alternative solutions becomes increasingly justified.

III. The Concept of Space-based Data centers

The prospect of deploying data centers in space is increasingly gaining traction as a feasible future alternative to terrestrial infrastructures. Several leading technology corporations and governmental bodies, including the European Union through its ASCEND project, are actively exploring this alternative. The ASCEND initiative envisions the deployment of orbital data center stations powered by high-capacity solar power plants, potentially generating several hundred megawatts. The primary objective is to significantly reduce the environmental impact of information technology infrastructure by harnessing solar energy in the space environment, thereby mitigating the carbon footprint associated with traditional, Earth-based data centers.

One potential solution is the utilization of a satellite networking system. This approach would entail the collection of data from Earth, followed by its transmission to space for processing and storage. The system would employ photonics and optical technology, thereby reduce energy consumption and increase in the data transmission speed. Such a system would be immune to the effects of adverse weather conditions or natural disasters, ensuring uninterrupted communication. In a collaborative effort, Japan's NTT has joined forces with SKY Perfect JSAT to develop a satellite network system, designated as the Space Integrated Computing Network. As stated in [13], the Space Integrated Computing Network is a novel infrastructure to be constructed by combining NTT's network and computing infrastructure with SKY Perfect JSAT's space assets and business. The system will integrate multiple orbits from the ground to High-Altitude Platform Stations (HAPSs) flying at high-altitude, Low Earth Orbit (LEO) satellites and Geostationary orbit (GEO) satellites. The constituent components will be connected to the ground via an optical wireless communication network, thereby forming a constellation that will enhance the processing of various data sets through distributed computing. Furthermore, it will facilitate access to terrestrial mobile devices, thereby extending the service coverage to an ultra-wide range. Each satellite will be equipped with computing functions that can process data, connecting to a network of satellites that perform the function of an optical communications data center. This eliminates the necessity for data to be transmitted back to the Earth for processing and analysis, which impedes data traffic and consumes a considerable amount of power. Here is a description of each potential satellite orbit: (i) Low Earth Orbit: It is located at an altitude of between 200 and 2000 km above the earth, and is characterized by low latency time, better performance for real-time communications, LEO satellites move quickly around the Earth, requiring satellite constellation networks to ensure continuous coverage, (ii) Medium Orbits (MEO–Medium Earth Orbit) Approximately located at 2000 to 35786 km: fewer satellites needed for global coverage compared to LEO, MEO satellites can operate at similar efficiencies to fibre optics, even in remote areas of the world, and they are most often used for GPS, (iii) Geostationary Earth Orbit (GEO): It is located about 5000 km above the earth. GEO has the following characteristics: (i) A single satellite can cover a large part of the Earth's surface, (ii) Satellites remain stationary concerning a fixed point on Earth, and (iii) GEO satellites yield high ROI thanks to their high reliability and long life spans.

The scenario of a space-integrated computing network showing the role of the space data center in the context of a space application is shown in Figure 1. The application context of the SDC presented in Figure 1 is that of enabling low latency via free space optical networks for a space-based IoT (Space IoT) application. The SDC executes IoT data storage and Artificial Intelligence (AI) related processing. In this case, it is identified that computing platform-related applications using SDCs do not experience service interruptions from natural disasters. This provides an additional layer of physical level protection against the occurrence of natural disasters. Hence, the use of SDC is beneficial for use in regions with high susceptibility to natural disaster occurrence. The realization of the SDC in the scenario presented in Figure 1 recognizes the capability of SDCs to integrate with other non-terrestrial network entities such as HAPSs.

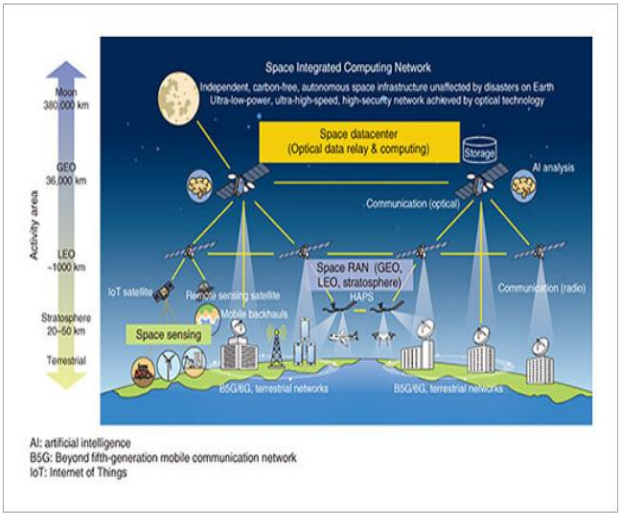

Figure 1. Space Integrated Computing Network.

The application context presented in Figure 1 is of enables the SDCs or a constellation of SDCs to accept, store, and process data from multiple Earth observation satellites. This is crucial considering the large number of in-orbit remote sensing satellites. The hosting of satellite systems enabling SDC functionality requires the derivation of global continuous coverage. This requires the specification of important orbital parameters such as: (i) Satellite orbital parameters, (ii) Orbital inclination, (iii) Number of orbital planes, and (iv) Satellite spacing. In the case of satellite orbital parameters, it is important to have global coverage with low latency. This requires ensuring proper organization of orbiting satellites. The important factors to consider are the orbital altitude. The orbital altitude influences: (i) Field of View: The higher the altitude, the wider the field of view of each satellite, but this can increase communication latency, (ii) Exposure to Space Debris: Lower altitudes may have a higher density of space debris, which increases the risk of collision, (iii) Lifespan: Satellites in lower orbit undergo greater atmospheric drag, which can reduce their lifespan unless regular orbit correction manoeuvres are performed.

The orbital inclination's related parameters are associated with latitudinal coverage and population access. The relevant parameters in this case are the latitudinal coverage and associated population access. The inclination determines the latitudes covered by the satellite. For example, an inclination of 90 degrees allows coverage up to the poles. The number of orbital planes also describes satellite distribution and influences the coverage density. Distributing satellites across multiple orbital planes allows for uniform coverage of the

Earth's surface, and Coverage Density. The more orbital planes there are, the denser the coverage, thus reducing poorly covered or uncovered areas. The use of multiple orbital planes also reduces interference between satellites. This is because it enables increased satellite spatial separation. In addition, the satellite spacing is an important orbital parameter as it influences : (i) Intervals: Satellites should be spaced evenly in each orbital plane to avoid interference and ensure uniform coverage, (ii) Revisit Time: Spacing affects the time it takes for a satellite to return above the same region (orbital period) in a given orbital altitude i.e., LEO, MEO and GEO, and (iii) Load Balancing. Proper spacing balances the communication load across the satellite network, ensuring optimal performance and reliability.

Besides the orbital aspects, the use of SDCs should consider their role in future network architectures alongside the evolution of crucial networking protocols. These aspects are crucial for describing the supported and realized data transmission. The data transmission protocol aspects are: (i) Photonics: These data centers will use photonics through Innovative Optical Wireless Network (IOWN) technology. This technology reduces satellite power consumption and allows satellites to withstand radiation better. (ii) TCP/IP over Satellite: The TCP/IP protocol, which forms the basis of the Internet, can be adapted for satellite communications. However, adjustments are necessary to accommodate latency and higher packet losses, and (iii) DTN (Delay-Tolerant Networking): This protocol is designed for environments where delays and interruptions are frequent, such as space. DTN stores and retransmits data until it can be delivered, ensuring reliable communication despite difficult conditions. The realization of low-latency and high data rate communications in SDC networks requires the use of optical communications. This is because optical communications offer higher data rates and better security. It is used for communications between satellites and between satellites and ground stations.

The realization of meaningful application-based communications with SDCs requires data exchange with ground stations. The required network architecture elements enabling this capability are: (i) ground stations, and (ii) relay (forwarding) satellites. Ground stations enable the establishment of a connection between higher-capacity terrestrial data centers and orbiting SDCs. This is useful in establishing uplink, and downlink connections. Relay satellites enable the execution of forwarding communications between orbiting SDCs and ground stations. This can include SDC satellites in low Earth orbit (LEO), medium Earth orbit (MEO), and geostationary Earth orbit (GEO), thereby ensuring global coverage, and reducing latency. In this case, the relations of SDCs with satellites in this context are presented in Figure 2.

The benefits and advantages of using SDCs are identified as: (i) Global coverage: The potential exists for SDCs to provide more equitable access to data services across the globe, (ii) Enhanced security: The physical isolation of SDCs offers additional protection against certain security threats, due to the reduced risk of external interference, (iii) Abundant solar energy. In the absence of atmospheric interference, solar panels in space could harness uninterrupted sunlight, thereby providing a constant and renewable energy source. The utilization of SDCs would not constitute an additional burden to the grid. In addition, the use of SDCs has a reduced environmental impact. By moving data centers off-planet, their direct impact on terrestrial ecosystems could be minimized.

Furthermore, SDCs benefit from natural cooling. The cold vacuum of certain locations in space may offer an optimal environment for cooling heat-generating SDC.

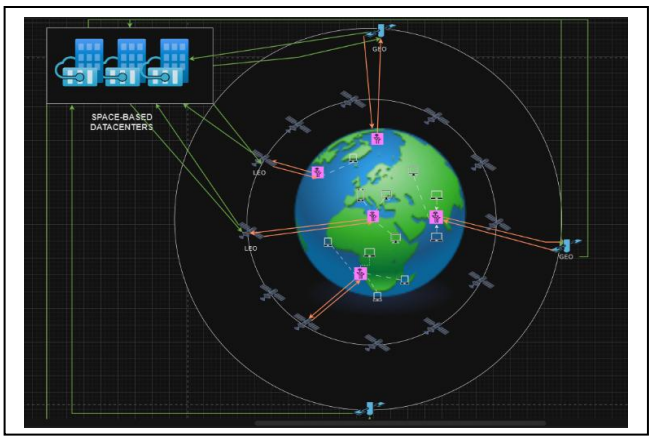

Figure 2. Space-based data center communications with satellites across the altitudes of LEO, MEO, and GEO.

Although the establishment of general-purpose large-scale SDCs (with power consumption more than 5.5 GW) may not be a realistic proposition soon, there are specific applications where large-scale SDCs could prove beneficial. Smaller SDCs with power consumption in the range of tens of MWs can be beneficial in (i) Scientific Research: The processing of data from space-based sources could prove advantageous for the analysis of extensive datasets generated by space telescopes, Earth observation satellites, and other scientific instruments in orbit, (ii) Disaster Recovery and Backup: Space-based data storage has the potential to serve as an ultra-secure backup solution for critical data, protected from terrestrial disasters, (iii) Edge Computing for Space Operations: As human activities in space increase, having computing resources in orbit could support various space operations, from satellite management to future lunar or Mars missions, and (iv) Global Communications Infrastructure: Although not yet at the level of full-scale data centres, satellite constellations providing Internet services are already exhibiting some of the principles that could lead to more sophisticated space-data processing capabilities.

IV. TECHNICAL CHALLENGES AND CONSIDERATIONS

The discussion in this section has two aspects. The first focuses on the SDC's components. The second focuses on the discussion of the application architecture for the SDCs in this case. The architecture is designed for SDC-anchored web development.

A. SDCs – Components and Enablers

SDCs present a promising alternative to traditional terrestrial facilities, particularly in addressing challenges related to energy consumption, environmental sustainability, and data latency. The development and deployment of SDCs involve a range of complex technical and operational challenges that must be addressed to realize their full potential.

One of the challenges is the launch and deployment of SDC infrastructure into orbit. Despite the significant reduction in launch costs in recent years, driven by advancements in aerospace technology and the increased availability of commercial launch vehicles [14], the expense associated with transporting sophisticated computing equipment to space remains substantial. In addition to cost, the physical design of these systems must account for the harsh launch conditions and the extreme space environment. Equipment must be robust enough to withstand vibrations and temperature fluctuations, while being resilient to radiation exposure, necessitating the use of radiation-hardened components.

The implementation of SDCs introduces complex regulatory considerations. The operation of such systems raises issues related to orbital rights, space debris mitigation, and international jurisdiction over data and communications. These challenges are compounded by the current lack of comprehensive global regulatory frameworks governing commercial data infrastructure in orbit. Effective policy will be essential in ensuring the responsible and sustainable use of orbital space for data storage and processing.

Energy management represents another critical aspect of SDC deployment. Although solar energy is abundant in space, the practical realization of a continuous and efficient power supply requires careful consideration. Energy systems must be designed to capture solar radiation effectively while accounting for panel degradation caused by radiation over time. Moreover, during eclipse periods, when solar power is temporarily unavailable, sufficient energy storage must be ensured through the integration of high-capacity batteries and intelligent power control systems.

Scalability is also a key consideration in SDC design. Unlike terrestrial data centers, which can expand horizontally by adding more physical infrastructure, orbital deployment is constrained by launch vehicle capacity and spaceborne volume limits. However, the relative abundance of orbital slots offers opportunities for distributed deployment. A constellation-based approach to SDCs could provide a scalable solution by enabling the deployment of multiple units across various orbital positions. This approach would also enhance redundancy and reduce latency by placing processing units closer to data sources in space.

Cost-effectiveness remains one of the most significant barriers to widespread adoption of SDCs. The high capital expenditure required for research, development, fabrication, launch, and maintenance must be weighed against long-term operational savings and environmental benefits. These benefits include a substantial reduction in terrestrial resource consumption, particularly water and land, and a minimized carbon footprint, especially when powered entirely by space-based solar energy systems. Nonetheless, whether SDCs can achieve economic competitiveness with terrestrial data centers in the near future remains an open question.

B. Proposed Low-Scale SDC Configuration

In this study, a low-scale SDC configuration is proposed as a viable and scalable architecture for orbital data processing. The model envisions a compact system with fewer than twenty servers, denoted as $N_s < 20$, optimized to minimize total mass and reduce launch costs. The overall system mass, M_{total}, is a summation of contributions from the servers, power subsystem, communication components, and thermal management infrastructure. The total SDC launch mass SDC is denoted M_{total} and given as:

$$M_{total} = M_s \cdot N_s + M_p + M_c + M_t \quad (1)$$

M_s is the mass of a single server, M_p corresponds to the power system mass, M_c is the mass of the communication subsystem, and M_t is the mass of the thermal management system.

The SDC's power budget is driven by the total average power consumption P_{avg}, which aggregates the power requirements of the computing, communication, and control electronics. This is expressed as:

$$P_{avg} = N_s \cdot P_s + P_{comm} + P_{control} \quad (2)$$

P_s denotes the power consumed per server, P_{comm} is the power consumption of the communication system. and $P_{control}$ is the power consumption of the operational control units.

Energy generation is handled through photovoltaic solar arrays. The total daily energy produced in orbit and accessible for SDC operation is denoted E_{day}, and given as:

$$E_{day} = \eta_{solar} \cdot A_{panel} \cdot G_{orbital} \cdot t_{illum} \quad (3)$$

where η_{solar} represents solar conversion efficiency, A_{panel} is the area of the solar panels, $G_{orbital}$ is the solar irradiance in orbit (approximately $1361\ W/m^2$), and t_{illum} is the duration of sunlight exposure per orbit.

During eclipse phases, the system relies on onboard batteries with energy storage capacity E_{bat} to maintain operations. This requires that $E_{bat} \geq P_{avg} \cdot t_{eclipse}$, ensuring sufficient energy availability when solar input is absent. Efficient thermal management is a critical design consideration, especially in the vacuum of space. Instead of traditional fluid-based cooling, the SDC relies on radiative heat dissipation, governed by the Stefan–Boltzmann law. The total radiated thermal power , Q_{rad} is modeled as

$$Q_{rad} = \epsilon \cdot \sigma \cdot A_{rad} \cdot \left(T^4 - T_{space}^4\right) \quad (4)$$

where ϵ is the emissivity of the radiator surface, σ is the Stefan–Boltzmann constant, A_{rad} is the surface area of the radiators, T is the radiator surface temperature, and T_{space}, approximately 3K, is the ambient temperature of space.

The relation in (4) serves to identify that the proposed SDC will be cooled via the radiation. In this case, heat pipes from the server interior will be connected to SDC exterior radiators. This ensures that server electronics are maintained at the requisite operational temperature.

Communication between the SDC and terrestrial stations is achieved through high-gain directional antennas, operating in the X or Ka bands. The link budget is constrained by the signal-to-noise ratio (SNR), given by the equation:

$$SNR = \frac{P_t \cdot G_t \cdot G_r \cdot \lambda^2}{(4\pi d)^2 \cdot k \cdot T_{sys} \cdot B} \quad (5)$$

P_t is the transmit power, G_t and G_r are the gains of the transmit and receive antennas respectively, λ is the operating SDC wavelength, d is the distance from the SDC to Earth, k is Boltzmann's constant, T_{sys} is the system noise temperature, and B is the communication bandwidth.

Through the coordinated integration of these subsystems into a cohesive and space-resilient design, the low-scale SDC illustrates the feasibility of deploying efficient and environmentally sustainable data processing infrastructure in orbit. This approach reduces reliance on land and water, and introduces new paradigms in distributed computing and remote sensing data processing, setting the stage for more ambitious deployments in the future.

C. SDC – Application Architecture

The proposed SDC architecture enables the remote development and hosting of websites using space-based computational infrastructure. Situated in LEO, the SDC is configured to support interactive web development tasks by Earth-based developers who access hosted tools, libraries, and development environments through satellite Internet connections.

A typical SDC pass over a ground station provides a communication window of approximately 7 minutes, or $T_w = 420$ seconds. Given the average round-trip latency of $L = 0.12$ seconds, the maximum number of potential bidirectional communication epochs per pass is $N_e = \frac{T_w}{L} \approx 3500$. This suggests that despite inherent latency, a significant number of interactions can be supported during each orbital pass, allowing developers to upload scripts, receive feedback, and test website functionality.

To manage this interaction efficiently, the SDC integrates interconnected logical entities. Commands issued by a developer from Earth are first received by the Script Receiver Entity (SSRE), which acts as the primary interface for incoming web development instructions. These commands, denoted $C_d(t)$, are passed to the Web Rendering Entity (WRE), responsible for executing them and rendering the resulting webpage. The rendered webpage state $R(t)$ evolves based on the input command as $R(t + \Delta t) = f(C_d(t))$, where $f(\cdot)$ is a function that maps developer input to the updated state of the web interface.

Given the latency sensitivity of space-based communication, the architecture includes a Render Pause Entity (RPE), which temporarily halts script execution when latency levels exceed tolerable thresholds. This behavior is informed by the Communication Profile Entity (CPE), which continuously monitors communication latency. Suppose the current latency measurements over a window of n epochs are $\{L_1, L_2, \ldots, L_n\}$; the moving average latency $\bar{L}$ is computed as $\bar{L} = \frac{1}{n}\sum_{i=1}^{n} L_i$. If the current latency L_c exceeds the average by a defined margin δ such that $L_c > \bar{L} + \delta$, then the CPE signals the RPE to activate, resulting in a temporary suspension of script execution, effectively setting a script execution flag $S_{exec} = 0$. This mechanism prevents unstable page rendering due to excessive delay.

In parallel, web content that has already been developed and is ready for access is managed by the Web Access Entity (WAE). When user requests are received, either from developers or end-users, the CPE determines whether the request pertains to command execution or content retrieval. If it concerns access, the request is routed to the WAE, which retrieves and prepares the corresponding webpage $P(t)$ for downlink. These components interact bidirectionally: the uplink allows for script transmission and page development, while the downlink returns rendered output and user-facing web pages. All data exchanges between users and the SDC are routed through an Internet Exchange Point (IXP), which connects the orbital system to terrestrial Internet infrastructure. The final content delivery process can be modeled as $P_{out}(t) = g(P(t), U(t))$, where $U(t)$ is the user request and $g(\cdot)$ denotes the IXP routing function.

Through this architecture, the SDC demonstrates the feasibility of low-latency, interactive web development and hosting in orbit. It addresses key latency challenges through computational buffering, latency-aware pausing, and an intelligent routing scheme. The resulting system not only serves as a viable complement to terrestrial web services but also establishes the foundation for scalable, environmentally responsible orbital computing platforms.

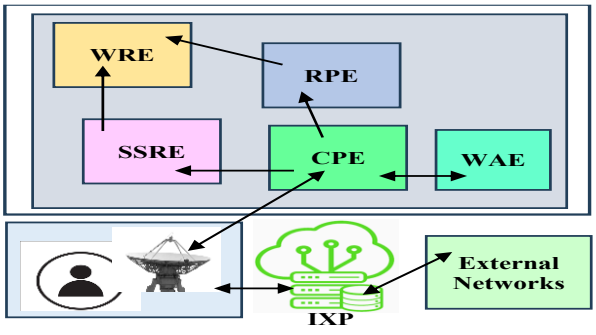

Figure 3. Architecture showing the relations between the entities WRE, SSRE, RPE, CPE, and the WAE.

V. CONCLUSION

The concept of Space Data Centers (SDCs) represents a convergence of data technology and space exploration. The miniaturization of electronics and a reduction in launch costs could render aspects of this concept feasible. From the perspective of near-term prospects, it is probable that there will be more advances in the processing capabilities of data in space rather than the establishment of comprehensive orbital data centres. Such developments could encompass the implementation of enhanced onboard computing for satellites and the utilization of limited-scale experimental platforms. SDC use in the medium term presents opportunities that should be further analyzed. From this perspective, we might see the deployment of small, specialized data centers in Low Earth Orbit (LEO). The discussion presents an application context for the LEO based SDC. The application is one in which the SDC enables web development and accessing developed web pages. Such an application is considered to have a significant role due to the pervasive deployment and use of web portals on existing data centers (with a high environmental toll). The research presents an architecture enabling subscribers to access the web page developed via rendering and execution aboard the SDC. Future work will focus on enabling additional functionalities for the SDCs. Advances in autonomous maintenance and space-based power generation will be key to making SDCs feasible, setting the stage for larger future projects. In the long term, as human space exploration and application development improves, the development of a space-based data infrastructure may become a necessity. Future work will address how the use of the medium earth orbit by the SDC can be sustainably realized. In addition, future work will address the need to design web development protocols that are suited to the space-based data center as a precursor to the conduct of performance analysis.

ACKNOWLEDGMENTS

The support of the French South Africa Institute of Technology (F'SATI) / African Space Innovation Centre (ASIC) and the Dept. of Electrical, Electronic & Computer Engineering, Cape Peninsula University of Technology is acknowledged. The role of the supervisors Dr. Ayodele A Periola and Dr. Gunjan Gupta is acknowledged for their conceptualization, technical input, guidance, and motivation.

REFERENCES

[1] R. Tariq, N. Sheikh, A. Livas-García, J. Xamán, A. Bassam, and Valeriy Maisotsenko, "Projecting global water footprints diminution of a dew-point cooling system: Sustainability approach assisted with energetic and economic assessment," Renewable & Sustainable Energy Reviews, vol. 140, pp. 110741–110741, Apr. 2021, doi: https://doi.org/10.1016/j.rser.2021.110741.

[2] H. S. Arunkumar, N. Madhwesh, S. Shenoy, and S. Kumar, "Performance evaluation of an indirect-direct evaporative cooler using biomass-based packing material," International Journal of Sustainable Engineering, vol. 17, no. 1, pp. 1–12, May 2024, doi: https://doi.org/10.1080/19397038.2024.2360451.

[3] D. Thangam et al., "Impact of Data Centers on Power Consumption, Climate Change, and Sustainability," Advances in Computational Intelligence and Robotics book series, pp. 60–83, Mar. 2024, doi: https://doi.org/10.4018/979-8-3693-1552-1.ch004.

[4] K. M. U. Ahmed, M. H. J. Bollen, and M. Alvarez, "A Review of Data Centers Energy Consumption and Reliability Modeling," IEEE Access, vol. 9, pp. 152536–152563, 2021, doi: https://doi.org/10.1109/ACCESS.2021.3125092.

[5] A. A. Periola and I. E. Davidson, "Networks and Smart Grids for Future Data Centres," 2025 33rd Southern African Universities Power Engineering Conference (SAUPEC), pp. 1–6, Jan. 2025, doi: https://doi.org/10.1109/saupec65723.2025.10944393.

[6] L. Barroso, Urs Hölzle, and P. Ranganathan, The Datacenter as a Computer. 2019. doi: https://doi.org/10.1007/978-3-031-01761-2.

[7] N. Kussul, A. Shelestov, and B. Yailymov, "Cloud Platforms and Technologies for Big Satellite Data Processing," Lecture Notes in Networks and Systems, pp. 303–321, 2023, doi: https://doi.org/10.1007/978-3-031-46880-3_19.

[8] "Data Centres in Space," ASCEND, [Online]. Available: https://ascend-horizon.eu/data-centres-in-space/. Accessed: May 16, 2025

[9] Thales Alenia Space, "Thales Alenia Space Reveals Results of ASCEND Feasibility Study on Space Data Centers," [Online]. Available: https://www.thalesaleniaspace.com/en/press-releases/thales-alenia-space-reveals-results-ascend-feasibility-study-space-data-centers-0. Accessed: May 16, 2025.

[10] J. Dumont, "What is a data center's carbon footprint?" [Online]. Available: https://greenly.earth/fr-fr/blog/actualites-ecologie/quelle-est-l-empreinte-carbone-d-un-data-center. Accessed: May 16, 2025.

[11] New Space Economy, "The potential and challenges of space-based data centers," [Online]. Available: https://newspaceeconomy.ca/2024/06/24/the-potential-and-challenges-of-space-based-data-centers/. Accessed: May 16, 2025

[12] I. Boglaev, "A numerical method for solving nonlinear integro-differential equations of Fredholm type," *J. Comput. Math.*, vol. 34, no. 3, pp. 262–284, May 2016, doi: 10.4208/jcm.1512-m2015-0241.

[13] Space Compass Corporation, "Overview of space integrated computing network," [Online]. Available: https://www.rd.ntt/e/research/JN202210_19855.html. Accessed: May 16, 2025.

[14] W. W. Baber and A. Ojala, "New Space Era: Characteristics of the New Space Industry Landscape," pp. 3–26, Jan. 2024, doi: https://doi.org/10.1007/978-981-97-3430-6_1.

[15] Esi Africa, "Deploying Data Centres for Sustainable Digitisation," [Online]. Available: https://www.esi-africa.com/magazine-article/deploying-data-centres-for-sustainable-digitisation/. Accessed: May 16, 2025

[16] A. Shehabi, S. Smith, D. Sartor, R. Brown, M. Herrlin, J. Koomey, E. Masanet, N. Horner, I. Azevedo, and W. Lintner, "United States Data Center Energy Usage Report," 2016

[17] S. Dongre, S. Mishra, C. Romanowski, and M. Buddhadev, "Quantifying the Costs of Data Breaches," in Critical Infrastructure Protection XIII: 13th IFIP WG 11.10 International Conference, ICCIP 2019, Arlington, VA, USA, Mar. 11–12, 2019, pp. 3-16, Springer International Publishing, 2019.

[18] I. Siddique, "Emerging Trends in Small Satellite Technology: Challenges and Opportunities," European Journal of Advances in Engineering and Technology, vol. 11, no. 2, pp. 42-48, 2024.

Adaptive Microgrid Architecture to Manage System Resiliency

Mobolaji Bello
Transmission Operations and Planning
Electric Power Research Institute (EPRI)
Palo Alto, USA
email: mbello@epri.com

Davis Montenegro
Power System Project
New Math Data
Maryville, USA
email: dmontenegro@newmathdata.com

Oladayo Bello
New Mexico State University
College of Engineering (ETSE)
Las Cruces, USA
email: oladayo@ieee.org

Abstract— **In light of the growing risks posed by high-impact, low-frequency events (such as those driven by climate change and other emerging hazards) utilities are increasingly deploying Distributed Energy Resources (DER), both utility- and customer-owned, to enhance grid reliability. These assets play a vital role in addressing system constraints during peak demand (thermal and voltage), mitigating power outage impacts, and improving overall resilience by supporting the formation of microgrids when distribution grid integrity is compromised. Yet, microgrid deployment presents its own technical challenges, particularly in coordinating the DERs involved. Critical functions such as grid separation (islanding), black start procedures, operational control, and eventual grid reconnection, must be executed with precision to ensure system stability. Poor coordination can exacerbate existing grid disturbances, extend recovery timeframes, and ultimately undermine the very resilience the microgrid is intended to deliver. To address these challenges, this paper proposes a microgrid architecture anchored by three resilience-enhancing pillars: (1) robust protection and power quality, (2) high-speed, reliable communication infrastructure, and (3) Machine Learning (ML) driven control and management. Each pillar is introduced through its operational goals and technical contributions, followed by a test case illustrating the integrated architecture in action.**

Keywords—Adaptive control; machine learning; microgrid; robust control; power system resilience.

I. INTRODUCTION

The conversion of energy into electricity remains a cornerstone of societal advancement. As underscored by the Rockefeller Foundation [1], electricity is now a more pivotal driver of economic growth and global competitiveness than ever before. Even in developing regions, robust, secure, and reliable power systems are essential to economic stability and social progress.

This widespread dependence on electricity has spurred innovation and improved quality of life, but it also exposes the urgency of updating aging infrastructure. A modernized grid is key to ensuring resilient and consistent power delivery to both everyday consumers and critical facilities.

Since the earliest stages of electrification, the safety and reliability of power systems have been foundational concerns, particularly within transmission networks, given their central role in system performance [2], [3], [4]. Within distribution systems, however, safety, strategic planning, and system availability are equally vital, enabling the efficient transfer of power from high-voltage transmission to end users.

To evaluate the reliability of electric power systems, regulators and utilities employ service quality indicators; quantitative metrics designed to measure system performance [2]. In distribution networks, reliability assessments typically draw on infrastructure and equipment data to estimate outage restoration times. While such outages are generally short and frequent, they are accounted for in power system design. However, their cumulative impact over the course of a year can degrade service quality metrics, potentially triggering financial penalties to offset disruptions experienced by customers.

Beyond these routine disturbances, power delivery can be compromised by rare but severe events capable of causing extensive damage. The system's ability to recover from such high-impact failures defines its resilience [5], [6]. Though a universally accepted technical definition is still lacking, experts broadly agree that resilience is associated with low-probability, high-consequence disruptions [7].

The U.S. Department of Energy (DOE) has noted the lack of universally accepted metrics for assessing grid resilience. As a result, federal policy does not prescribe specific resilience standards for electric systems [8], [9]. Instead, resilience (defined by the grid's ability to adjust to evolving conditions and recover rapidly from disruptions) is treated as an integral aspect of the broader reliability framework.

Multiple factors affect the resilience of distribution systems, ranging from natural disasters to human-induced risks, such as cyber-attacks, labor shortages, and other societal dynamics. When resilience is diminished, the repercussions often extend beyond infrastructure damage, posing risks to vulnerable communities and broader societal functions.

In response, utilities are increasingly integrating Distributed Energy Resources (DER), whether utility-owned or customer-owned, to improve grid reliability. These resources help mitigate system violations during periods of high demand (thermal and voltage), reduce the impact of power outages, and bolster resilience by enabling the formation of microgrids when the integrity of the distribution grid is disrupted [10].

However, forming a microgrid introduces its own set of challenges, primarily due to the need for precise coordination of the DER involved. Key activities (including islanding from the distribution grid, black start procedures, operational control, and reconnection) must be carefully managed to ensure stability. Poor coordination during these stages can exacerbate grid issues, prolong restoration efforts, and turn a potential solution for reliability and resilience into a source of additional disruption.

This paper proposes a microgrid architecture grounded in three key pillars of resilience: (1) robust protection and power quality, (2) fast and reliable communication infrastructure, and (3) Machine Learning (ML) based management for intelligent microgrid control. Each pillar is briefly examined through its objectives, culminating in a test case to demonstrate the proposed approach in practice. Section II describes the resilient microgrid architecture, Section III highlights a case study while Section IV concludes the paper and highlights future work.

II. RESILIENT MICROGRID ARCHITECTURE

As outlined earlier, the three foundational pillars of resilient microgrid architecture serve as a framework for the effective coordination of DER. These pillars encompass key technical recommendations designed to address the following operational challenges and performance objectives (illustrated in Fig. *1*):

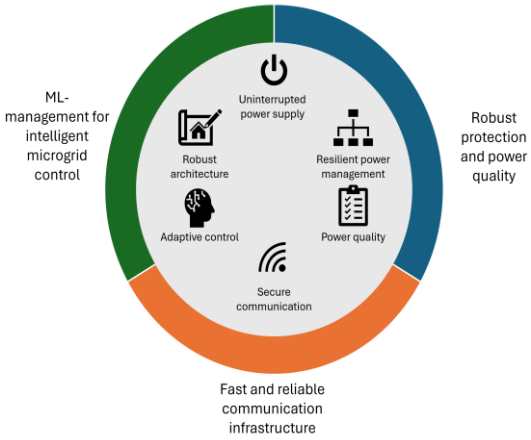

Fig. 1. Performance objectives of the resilient microgrid architecture.

- Uninterrupted Power Supply: Maintain service to controllable loads despite intermittent generation.
- Resilient Power Management: Leverage energy storage and dynamic load control to mitigate generation variability.
- Current Imbalance Minimization: Apply targeted techniques to preserve power quality across phases.
- Secure Communication: Ensure the integrity, reliability, and responsiveness of control signal transmission.
- Adaptive Control: Incorporate predictive control, anomaly detection, and optimization strategies to enhance operational intelligence.
- Robust System Architecture: Define the microgrid's structural design and its supporting communication network.

These performance objectives are described as follows.

A. *Uninterrupted power supply*

This goal of the resilient microgrid architecture is composed by the DER deployed within the microgrid. It includes:

- Intermittent DER such as solar Photovoltaic (PV), wind turbines.
- Energy storage systems, such as battery banks or buffer intermittent generation.

These energy resources, whether utility-owned or customer-owned, must be coordinated during microgrid formation to account for their availability and operational roles. This includes identifying devices that provide a grounding reference, such as Grid Forming Inverters (GFM), as well as supporting generation sources configured to follow the reference, such as Grid Following Inverters (GFL) [11], [12], [13].

The available DER capacity within the microgrid influences its charge and discharge cycles, as well as the operational usage rate while grid connected. This coordination supports preparation for potential islanding events. Additionally, DER capacity determines the microgrid's autonomy (defined by the number of hours it can supply energy independently) and governs the usage rate sustainable during island mode.

B. *Resilient power management*

In addition to DER, intelligent load-controlling devices play a key role in managing energy within a microgrid. Examples include smart thermostats, switches, and heat strips, controllable loads that the microgrid controller can leverage to reduce demand and extend the duration of available energy resources [14], [15]. See an illustration in Figure 2.

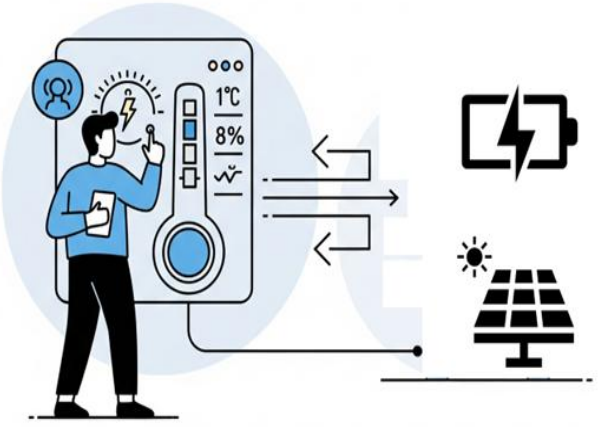

Fig. 2. Demand side management from within a microgrid.

This is especially relevant for energy storage systems, where discharge rates depend directly on the load profile. Coordinated adjustments (such as slightly lowering thermostat setpoints) can help batteries deliver additional service hours during periods without local generation, all while minimizing customer discomfort.

Traditionally, demand-side management has been used to help utilities mitigate system violations (thermal and voltage) while connected to the grid, often enabling deferral of large infrastructure investments. Within the context of microgrids, controllable loads offer an additional advantage: they transform demand into a dynamic energy management tool, enabling finer optimization of energy use and supporting resilient island-mode operation.

Power intermittency, stemming from the variability of available generation types, is a key consideration in microgrid power management. To maintain uninterrupted supply and improve system resilience, the following strategies can be implemented:

- Energy Storage Sizing: Leverage historical generation and load data to appropriate size Energy Storage Systems (ESS), ensuring coverage of worst-case generation deficits.
- Load Prioritization: Categorize controllable loads into tiers (critical, semi-critical, and non-critical) and implement load shedding for non-critical demands during supply shortfalls.
- Demand Response (DR): Dynamically adjust controllable loads to align with real-time generation availability.
- Together, these measures support optimized energy utilization within the microgrid, helping ensure reliable performance and extended autonomy during islanded operation.

Equation (1) serves as a reference point for assessing the microgrid's operational status by evaluating the current load relative to the available DER.

$$Pgen + Pdis(t) - Pch(t) = Pload(t) - Pshed(t) \quad (1)$$

Where:

Pgen(t): Power from renewable sources.
Pdis(t), Pch(t): Discharging and charging power of ESS.
Pload(t): Total load demand.
Pshed(t): Sheddable load (non-critical loads).

The ESS State of Charge (SOC) is updated as indicated in (2).

$$min \sum_{t=1}^{T} Pshed(t) \quad (2)$$

This is subject to the constraint of the ESS state of charge, SoC:

$$SoCmin \leq SoC(t - 1) \leq SoCmax \quad (3)$$

The objective of this equation is to minimize load shedding by maximizing the utilization of renewable generation.

C. Current imbalance minimization

Managing current imbalance is a critical operational objective within microgrids. It supports maintaining voltage levels within acceptable ranges, prevents conductor overload, and minimizes zero-sequence current; factors that, if left unaddressed, can contribute to system faults and compromised reliability [16], [17].

In a three-phase system, current imbalance can lead to voltage imbalance and equipment damage. Current imbalance is defined as the deviation from balanced three-phase currents. This impact can be calculated using the Current Imbalance Impact (CII) [18] shown at (3).

$$CII = \frac{|I_{max} - I_{avg}| + |I_{min} - I_{avg}|}{I_{avg}} * 100\% \quad (4)$$

Where Imax, Imin, Iavg are the maximum, minimum, and average of the three-phase currents.

A robust control strategy is essential for maintaining phase balance and minimizing disruptions within microgrid operations. This strategy begins with actionable interventions, such as applying phase swapping for single-phase loads where technically feasible and dynamically adjusting the operation of controllable loads across phases to correct imbalance. These approaches provide the groundwork for a more intelligent and responsive microgrid framework.

Building on this, the control strategy is formalized through an optimization problem aimed at minimizing CII. The optimization targets load-level power adjustments on each phase, governed by system-level constraints that ensure total power demand is met, either fully or within allowable shedding margins, and that device-specific constraints, such as minimum on/off durations, are respected.

By harmonizing device-level control with system-wide optimization, this framework supports both operational reliability and efficiency. It enables microgrids to handle variable demand profiles and DER with greater agility, paving the way for more resilient and adaptive energy ecosystems.

D. Secure Communication

Microgrid operation relies on uninterrupted, low-latency data exchange among controllers, sensors, DERs, and loads. Secure communication is crucial to ensure control commands, measurements, and system updates (as shown in Figure 3) are delivered accurately and promptly enabling essential functions such as voltage regulation, frequency control, and seamless islanding transitions,.

During island operation, microgrids must function independently, without assistance from the main grid. To maintain system integrity and extend autonomy, secure communication is essential. It enables seamless coordination

among distributed assets such as energy storage systems, grid-forming inverters, and controllable loads.

Critical functionalities like demand response, load prioritization, and predictive dispatch rely on the accuracy and timeliness of real-time data. In the absence of secure communication, optimization algorithms may receive corrupted or delayed inputs, potentially leading to inefficiencies or operational faults. Far more than a background utility, secure communication serves as the microgrid's nervous system, empowering intelligent control, defending against emerging threats, and ensuring that distributed resources operate as an integrated, resilient whole [19].

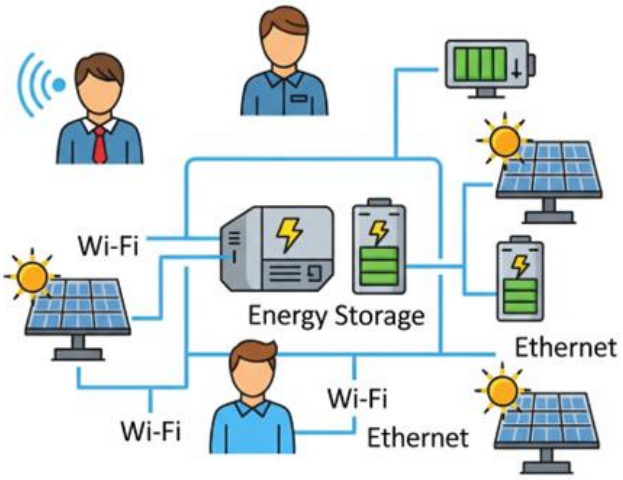

Fig. 3. Communication as the microgrid nervous system.

Recent advancements in Internet of Things (IoT) technologies have enabled the integration of heterogeneous devices and control strategies within microgrid architectures. To achieve the full functionality of the IoT, intelligent protocols/algorithms are needed for Device to Device (D2D) communications in the IoT [20]. In this work, a foundational communication framework to support interoperability, scalability, and security across diverse technologies is proposed.

The framework outlines a set of minimum requirements for communication infrastructure, including: confidentiality and data integrity, enforced through industry-standard encryption algorithms such as Advanced Encryption Standard (AES) -256 and secure transport protocols (e.g., (Datagram Transport Layer Security) [DTLS], (Transport Layer Security) [TLS]); device authentication, achieved via digital certificates or pre-shared keys; and low-latency communication, facilitated by Quality of Service (QoS) prioritization for control messages and edge computing for distributed decision-making. To enhance resilience against cyber threats, the architecture incorporates Intrusion Detection Systems (IDS) and network redundancy mechanisms.

Additionally, the proposed solution leverages a multilayered IoT communication stack, comprising: the application layer, employing lightweight messaging protocols, such as Message Queuing Telemetry Transport (MQTT) or Constrained Application Protocol (CoAP); the network layer, utilizing Internet Protocol version 6 (IPv6) with IPv6 over Low-Power Wireless Personal Area Networks (6LoWPAN) to enable efficient header compression and address allocation; and the link layer, based on low-power communication standards such as IEEE 802.15.4 or Low-Rank Adaptation (LoRa) to support constrained and distributed environments. Collectively, these components establish a secure, responsive, and extensible communication backbone suited for next generation microgrid systems [21], [22], [23], [24], [25], [26].

E. Adaptive control and Robust architecture

Microgrids function in highly dynamic environments where variables such as solar irradiance, wind conditions, load profiles, and grid connectivity can change rapidly. To sustain optimal system performance, adaptive control mechanisms adjust control parameters in real time, eliminating the need for predefined system models.

This approach enhances voltage and frequency regulation, particularly during critical transitions between grid-tied and islanded operation. A key example is adaptive droop control, which achieves more balanced current sharing and improved bus voltage stability compared to static control schemes [19].

Unlike conventional controllers that depend on accurate, fixed system representations, adaptive control accommodates incomplete or fluctuating system data, making it especially effective in settings with plug-and-play DERs or continuously evolving network topologies.

This work suggests that a suite of Machine Learning (ML) techniques can be designed to enhance microgrid intelligence across forecasting, control, protection, and cybersecurity domains.

Generation and Load Forecasting leverages Long Short-Term Memory (LSTM) networks to predict renewable energy generation and load demand. The models use inputs such as weather data, historical generation and consumption patterns, and temporal factors (e.g., time of day) to improve forecasting accuracy and enable more informed operational decisions.

Optimal Power Dispatch is approached through Reinforcement Learning (RL), where a trained agent optimizes Energy Storage System (ESS) charging and discharging, along with load control actions, to minimize costs and load shedding. The agent observes system states, including State of Charge (SOC), current generation, load, and time; and executes actions involving ESS power and load control signals. The reward function penalizes a combination of shedding cost, power imbalance, and ESS degradation, driving the agent toward efficient and resilient dispatch strategies.

To ensure timely fault mitigation, a Fast Protection System is also suggested to enhance response speed and reduce system vulnerability to electrical disturbances or equipment failures.

For phase balancing, clustering algorithms such as k-means are used to group loads with similar demand patterns and strategically assign them across phases to improve balance. Reinforcement Learning can also be deployed to enable real-time phase switching decisions, allowing adaptive control based on evolving operational conditions.

Finally, Anomaly Detection in Communication can be addressed using unsupervised learning methods like

autoencoders or isolation forests. These algorithms identify abnormal traffic patterns that may indicate cyber-attacks or system faults, enhancing the microgrid's security posture and operational reliability.

III. Case Study and Analysis

This section presents a simple test case based on the IEEE 8500 node test system where a microgrid is formed. In this microgrid, there are 50 residential customers (2.5 kW nominal each, power factor 0.95 – summer load assumed) plus an commercial customer (50 kVA, power factor 0.9). In this microgrid, the residential customers have rooftop solar panels installed. Their installations vary between 1.5 (60% of customers) and 3 kW (40% customers), allowing customers to supply their own demand and some of them being able to deliver a few kW to the grid.

The commercial customer is supported by its own installation of rooftop solar (100 kW) and a battery energy storage systems with a capacity of 2 MWh with an interfacing inverter of 250 kVA. The system schematic is shown in Figure 4.

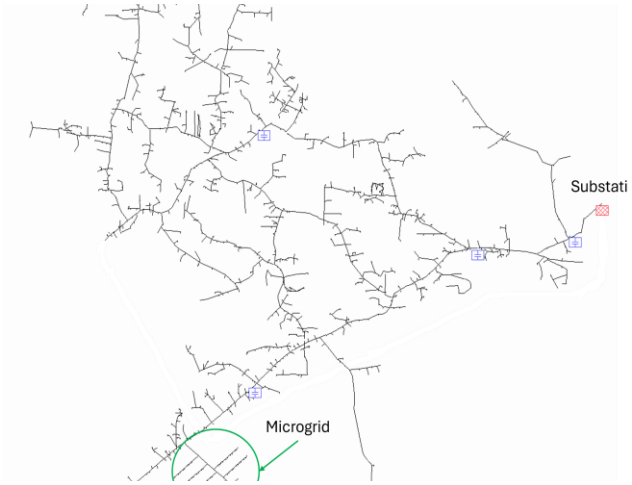

Fig. 4. IEEE 8500 nodes test system including microgrid.

After an outage event, the microgrid separates from the grid through the recloser installed at the edge of the microgrid. Once this occurs, the adaptive control determines the SoC of the Battery Energy Storage System (BESS) installed at the commercial customer. Simultaneously, it disconnects all the solar PV delivering power to the microgrid.

Once the microgrid is OFF, the adaptive algorithm based on the SoC uses the BESS as GFM for referencing the microgrid. Once the BESS inverter is connected in GFM configuration the black start operation begins, as shown in Figure 5. The voltage increase at the BESS point of connection, as shown in Figure 5.

During the initial 60 milliseconds of operation, photovoltaic (PV) systems remain intentionally disconnected, allowing system voltage to stabilize within acceptable limits. This delay is essential for enabling Grid Following (GFL) devices to synchronize with the GFM inverters, thereby preventing faults or voltage oscillations that could compromise microgrid stability.

The current profile at the BESS is illustrated in Figure 6, highlighting a reduction in delivered current as GFL devices (solar PV systems) begin to contribute power to the microgrid. This interaction supports the overall demand and effectively extends the operational capacity of the BESS. The coordination between GFM and GFL components is managed by the adaptive control system, ensuring seamless integration and balanced power sharing.

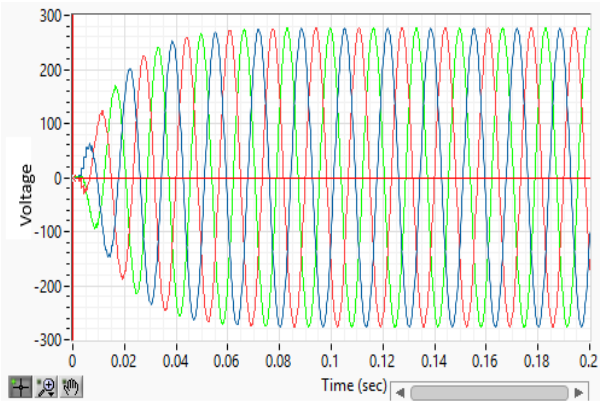

Fig. 5. Voltage during black start led by BESS.

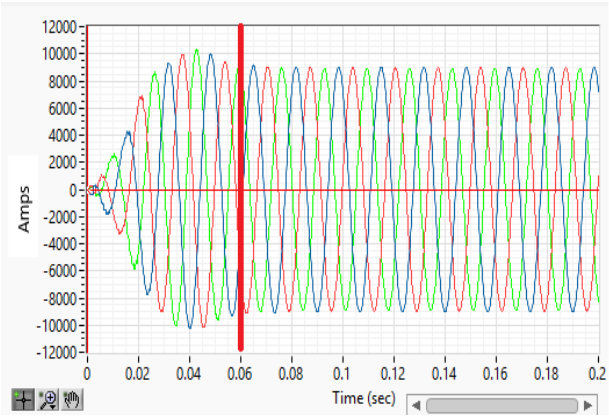

Fig. 6. BESS current during the power restoration.

IV. Conclusion and Future Work

This paper presented a microgrid architecture anchored by three resilience-enhancing pillars: (1) robust protection and power quality, (2) high-speed, reliable communication infrastructure, and (3) Machine Learning (ML)-driven control and management. Each pillar was introduced through its operational goals and technical contributions. A simulated test case illustrating the benefits of the proposed framework was briefly presented, highlighting the energy interactions occurring during the microgrid islanding and later black start.

The Future work will entail quantitative results on D2D Quality of Supply (QoS), using formal optimization models to ensure chance-constrained guarantees in the network. Other goals, such as current imbalance, secure communications will be discussed in further publications.

 ISBN: 978-1-68558-288-3

REFERENCES

[1] T. Moss et al., "The Modern Energy Minimum: The case for a new global electricity consumption threshold," The Rockefeller foundation, p. 15, Energy growth gub, Washington D.C. 2020.

[2] "IEEE Guide for Electric Power Distribution Reliability Indices," IEEE Std 1366-2012 (Revision of IEEE Std 1366-2003), pp. 1-43, 2012, doi: 10.1109/IEEESTD.2012.6209381.

[3] I. Abdulhadi, F. Coffele, A. Dysko, C. Booth, and G. Burt, "Adaptive protection architecture for the smart grid," in 2nd IEEE PES International Conference and Exhibition on Innovative Smart Grid Technologies (ISGT Europe), 2011, 5-7 Dec. 2011, pp. 1-8, doi: 10.1109/ISGTEurope.2011.6162781.

[4] M. H. J. Bollen, "Voltage sags: effects, mitigation and prediction," Power Engineering Journal, vol. 10, pp. 129-135, 1996, doi: 10.1049/pe:19960304.

[5] C. Insight, "Electric Reliability and Power System Resilience," Congressional Research Service, Online, 2018. [Online, August, 2025]. Available: https://www.everycrsreport.com/files/20180502_IN10895_b74bbaf13d1c87cf3bcd377022a1596667834782.pdf.

[6] M. Panteli, P. Mancarella, D. N. Trakas, E. Kyriakides, and N. D. Hatziargyriou, "Metrics and Quantification of Operational and Infrastructure Resilience in Power Systems," IEEE Transactions on Power Systems, vol. 32, no. 6, pp. 4732-4742, 2017, doi: 10.1109/TPWRS.2017.2664141.

[7] M. Mahzarnia, M. P. Moghaddam, P. T. Baboli, and P. Siano, "A Review of the Measures to Enhance Power Systems Resilience," IEEE Systems Journal, vol. 14, no. 3, pp. 4059-4070, 2020, doi: 10.1109/JSYST.2020.2965993.

[8] M. Panteli, C. Pickering, S. Wilkinson, R. Dawson, and P. Mancarella, "Power System Resilience to Extreme Weather: Fragility Modeling, Probabilistic Impact Assessment, and Adaptation Measures," IEEE Transactions on Power Systems, vol. 32, no. 5, pp. 3747-3757, 2017, doi: 10.1109/TPWRS.2016.2641463.

[9] L. Rodriguez-Garcia, M. M. Hosseini, T. M. Mosier and M. Parvania, "Resilience Analytics for Interdependent Power and Water Distribution Systems," in IEEE Transactions on Power Systems, vol. 37, no. 6, pp. 4244-4257, Nov. 2022, doi: 10.1109/TPWRS.2022.3149463.

[10] "Grid Forming Inverters: EPRI Tutorial (2021)," EPRI, Palo Alto, CA, 2021, 3002021722.

[11] J. Matevosyan et al., "Grid-Forming Inverters: Are They the Key for High Renewable Penetration?," IEEE Power and Energy Magazine, vol. 17, no. 6, pp. 89-98, 2019, doi: 10.1109/MPE.2019.2933072.

[12] Y. Lin et al., "Research Roadmap on Grid-Forming Inverters," NREL, Online, 2020. [Online, August, 2025]. Available: https://www.nrel.gov/docs/fy21osti/73476.pdf

[13] D. Montenegro, A. O'Connell, and J. Taylor, "Effects of demand side management programs in modern distribution planning - challenges and opportunities," IET Conference Proceedings, CP823 2023 (6), 3977-3981

[14] D. Montenegro, A. O. Connell, J. Deboever, and J. Taylor, "Localized Yearlong Power Flow Estimation Based on Limited Data Using Adaptive Filtering," in 2021 IEEE Power & Energy Society General Meeting (PESGM), 26-29 July 2021 2021, pp. 01-05, doi: 10.1109/PESGM46819.2021.9637965.

[15] T. D. Mai, T. Verschelde, and J. Driesen, "Comparative study of current redistributor's topologies for mitigating unbalanced currents in bipolar DC microgrids," in 2017 IEEE Second International Conference on DC Microgrids (ICDCM), 27-29 June 2017 2017, pp. 242-247, doi: 10.1109/ICDCM.2017.8001051.

[16] M. H. Karimi, S. A. Taher, Z. D. Arani, and J. M. Guerrero, "Imbalance Power Sharing Improvement in Autonomous Microgrids Consisting of Grid-Feeding and Grid-Supporting Inverters," in 7th Iran Wind Energy Conference (IWEC2021), 17-18 May 2021 2021, pp. 1-6, doi: 10.1109/IWEC52400.2021.9466976.

[17] H. J. Chiu, T. H. Wang, L. W. Lin, and Y. K. Lo, "Current Imbalance Elimination for a Three-Phase Three-Switch PFC Converter," IEEE Transactions on Power Electronics, vol. 23, no. 2, pp. 1020-1022, 2008, doi: 10.1109/TPEL.2007.917958.

[18] E. D. Ayele, J. F. Gonzalez, and W. B. Teeuw, "Enhancing Cybersecurity in Distributed Microgrids: A Review of Communication Protocols and Standards," Sensors, vol. 24, no. 3, 2024. doi: 10.3390/s24030854.

[19] T. V. Vu, D. Perkins, F. Diaz, D. Gonsoulin, C. S. Edrington, and T. El-Mezyani, "Robust adaptive droop control for DC microgrids," Electric Power Systems Research, vol. 146, pp. 95-106, 2017/05/01/ 2017, doi: https://doi.org/10.1016/j.epsr.2017.01.021

[20] O. Bello and S. Zeadally, "Intelligent Device-to-Device Communication in the Internet of Things," in *IEEE Systems Journal*, vol. 10, no. 3, pp. 1172-1182, Jan. 2014, doi: 10.1109/JSYST.2014.2298837.

[21] U. Tariq, I. Ahmed, A.K. Bashir, and K. Shaukat, "A Critical Cybersecurity Analysis and Future Research Directions for the Internet of Things: A Comprehensive Review," *Sensors*, *23*(8), 4117, 2023. https://doi.org/10.3390/s23084117

[22] B. Zhang, Z. Chen and A. M. Y. M. Ghias, "Deep Reinforcement Learning-based Energy Management Strategy for a Microgrid with Flexible Loads," *2023 International Conference on Power Energy Systems and Applications (ICoPESA)*, Nanjing, China, 2023, pp. 187-191, doi: 10.1109/ICoPESA56898.2023.10141490.

[23] M. S. Hossain and H. Mahmood, "Intelligent Energy Management of a Microgrid Using Reinforcement Learning," *2024 IEEE Power & Energy Society General Meeting (PESGM)*, Seattle, WA, USA, 2024, pp. 1-5, doi: 10.1109/PESGM51994.2024.10689163.

[24] A. Charalambous, L. Hadjidemetriou, L. Zacharia, A.D. Bintoudi, A.C. Tsolakis, D. Tzovaras, and E. Kyriakides, "Phase Balancing and Reactive Power Support Services for Microgrids,"*Applied Sciences*, vol. 9 issue 23, pp. 5067, 2019. https://doi.org/10.3390/app9235067

[25] K. Kimani, V. Oduol and K. Langat, "Cyber security challenges for IoT-based smart grid networks," International Journal of Critical Infrastructure Protection, Volume 25, Pages 36-49, 2019, ISSN 1874-5482, https://doi.org/10.1016/j.ijcip.2019.01.001.

[26] R. C. Borges Hink, J. M. Beaver, M. A. Buckner, T. Morris, U. Adhikari and S. Pan, "Machine learning for power system disturbance and cyber-attack discrimination," 2014 7th International Symposium on Resilient Control Systems (ISRCS), Denver, CO, USA, 2014, pp. 1-8, doi: 10.1109/ISRCS.2014.6900095.

Open Source Real-Time Automatic Modulation Classification with Deep Learning for Internet of Things Devices

Simon Boka
Tickle College of Engineering
University of Tennessee
Knoxville, U.S.A
email: sboka@vols.utk.edu

Oladayo Bello
College of Engineering | New Mexico State University
Cape Town, South Africa | Las Cruces, U.S.A
email: oladayo@ieee.org

Innocent Davidson
Cape Peninsula University of Technology
Cape Town, South Africa
email: davidsoni@cput.ac.za

***Abstract*— Deep Learning (DL) has redefined Automatic Modulation Classification (AMC) by replacing traditional hand-engineered features with end-to-end neural networks that process raw signal data, thus demonstrating high accuracy at moderate-to-high Signal-to-Noise Ratios (SNRs). While contemporary convolutional and hybrid recurrent network architectures achieve excellent performance, they often incur significant computational costs that hinder deployment on resource-constrained Internet of Things (IoT) edge devices. To address this challenge, this work proposes and presents a low-cost, open-source radio platform that performs signal acquisition and utilizes vector extensions for accelerated inference. The platform integrates a commodity Realtek Software-Defined Radio (RTL-SDR) with Reduced Instruction Set Computer – Five (RISC-V) processors. The workflow methodology for the proposed approach is a reproducible, end-to-end pipeline for deploying signal classification models on resource-constrained devices in IoT networks. The pipeline's primary strength is its deterministic dataset assembly. The workflows process establishes a coherent baseline for embedded classification under strict memory and processing power constraints typical in IoT devices.**

Keywords-Automatic modulation classification; Signal to noise ratio; RISC-V; interference.

I. INTRODUCTION

Deep learning has reshaped AMC by replacing hand-engineered features with end-to-end models that operate directly on raw In-phase and Quadrature (I/Q) sequences, that achieve strong performance at moderate-to-high SNRs. Convolutional Neural Networks (CNNs) and hybrid Convolutional Neural Network–Recurrent Neural Network (CNN–RNN) models trained on datasets such as RadioML have shown much higher accuracy, often above 90–98% at stronger SNRs. However, these models come with significant computational costs during inference, making efficient IoT edge device deployment a big challenge. Though CNN variations and spectrogram-based techniques are continually introduced, showing the clear shift toward deep learning. Yet, persistent issues remain, including performance drops at low SNR, difficulty generalizing beyond synthetic datasets, and the need to sustain real-time processing under hardware limits [1],[2].

Open RISC-V platforms with the RISC-V Vector (RVV) 1.0 vector extension help accelerate vector-heavy signal-processing and inference workloads central to intelligent radio. The acceleration is done via RVV's Vector-Length Agnostic (VLA) programming model, flexible register grouping, and support for mixed-precision arithmetic. These processes enable scalable Single Instruction, Multiple Data (SIMD) style parallelism tuned from embedded to High-Performance Computing (HPC) class implementations. RISC-V vector cores illustrate how RVV-backed designs pair a scalar pipeline with a decoupled vector unit and high-throughput memory subsystems. The pairing facilitates efficient IoT edge device inference with publicly documented configurations touting 512-bit vector registers, BFloat16 (BF16)/16-bit Floating Point (FP16)/8-bit Integer (INT8) support, and Machine Learning (ML) oriented instruction extensions for neural kernels and matrix operations [3],[4],[5],[6],[7],[8].

The RTL-SDR, which is a USB Software-Defined Radio (SDR) derived from Digital Video Broadcasting – Terrestrial (DVB-T) tuner chipsets, provides wide coverage and stable sample rates up to roughly 2.56 Mega Samples per second (MS/s) for reliable demodulation. The wide coverage provided is commonly between 24 MHz and 1766 MHz with popular tuners. These attributes make the RTL-SDR a practical, inexpensive front end for collecting real I/Q datasets to complement synthetic corpora during development and testing. As a commodity device with 8-bit Analog-to-Digital Converter (ADC) samples and ubiquitous host support, it enables rapid, repeatable data capture across bands of interest for model pre-training, augmentation, and validation. These enable it to keep total system cost low enough and allow it to scale benchtop experiments to distributed field measurements [2],[9].

Therefore, this work proposes and presents a new approach for real time automatic modulation classification using open-source platforms. The method utilizes inexpensive RTL-SDR USB dongles for capturing signals and RISC-V vector chips for fast speed running of Artificial Intelligence (AI) models on miniature, power-constraint devices. The contribution and significance of this approach is fourfold. First, it equips IoT nodes and gateways with on-device spectrum intelligence. Such capability allows distributed IoT devices to monitor, classify, and react to the radio frequency environment in real time without relying on centralized cloud processing. Second, it shortens the path from simulated data to real over-the-air recordings, by improving real-time speed and reliability for AMC.

Particularly, this reduction in time supports IoT use cases such as local interference detection on smart-city lampposts, factory floor coexistence monitoring, and edge device anomaly alert transmission without the need for constant cloud backhaul. Third, implementing AMC at the ultra-edge also facilitates the adaptation of radio in situ for IoT device deployments. For example, selecting robust modulation techniques under congestion, flagging unauthorized emitters near industrial assets, or triaging spectrum events in environmental sensor networks are possible. All these capabilities reduce latency, bandwidth, and power while maintaining service quality. Lastly, the approach makes wide-area spectrum monitoring become feasible due to the adoption of distributed receivers on battery-powered or solar-powered IoT gateways. These gateways classify signals locally and share only compact summaries, in order to improve scalability and privacy while preserving situational awareness. IoT applications that benefit from this capability include utility metering, telehealth, telemetry backhaul, and campus-scale asset tracking.

Overall, the proposed approach makes spectrum intelligence more accessible by pairing modern deep learning with vectorized execution on widely available RISC-V hardware. Both concepts have been explored separately, but until now, have not been paired together as explored in this work. The methodology leverages the growing adoption of RISC-V in IoT edge devices due to cost, openness, and efficiency. Specifically, the aforementioned capabilities support IoT deployments in smart cities, industrial environments, and environmental monitoring sensor networks. They facilitate localized decision-making, latency reduction, bandwidth conservation and network reliability.

The remainder of this paper is organized as follows: Section II surveys and motivates the selection of hardware platforms. A comparison of RISC-V compute modules and SDR front-ends is given to highlight trade-offs in cost, performance, and suitability for edge deployment. Section III presents the end-to-end workflow of the proposed approach, which includes data collection, signal preprocessing, spectrogram-based CNN training, and compilation to kmodel for execution on constrained K210 microcontrollers. Section IV outlines directions for future work, including implementation in advanced and alternative architectures, utilizing expanded over-the-air datasets, and signal intelligence testing in broader applications. Section V concludes the paper by highlighting the work's contributions to accessible edge spectrum intelligence for typical miniature IoT devices.

II. HARDWARE SURVEY AND SELECTION

There are potentially different types of hardware that can be used as the platform for this work. Thus, it is important to evaluate candidate platforms along both performance and integration dimensions. For an IoT-oriented pipeline, cost, power consumption, and form factor are just as critical as raw computational throughput. Accordingly, two categories of hardware are reviewed which are RISC-V and software-defined radio (SDR). RISC-V compute platforms are capable of running machine learning inference at the edge, while SDRs are front-ends for signal capture.

A. RISC-V compute platforms

These are low-cost platforms that facilitate efficient on-device inference for edge Digital Signal Processing (DSP) classification. Their key differentiators include the CPU microarchitecture, availability of vector or Neural Processing Unit (NPU) acceleration, memory capacity, and indicative pricing for Bill Of Materials (BOM) planning. These platforms are particularly well-suited for IoT nodes because they balance affordability with power efficiency, making it feasible to deploy spectrum-aware intelligence across a large number of distributed IoT devices. By handling feature extraction and inference locally, such platforms reduce the need for continuous backhaul to the cloud, improving both scalability and responsiveness. The results of this survey are summarized in Table I.

B. SDR front-ends (RX/TX)

On the RF side, SDR front-ends were surveyed to identify capture devices that complement lightweight RISC-V compute platforms. Available SDRs span ultra-low-cost USB dongles through to higher-end lab-grade radios. Selection criteria included frequency coverage, converter depth and sampling rate, frequency stability, front-end filtering, duplex capability (receive-only or full transmit/receive), and cost trade-offs for system integration. For IoT deployments, receive-only devices often suffice, since the primary task is passive spectrum monitoring and classification rather than active transmission. Low-cost SDRs with stable frequency control and sufficient bandwidth can therefore enable practical large-scale sensing deployments while keeping per-node costs minimal. Table II compares candidate SDR devices.

C. Selection rationale

For cost-effective, edge-deployed classification, the MaixCAM provides enough integer SIMD and a small NPU. These features accelerate lightweight DSP and inference under tight power and memory budgets, while maintaining a compact BOM and integrated camera-oriented I/O for data capture. The RTL-SDR Blog V4 pairs well by offering stable frequency control, improved high frequency performance, and integrated filtering at a fraction of the cost of wideband Transmit/Receive (TX/RX) radios whose transmit capability is unnecessary for receive-only classification pipelines. Together, these devices can be used to create a compact, IoT-ready sensing node capable of autonomous spectrum monitoring, which is critical for distributed edge applications where network connectivity may be intermittent or bandwidth-limited. IoT application examples include Health Internet of Things where devices are miniature and resource constrained.

TABLE I - COMPARISON OF RISC-V COMPUTE PLATFORMS

Device	SoC / cores	Vector / NPU	RAM	Storage	I/O highlights	Notes
Sipeed MaixCAM	Sophgo SG2002, dual T-Head C906 (1.0 GHz + 0.7 GHz)	Legacy RVV 0.7.x; 1 TOPS NPU	256 MB DDR3	microSD	MIPI CSI, DVP cam, USB-C	Compact module for edge vision/DSP; selected compute node
StarFive VisionFive 2	StarFive JH7110, quad SiFive U74 (up to 1.5 GHz)	No RVV; RV64GC	2–8 GB LPDDR4	microSD	GbE, HDMI, M.2 (PCIe 2.0)	Mature RISC-V SBC with broad Linux support
Milk-V Duo S	Sophgo SG2000, dual C906 + 1× Cortex-A53	Legacy RVV 0.7.x; vendor NPU	512 MB SIP DRAM	microSD	MIPI CSI/DSI, USB	Tiny hybrid RISC-V + ARM for I/O flexibility
Sipeed LicheePi 4A	T-Head TH1520, quad C910	Vendor vector ext; NPU present	4–16 GB LPDDR4	microSD/eMMC	PCIe 3.0, HDMI, MIPI	Higher-end RISC-V SBC for heavier workloads
Milk-V Mars	T-Head TH1520, quad C910	Vendor vector ext; NPU present	4–16 GB	microSD/eMMC	PCIe, HDMI, MIPI	Dev board variant around TH1520
Pine64 Star64	StarFive JH7110, quad U74	No RVV; RV64GC	4–8 GB LPDDR4	microSD	GbE, PCIe, HDMI	JH7110 platform in Pine64 ecosystem
Banana Pi BPI-F3	SpacemiT K1 (multi-core RISC-V)	Vendor vector/NPU (SoC-dependent)	up to 8 GB	microSD/eMMC	GbE, PCIe, HDMI	Newer RISC-V SBC line; specs evolving
MangoPi MQ-Pro (D1)	Allwinner D1, single XuanTie C906 (~1 GHz)	Legacy vector ext	512 MB DDR3	microSD	GPIO, USB OTG	Ultra-low-cost entry RISC-V Linux
HiFive Unmatched	SiFive FU740 (quad U74 + S7)	No RVV; RV64GC	8 GB DDR4	M.2 NVMe	PCIe x8 (x4 elec), GbE	High-end dev board; limited availability
BeagleV Ahead	T-Head TH1520	Vendor vector ext; NPU present	4–8 GB	microSD/eMMC	PCIe, HDMI, MIPI	Community SBC with TH1520
StarFive VisionFive (v1)	StarFive JH7100 (dual U74)	No RVV	up to 8 GB	microSD	GbE, HDMI	First-gen predecessor to VF2
Milk-V Meles	CVITEK CV1800B (RISC-V C906)	Vendor vector; ISP/NPU (SoC)	512 MB	microSD	Dual MIPI CSI, Ethernet	Camera-centric edge module

TABLE II - COMPARISON OF SOFTWARE DEFINED RADIO FRONT-ENDS

Device	Frequency coverage	ADC / sample rate	TCXO	Preselection / filters	Notes
RTL-SDR Blog V4	~0.5–30 MHz (direct) + ~24/28–1766 MHz	8-bit (RTL2832U), up to ~2.4–3.2 Msps	1 ppm	Improved HF path, FM notch	Bias-T; RX only; Selected RF frontend; stable, low cost
RTL-SDR Blog V3	~0.5–30 MHz (direct) + 24–1766 MHz	8-bit, up to ~2.4 Msps	1 ppm	Basic, optional FM notch	Bias-T; RX only; Proven baseline dongle
HackRF One	~1 MHz–6 GHz	8-bit, up to 20 Msps	~20 ppm	Minimal onboard filtering	No Bias-T; Half-duplex TX/RX Wideband, experimental TX
Airspy Mini	~24–1800 MHz	12-bit, up to 6–10 Msps	0.5 ppm	Moderate front-end filtering	No Bias-T; RX only; High dynamic range for VHF/UHF
Airspy R2	~24–1800 MHz	12-bit, up to 10 Msps	0.5 ppm	Improved linearity/filtering	No Bias-T; RX only; Performance-oriented dongle
Airspy HF+ Discovery	~0.5 kHz–31 MHz + 60–260 MHz	16-bit MF stages, high effective ENOB	0.5 ppm	Strong HF preselection	No Bias-T; RX only; Elite HF sensitivity and selectivity
SDRplay RSP1A	~1 kHz–2 GHz	12–14-bit, up to 10 Msps	0.5 ppm	Multi-band preselection	No Bias-T; RX only; Versatile coverage with filtering
SDRplay RSPdx	~1 kHz–2 GHz	12–14-bit, up to 10 Msps	0.5 ppm	Enhanced HF front-end	No Bias-T; RX only; Improved LF/MF/HF robustness
SDRplay RSPduo	~1 kHz–2 GHz (dual tuners)	12–14-bit, up to 10 Msps	0.5 ppm	Preselection per tuner	No Bias-T; RX only (dual coherent) Diversity/DF use cases
LimeSDR Mini 2.0	~10 MHz–3.5 GHz	12-bit, up to ~30.72 Msps	1 ppm	Basic, external filtering advised	No Bias-T; Full-duplex Compact TX/RX platform
ADALM-Pluto (PlutoSDR)	~325 MHz–3.8 GHz (70 MHz–6 GHz mod)	12-bit, up to ~61.44 Msps RX	1 ppm	Minimal onboard filtering	No Bias-T; Full-duplex Flexible teaching/experimental SDR
USRP B200mini-i	~70 MHz–6 GHz	12-bit, up to ~56 Msps	2.5 ppm OCXO (-i)	External filtering recommended	No Bias-T; Full-duplex; Lab-grade, UHD ecosystem
KrakenSDR (coherent)	~24–1766 MHz (5 coherent tuners)	8-bit, per-tuner ~2.4 Msps	0.5–1 ppm	FM notch options	Bias-T; RX only; DoA/beamforming with 5-way phase-coherence
KerberosSDR (coherent)	~24–1766 MHz (4 tuners)	8-bit, per-tuner ~2.4 Msps	0.5–1 ppm	Optional filtering	No Bias-T; RX only; Earlier 4-tuner coherent array

 ISBN: 978-1-68558-288-3

III. WORKFLOW- METHODOLOGY

In this section, the proof-of-concept implementation of the proposed approach is presented. The flowchart in Figure 1 illustrates the workflow methodology. The end-to-end pipeline runs on a desktop host and outputs a compact kmodel artifact for the K210's Kendryte Processing Unit (KPU). It starts with raw I/Q captures and finishes with a compiled model that adheres to the memory and operator constraints documented for Maix/MaixPy deployments on the K210. Although development occurs on a desktop host, the resulting models are fully compatible with IoT edge devices. Consequently, the models facilitate the implementation of autonomous spectrum classification in IoT devices deployed in remote or power-constrained environments. The workflow uses SDR# as shown in Figure 2 with an RTL-SDR to collect labeled I/Q recordings, Scientific Python (SciPy) to generate time-frequency spectrograms from complex baseband arrays, TensorFlow/Keras to train a CNN on those images, and TensorFlow Lite plus *nncase*/KPU tooling to export and compile an embedded-ready kmodel which is standard practice supported by [1],[2],[3],[4],[5].

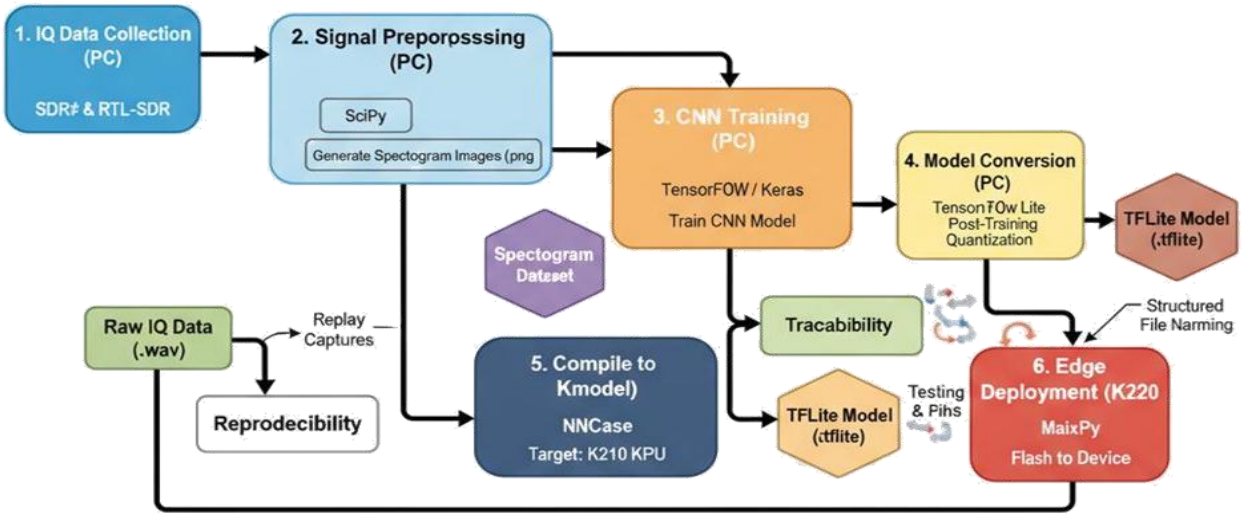

Figure 1. Workflow Approach.

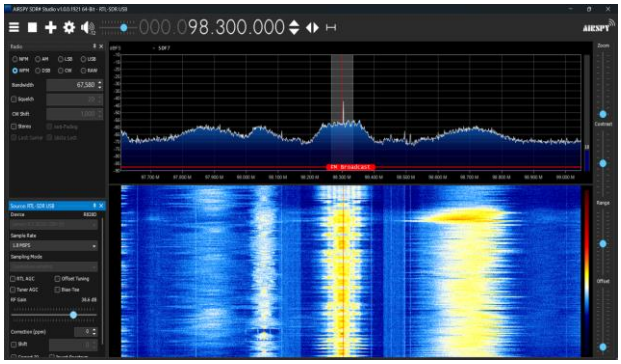

Figure 2. SDR Sharp (SDR#) Interface.

A. *Data collection on PC*

During this process, raw complex baseband streams are recorded interactively in SDR# via the Recording tab, which supports baseband I/Q capture to Waveform Audio File Format (WAV). The WAV file is used for later offline processing and precise replay in SDR# to enable deterministic dataset curation for downstream steps. To assemble a minimal labeled corpus aligned with target classes, one can capture a strong local FM broadcast segment, record a National Oceanic and Atmospheric Administration (NOAA) Weather Radio transmission within the 162.400–162.550 MHz Very High Frequency (VHF) allocation. Then gather a clip from an unoccupied channel to form a noise baseline, with files organized into class-named directories and metadata preserved in filenames to aid traceability. Alternate SDR ecosystems and guidance on I/Q data handling reinforce the objective of producing contiguous, timestamped baseband data suitable for reproducible post-processing and later validation, independent of the specific GUI tool used [1],[6],[7].

B. Signal preprocessing and spectrograms

At this stage, complex I/Q arrays are transformed into two-dimensional time–frequency images using a Short-Time Fourier Transform (STFT) spectrogram. The SciPy's documentation specifies outputs as frequency bins, time-frames, and a non-negative spectral representation suitable for learning and visualization. A practical implementation loads each I/Q recording, computes spectrograms with fixed window and overlap for consistent resolution. Log scaling and normalization are performed, and standardized images are written to per-class folders. These align with established spectrogram practice and tutorials [4],[8].

C. CNN training on spectrograms

During this task, with a directory of labeled spectrogram images, a compact CNN is defined and trained using Keras. The task follows TensorFlow's canonical model creation and training patterns that interoperate cleanly with subsequent TensorFlow Lite conversion. In addition, the training routine uses consistent image dimensions and a simple stack of convolutional and pooling layers ending in a softmax head. Then, it saves a validated host-side model artifact before any edge-oriented conversion, which cleanly separates algorithm development from deployment concerns as suggested [5].

D. Model conversion to TensorFlow Lite

After host-side validation, the Keras model is converted into a *.tflite* FlatBuffer using the TensorFlow Lite Converter API, which converts Keras models to byte buffer for exporting and saving the result. To meet embedded constraints, post-training quantization can be enabled during conversion to reduce model size and improve inference efficiency. These steps establish both float and quantized TFLite variants for rapid A/B checks prior to device-specific compilation as proven in [5],[9].

E. Compilation to kmodel for K210

At this level, the Kendryte K210's KPU executes models in the vendor-specific kmodel format generated by *nncase*. The Sipeed's Maix/MaixPy documentation outlines KPU loading modes, typical memory ceilings by firmware variant, and the expectations for kmodel artifacts compiled from TFLite. In practice, the TFLite model is compiled with *nncase* to produce a kmodel and then verified against the host TFLite baseline on representative inputs. The process ensures operator support and adherence to KPU memory limits described in Sipeed guidance for C software development kit (C SDK) or MaixPy runtimes. Community implementation notes also emphasize using a compatible *nncase* release for K210 workflows and cross-checking inference numerics between TFLite and kmodel before flashing or SD-card deployment [2],[3],[10].

F. Setting up the edge device to run the converted model

The sub-steps for this task are 1) on the edge device, a suitable MaixPy firmware or a C SDK–based firmware is installed, 2) on either the SD card or in on-board flash, the kmodel is provisioned, as supported by MaixPy's KPU loader and the Kendryte flashing utility, 3) loading models are placed on an SD card with the MaixPy KPU API using a filesystem path; models flashed to a designated offset can be loaded from flash, with Sipeed documentation to describe memory limits and firmware variants, 4) modules are deployed, which typically involves flashing firmware with kflash.py or its GUI, copying the kmodel to SD or embedding it in flash, and writing a minimal runtime that initializes the sensor or input pipeline, 5) frames are pre-processed to the model's input shape and data layout; this invokes the KPU inference, and emits results over serial, display, or General-Purpose Input/Output (GPIO) as applicable to scenarios discussed in [2],[3],[11].

G. Reproducibility considerations

Replaying recordings to validate labeling and preprocessing is facilitated by SDR#'s ability to open baseband I/Q WAV files in order to enable confirmation of tuned stations, SNR, and channel occupancy prior to batch spectrogram generation and training. Moreover, informative file naming, directory schemes, and general I/Q data management best practices support traceability across data collection, preprocessing, and inference stages without changing the core methods described here. Retaining both original I/Q archives and derived spectrograms ensures experiments can be reconstructed or extended. The maintenance of float and quantized TFLite baselines provides stable references for evaluating compiler effects prior to kmodel flashing and device trials. By producing compact, portable models, this workflow allows IoT devices to perform on-site classification, anomaly detection, and local interference management while maintaining minimal power and memory usage.

IV. Conclusion And Future Work

This work has aimed to present an end-to-end, reproducible pipeline for converting raw SDR# baseband recordings into an optimized model for inference on resource-constrained Kendryte K210 microcontrollers. The methodology integrates four key stages and a total of six tasks including the key stages. These key stages are standardized spectrogram preprocessing, compact CNN training, post-training quantization via TensorFlow Lite, and final compilation using *nncase*. The resulting workflow establishes a tractable and verifiable pathway from RF signal acquisition to on-device classification, explicitly addressing the memory and operator limitations inherent to edge hardware. This approach makes spectrum intelligence accessible to a wide range of IoT deployments, enhances local decision-making, reduces latency, and conserves bandwidth. However, areas for future work include implementing formal dataset quality assurance beyond manual replay, empirically justifying spectrogram parameters, enforcing numerical parity between the TFLite and kmodel outputs, and conducting instrumented on-device profiling to measure real-world performance.

The next steps of this proof of concept will focus on the following activities:

- **Porting to a Ratified RVV 1.0 Platform:** The highest priority is to migrate the entire workflow to a newer RISC-V platform that implements the fully ratified version 1.0 of the RVV Extension. Migration will enable a quantitative analysis of the performance gains achievable using the more powerful and flexible VLA programming model rather than the legacy vector implementation used in this work. This step is crucial for demonstrating the full potential of standardized RISC-V vector processing for ML workloads. Porting to RISC-V platforms could further enhance IoT nodes by enabling larger or more sophisticated models to execute on small, distributed devices at the network edge.
- **Comparative Benchmarking:** A performance benchmark will be conducted to compare RISC-V-based platform with a low-cost edge AI accelerator, such as the Raspberry Pi with a Google Coral Tensor Processing Unit (TPU) or the NVIDIA Jetson Nano. This will provide insightful trade-offs between performance, power consumption, cost, and openness across different edge computing paradigms.
- **Advanced Model Architectures:** Models like MobileNets, SqueezeNets, quantization-aware networks could potentially improve classification accuracy on more complex modulation schemes while maintaining or even reducing inference latency. Therefore, more sophisticated and computationally efficient neural network architectures will be explored.
- **Expanded Over-the-Air (OTA) Dataset and Classification Tasks:** Training on a larger, more diverse dataset, which includes various modulation techniques like QPSK, GMSK, 16-QAM, and 64-QAM, will enhance the platform's ability to operate reliably in several RF conditions. This will improve the performance of IoT applications such as smart city infrastructure, telehealth, telemetry, and environmental monitoring.
- **Exploration of New Applications:** The validated platform serves as a foundation for exploring other signal intelligence tasks beyond AMC. The platform can also support IoT-specific applications such as autonomous anomaly detection, RF fingerprinting for device authentication, interference localization in distributed sensor networks, and automated spectrum-aware control of edge devices, or automated signal protocol identification.

ACKNOWLEDGMENT

The authors would like to thank the French South African Institute of Technology (F'SATI), Cape Peninsula University of Technology, Cape Town, South Africa for funding this research.

REFERENCES

[1] O. F. Abd-Elaziz, M. Abdalla, and R. A. Elsayed, "Deep learning-based automatic modulation classification using robust CNN architecture for cognitive radio networks," Sensors, vol. 23, art. no. 9467, Nov. 2023, doi: 10.3390/s23239467.

[2] Ashishware, "Creating a CNN to classify cats and dogs for Kendryte K210 boards," Ashishware.com, Aug. 28, 2024. [Online]. Available: https://ashishware.com/2024/08/29/k210CatDog/.

[3] Android Open Source Project, "TensorFlow Lite converter," GoogleSource, n.d. [Online]. Available: https://android.googlesource.com/platform/external/tensorflow/+/HEAD/tensorflow/lite/g3doc/convert/index.md. (Accessed: Sep. 3, 2025).

[4] RTL-SDR.com, SDRSharp User's Guide, May 7, 2018. [Online]. Available: https://www.rtl-sdr.com/sdrsharp-users-guide/.

[5] A. Frame, "SiFive intelligence X280: Optimized efficiency and control for the modern workload" [Product Brief], presented at the Reduced Instruction Set Computer-Five In Space Conference, ESA, Dec. 2022. [Online]. Available: http://microelectronics.esa.int/riscv/rvws2022/presentations/04-SiFive_Intelligence_X280_for_Space_Exploration_v2.0_Dec_22.pdf.

[6] Google AI Edge, "Convert TensorFlow models," Aug. 29, 2024. [Online]. Available: https://ai.google.dev/edge/litert/models/convert_tf.

[7] Kendryte, kflash.py: A Python-based Kendryte K210 UART ISP utility, 2018. [Online]. Available: https://github.com/kendryte/kflash.py.

[8] Osmocom, "rtl-sdr wiki," n.d. [Online]. Available: https://osmocom.org/projects/rtl-sdr/wiki. (Accessed: Sep. 3, 2025).

[9] H. Ouamna, A. Kharbouche, Z. Madini, and Y. Zouine, "Deep learning-assisted automatic modulation classification using spectrograms," Eng. Technol. Appl. Sci. Res., vol. 15, no. 1, pp. 19925–19932, Feb. 2025, doi: 10.48084/etasr.9334.

[10] Π Node, "RTL-SDR," n.d. [Online]. Available: https://p-node.org/documentation/ressources/rtl-sdr. (Accessed: Sep. 3, 2025).

[11] RISC-V International, RISC-V "V" vector extension, Version 1.0, 2021. [Online]. Available: https://github.com/riscv/riscv-v-spec/releases/tag/v1.0.

Towards Optimized Connectivity in Health Internet of Things Device-to-Device Networks

Oladayo Bello
Cape Peninsula University of Technology|
College of Engineering, New Mexico State University
Cape Town, South Africa | Las Cruces, U.S.A
Email: oladayo@ieee.org

Innocent Davidson
Cape Peninsula University of Technology
Cape Town, South Africa
Email: davidsoni@cput.ac.za

***Abstract*— The Health Internet of Things (HIoT) enables Device-to-Device (D2D) communication among heterogeneous medical devices. However, optimal D2D connectivity is challenging due to traffic demand, the inherent environmental and device constraints. Prior works have characterized HIoT networks with single objective optimization models and either simplify or ignore device and environmental constraints, thus yielding poor scalability and limited practical value. Thus, this paper casts optimal HIoT D2D connectivity as a stochastic Multi-Objective, Mixed-Function and Mixed-Constraint (MO-MF-MC) problem. An analysis of why the HIoT D2D network is fundamentally stochastic is presented. In addition, the paper presents and formalizes two views to model optimal D2D connectivity. These are the Constraint Based (CB) and the Pareto Optimal Vector (POV) perspectives. The paper supports POV as most suitable. The contributions of this paper are: (1) an analysis of the challenges of modeling optimal HIoT D2D connectivity (2) the formulation of the stochastic D2D optimal connectivity from CB and POV perspectives, (3) justification of POV modeling for optimal D2D connectivity in HIoT. This work establishes the need for the design of lightweight, scalable, and adaptive protocols for sustainable, reliable real-time and optimal connectivity in HIoT D2D networks.**

Keywords- Constraint based; Device-to-Device; Health Internet of Things; Optimization; Pareto vector.

I. INTRODUCTION

The Health Internet of Things (HIoT) connects medical sensors, wearables, clinical instruments and infrastructure for real-time patient diagnosis, treatment, and monitoring. A key enabler of the HIoT ecosystem is Device-to-Device (D2D) networks, which facilitate direct data transfer between devices, thereby reducing dependency on centralized infrastructure [1][2]. In healthcare scenarios, this is crucial because timely and reliable data transmission are essential for clinical decisions and emergency response. Thus, low latency, loss and jitter along with high data rate are required Quality of Service (QoS) [2][3]. However, HIoT D2D networks face unique challenges due to constraints imposed by their operational environment, the type of devices and traffic they support. Typically, these networks operate in Not-For-Wire (NFW) environments, which refer to any domain where wired connections are either infeasible, impractical, or undesirable. In such domains, devices exchange data by leveraging the wireless medium, which is shared, inherently unstable, and resource constrained. It is also characterized with limited bandwidth and data transmissions are prone to interference and high path loss. These conditions degrade and affect the network's performance to guarantee optimal connectivity essential for reliable communication within healthcare systems. Additionally, HIoT D2D devices are unconventional, miniature, and constrained in resources, such as computational power, memory, and battery life [4][5]. Traffic is diverse, ranging from data generated by patient monitoring, mission-critical and real-time operations to emergency alerts. These traffic streams require differentiated treatment and stringent QoS guarantees. However, the constraint imposed by devices, the unpredictability of the NFW environment coupled with the unique traffic types, introduces unpredictable conditions that cause stochastic connectivity and thus makes it difficult to guarantee QoS.

In HIoT D2D networks, connectivity implies that QoS demands by active traffic flows are simultaneously satisfied. QoS metrics include latency, jitter, throughput, and packet loss. The basic expression for connectivity is given by equation (1)

$$Connectivity \Leftrightarrow \forall i,\ f_i(x) \leq b_i \quad (1)$$

where

- i: index over all QoS metrics
- $f_i(x)$: objective function of QoS metric i.
- b_i: the bound value for QoS metric i

Equation (1) states that connectivity is achieved, if and only if (iff), all QoS metrics indexed by i satisfy their respective bound (threshold). $f_i(x)$ represents the QoS performance under a given network configuration x, while b_i denotes the required bound that must be satisfied for each metric. For example, in a healthcare scenario, latency measured using $f_{latency}(x)$ must not exceed its critical bound $b_{latency}$ and similarly, packet loss must remain below its acceptable threshold. Quantifier ∀i ensures that QoS demand is simultaneously satisfied.

Consequently, sustaining QoS in HIoT D2D network requires protocols that utilize Multi-Objective, Mixed-function, Mixed-constraint (MO-MF-MC) optimization

approach. The approach ensures that trade-offs between conflicting goals are carefully balanced. However, there is a lack of such protocols because most networking protocols were not designed to handle the multi-layer dynamics of QoS objectives, device constraints and uncertainty that exists in the NFW environments [6].

These dynamics highlight the importance of treating optimal D2D connectivity in HIoT as a stochastic MO-MF-MC problem. It is also desirable to have protocols that facilitate optimal D2D connectivity. To address these gaps, this paper focuses on:

"How optimal connectivity can be achieved despite the tradeoff that exists in meeting conflicting and stringent QoS demands of mission-critical traffic traversing the constrained HIoT D2D network operating under stochastic conditions"

The contributions of this paper are: 1) analysis of the inherent challenges for optimal connectivity and the limitations of single-objective optimization models in HIoT D2D networks 2) formulation of optimal connectivity with a stochastic MO-MF-MC model under the Constraint Based (CB) and Pareto Optimal Vector (POV) perspectives. 3) justification of POV as the perspective that best captures the realistic trade-offs among QoS metrics subject to device and environment constraints. Moreover, one of the challenges for optimal connectivity identified and introduced in this paper, is the unique characteristic of HIoT D2D traffic flow, which has been termed "Mixed-criticality, Bound-assured, Mission-synchronous" (MC-BAMS). The term is explained in Section II. Lastly, the paper provides insight into a framework to be adopted in the design of next generation communication protocols for HIoT D2D networks. The future work that builds upon this paper includes a lightweight protocol that operationalizes the POV framework. The paper's content is as follows: Section II presents the challenges for optimal D2D connectivity in HIoT, Section III discusses optimization in HIoT, Section IV presents the optimal connectivity model and Section V concludes the paper.

II. Challenges for Optimal Connectivity

Within the HIoT D2D networks, three main challenges impose the need for tailored protocols to facilitate optimal connectivity. These are operational challenges, which affect QoS performance objectives and in turn impacts connectivity. They stem from environmental and device constraints, and heterogeneity of data traffic, which are discussed as follows.

A. Not-For-Wire (NFW) Environmental conditions

The operational domain of HIoT D2D networks is often a NFW setting where links are wireless. Conditions within such settings are inherently unpredictable due to co-located medical systems, patient movement and deteriorating signal strength. These conditions introduce interference and fluctuations that cause connectivity to be stochastic thus, making QoS guarantees difficult to sustain [3][7]. While deterministic connectivity models may suffice in stable networks, the instability of NFW environmental conditions favors stochastic modeling especially in healthcare systems where millisecond delays can impact outcomes [7][8]. An example of the NFW environment is smart Intensive Care Units (ICUs) where ventilators, infusion pumps, and monitors exchange critical data simultaneously over the shared wireless spectrum. The setting reduces cable clutter and improves safety but raises signal interference risk [8]. In homecare, data generated by wearable ECG patches and implantable glucose sensors is wirelessly sent to smartphones or clouds systems and thus allow patient mobility. However, these medical devices contend with home appliance operating in the same frequency bands and patient mobility can affect link quality [9]. Mobile emergency care further highlights the stochasticity in NFW environments. The ambulances stream vital signs en route, so low latency and negligible error rates are essential for pre-arrival interventions, yet handoffs and fading continually perturb the wireless links [7][10]. HIoT D2D networks need robust, adaptive mechanisms that handle environmental variability while preserving the performance of life-critical traffic. Therefore, optimization frameworks should explicitly model NFW uncertainty and guarantee QoS bounds [7][8][9].

B. Device Constraints

In D2D networks, devices are often miniature embedded systems designed with strict size for comfort and usability requirements. Smartwatches, biosensors, and implantable medical devices prioritize patient convenience and portability but at the cost of battery capacity, memory, and processing power [11]. Limited energy prevents prolonged high data rate thus making it challenging to guarantee continuous, low-latency transmission. Memory and computational limitations further restrict the use of conventional protocols, which often require data buffering, complex computations and large memory [4]. For instance, real-time ECG monitoring generates massive data streams, but devices often lack the capacity to buffer or preprocess data locally [12]. This constraint forces reliance on lightweight, efficient communication mechanisms tailored for low-resource devices. Additionally, battery longevity is a critical factor. Many implantable wearable devices must function for months or even years without replacement and frequent recharging is impractical. Battery power constraint impacts not just transmission occurrence rate but also the complexity of protocols that can be executed.

C. Traffic Characteristics

The data traffic in HIoT D2D networks is highly heterogeneous. It includes data generated by routine updates, monitoring devices and mission-critical alerts from pacemakers. Diversity means that different traffic streams require differentiated QoS guarantees. For healthcare traffic, timeliness is as crucial as accuracy [13]. Inherently, traffic is generated in real-time and delayed data may become irrelevant, thus reducing their utility for clinical decisions. For example, a physician monitoring a remote patient's heart rhythm requires data to be streamed in near real time. A

delayed transmission of the same data will lose diagnostic value. However, while routine patient monitoring data can tolerate modest delays, mission-critical signals must be delivered with minimal latency and low jitter [13]. The diverse traffic requirements make prioritizing traffic during resource allocation difficult. High-priority emergency traffic must preempt less urgent transmissions without entirely starving background data streams, such as periodic wellness updates. Thus, the nature of traffic flow is such that they are "mixed-criticality, bound-assured, mission-synchronous" (MC-BAMS). MC-BAMS implies that "At any time, there exist diverse traffic flow with different criticality level and QoS bounds that must be simultaneously guaranteed and transmitted in a shared, unpredictable, and resource-constrained environment, where no traffic can be deferred". This traffic flow characteristic is unique to HIoT D2D networks. Consequently, designing scheduling and resource allocation protocols to facilitate fair differentiated service by supporting MC-BAMS traffic flows under device and environmental constraints remains a challenge.

D. Architectural Overview of HIoT D2D Networks

Figure 1 depicts a simplified architectural overview of the challenges for optimized connectivity in HIoT D2D networks. Typically, such networks integrate multiple types of medical and wearable devices that communicate directly without relying exclusively on centralized infrastructure. The devices include implantable sensors, wearable glucose monitors, smartwatches, infusion pumps, ventilators, and imaging systems. Each device is constrained by size, memory, computational capacity, and battery power, limiting its ability to process and transmit continuous high-volume traffic.

These limitations necessitate lightweight optimization strategies to maintain network reliability. The environment is depicted as a Not For Wire (NFW) medium, characterized by interference, unpredictability, mobility, and shared spectrum resources. Within this environment, different types of traffic coexist.

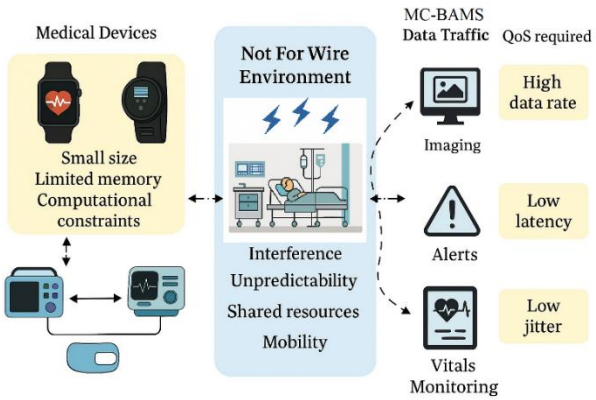

Figure 1. Architectural Challenges for Optimized Connectivity in HIoT D2D Networks.

Each traffic type has unique QoS requirements as outlined:

- Imaging data (high bandwidth, moderate latency tolerance).
- Alerts (ultra-low latency, mission-critical).
- Vitals monitoring (periodic updates, moderate QoS).

The figure illustrates how device limitations, volatile environments, and heterogeneous traffic demands combine to create optimization challenges for HIoT D2D networks.

E. Need for lightweight protocols

Commonly used standardized networking protocols were not designed for the highly unstable and constrained condition of the HIoT D2D network. Those protocols function with high signaling overhead, computational processes and memory resources, which cannot be supported by miniature medical devices. Moreover, most do not adapt to the NFW environments where connectivity is stochastic. Furthermore, these protocols do not implement mechanisms that can cater for the unique nature of MC-BAMS traffic flow in HIoT D2D. Thus, making them inadequate for mission-critical health applications where real-time medical signals require stringent QoS guarantees simultaneously. Due to these limitations, HIoT D2D networks require lightweight, adaptive protocols that will optimize and allocate resources fairly while meeting stringent QoS requirements of heterogeneous medical traffic. Such tailored protocols should be based on optimization models that will ensure that life-critical communications are reliably sustained under device, environmental and traffic requirement constraints.

III. Optimization in Health Internet of Things (HIoT)

In this section, a comparison of single and multi-objective optimization techniques for HIoT is presented. Existing approaches and gaps are also discussed.

A. Single objective vs. Mult objective Optimization

Single-objective optimization approaches focus on one metric at a time, for example, minimizing latency or maximizing data rate. They are simple, computationally less intensive and easier to interpret thus appear appealing for modeling constrained environments [14]. However, their weakness lies in oversimplification and the inability to combine multiple metrics' objectives simultaneously. For example, minimizing latency without regarding data rate forces very short traffic inter-arrival times and high scheduling frequency, which inflates protocol overhead and reduce effective data rate. Conversely, if data rate is maximized without regard for latency, large data aggregation, and buffering raise queueing delays (and jitter), thus impacting end-to-end latency.

In real-world HIoT D2D network applications, where performance dimensions are highly interdependent, single-objective approaches often fail to capture the true complexity of the problem. By contrast, multi-objective optimization models recognize trade-offs across multiple metrics [14][15]. These models enable the design of protocols that can generate sets of optimal solutions instead of committing to a single "winner." This is valuable in HIoT D2D networks, where objectives of guaranteeing multiple QoS demands such as minimizing latency, maximizing data rate and minimizing data error and loss are simultaneously critical.

While single-objective optimization provides clarity, it lacks realism for HIoT D2D applications. Multi-objective models, though more complex, provide the flexibility and adaptability needed to balance diverse, and often conflicting requirements exhibited by the MC-BAMS traffic flow in a NFW, device constrained healthcare-focused D2D networks.

B. Optimization aproaches in IoT and HIoT

Optimization in the IoT has been widely studied. Particularly, wireless sensor networks (WSNs) formed the basis of many early optimization frameworks. These networks highlighted the challenges of balancing multiple objectives such as data rate and latency. In [15] the authors provided comprehensive taxonomies of Multi-Objective Optimization (MOO) in WSNs. They examined both scalarization approaches (e.g., weighted sums) and evolutionary algorithms. Such methods laid the groundwork for extending optimization approaches into the more complex domain of HIoT.

In the HIoT context, optimization approaches have focused on areas such as task scheduling, device allocation, and network resource management. Nucci et al presented a bi-objective scheduling framework. It minimizes operational costs, maximizes quality of care simultaneously with non-dominated sorting genetic algorithm II (NSGA-II**)** heuristics. The framework takes into consideration device constraints such as compatibility, limited battery capacity, and setup overheads. This dual-focus design underscores the necessity of balancing operational efficiency with service quality in life-critical environments. Similarly [17] focused on multi-objective model for IoT application placement (MAPO) to address application placement by balancing latency, energy consumption, and operational costs. Such approach demonstrates particular relevance for medical applications, where offloading and distributed computing reduces stress on resource-constrained devices while still meeting stringent latency and reliability requirements.

More recently, adaptive strategies leveraging evolutionary and reinforcement learning have been introduced. Evolutionary Multi-Objective Optimization (EMO) techniques have shown promise in handling the scale and complexity of large HIoT deployments [18]. Furthermore, Multi-Objective Reinforcement Learning (MORL) has emerged as a dynamic solution, capable of learning adaptive trade-off policies under uncertain environments without relying on predefined training data [19]. Such methods are increasingly attractive for HIoT networks.

C. Gaps: lack of real world conditions

Despite advancements in optimization frameworks, existing models often fall short in their applicability in real-world HIoT scenarios. Many rely on idealized assumptions such as stable wireless channels, unconstrained processing power, and predictable traffic patterns. These assumptions do not reflect the stochastic nature of the NFW environment of HIoT D2D networks. Most models assume deterministic QoS guarantees by discounting the reality of fluctuating NFW settings and the constraints of miniature devices.

Scalability presents another significant challenge. While evolutionary algorithms are robust in generating optimal solutions, their computational complexity grows exponentially with the number of objectives or devices [18]. Consequently, this can create a bottleneck in large-scale deployments of real-time HIoT D2D applications, where instantaneous decision-making is critical. Optimizing all data traffic QoS demand across hundreds of devices simultaneously can exceed the practical computational capacity of even fog-enabled networks.

Moreover, most of the existing frameworks do not integrate dynamic, stochastic constraints into their models. Real-world HIoT D2D networks may be subject to fluctuating interference and unpredictable error rate, which leads to unstable connectivity for traffic flows with diverse criticality levels. Yet most optimization studies often treat constraints as static, ignoring temporal variations and unpredictability [17]. Finally, the lack of lightweight protocols derived from optimization insights remains a critical concern. Current optimization research tends to stop at theoretical modeling, without translating results into deployable protocols that resource-constrained devices can implement.

These gaps undermine the delivery of consistent QoS in real medical environments, where momentary lapses in connectivity can jeopardize patient safety. Addressing these gaps require developing scalable, adaptive, and lightweight MO-MF-MC frameworks that integrate stochastic constraints and translate directly into operational protocols suitable for HIoT D2D networks.

IV. Optimal Connectivity Model

A. Stochastic Nature of Connectivity

Connectivity is achieved in HIoT D2D networks if and only if all QoS requirements are simultaneously satisfied. This reflects the mission-critical nature of such networks, where failing in one metric implies network failure. However, the stochastic nature of connectivity, due to the NFW environmental conditions, makes it challenging to satisfy simultaneous demands. As a result, QoS metrics cannot be taken on fixed values but rather be modeled as random variables with probability distributions [20]. For example, the probability of maintaining latency below a certain threshold may vary significantly depending on interference levels, which means deterministic guarantees are impossible. Instead, probabilistic QoS guarantees, e.g., $P(\text{latency} \leqslant \tau) \geqslant 0.95$ must be incorporated into optimization formulations to account for the stochastic nature of connectivity. However, the stochasticity also complicates optimization because even when devices operate under optimal configurations, performance guarantees may not be met due to environmental variations. For example, significant loss of data may occur due to unpredictable interference bursts, even if optimal QoS targets have been initially met. Hence, connectivity optimization frameworks must be designed to adapt

dynamically to changing states while tolerating uncertainty. Stochasticity also arises from device mobility and human activity patterns. Wearable sensors and implantable devices attached to patients move unpredictably, thus making deterministic assumptions impractical. Moreover, the pattern of the MC-BAMS traffic flow validates the importance of accounting for stochasticity. Traffic flow competes for resources that must be shared fairly and optimally.

B. Network Objectives function (QoS Metrics)

The primary goal of HIoT D2D network is to ensure connectivity by guaranteeing QoS requirements of traffic flows generated for healthcare service delivery. Latency and jitter are among the most vital metrics. The transmission of ECG data during cardiac arrest must occur timely with strict bounds on delay variation. Data rate and data loss are equally important, as compromised arrival rate and integrity of medical data can lead to unintended consequences. Data rate becomes very critical when devices are streaming medical data in images and videos format. The network must balance the requirements from all traffic flows simultaneously without sacrificing any traffic demand. In summary, connectivity is achieved by simultaneously meeting multiple QoS objectives in HIoT D2D networks, which extend beyond conventional communication goals. Meeting these combined stringent performance metrics' objectives, which are imperative for timeliness and accuracy of healthcare services demands tailored frameworks.

C. Network Constraints (Device and Environmental)

While objectives define "what" should be achieved, constraints determine "how" or "if" such objectives are possible. In other words, in HIoT D2D networks, QoS metrics establish performance targets, however, they can only be achieved within the bounds of multiple layers of constraints. The first set of constraints are device-level constraints, which are due to the limitations of medical devices' hardware. These devices are miniature, and resource constrained. Limited operating power creates tension between sustaining QoS metrics and preserving device lifetime. Continuous and frequent high data rate transmissions can drain power. Furthermore, computational and memory limitations restrict the complexity of algorithms and the size of data buffers that can be deployed on these devices. Environmental constraints also influence performance. HIoT D2D networks function in NFW medical environments where unpredictable channel conditions, multipath fading and interference impact the stability of network connectivity. Therefore, these constraints must be appropriately modelled.

D. Formulation of Optimal Connectivity

To achieve optimal D2D connectivity, the conflicting goal is to minimize and maximize multiple objective functions simultaneously. Thus, the connectivity problem can be defined as a stochastic MO-MF-MC optimization problem. The objectives functions are the QoS performance metrics that capture the QoS goals, which are to minimize latency, jitter, loss and maximize data rate simultaneously. Constraints functions, which are either deterministic or stochastic, model the limitations imposed by devices and the NFW environmental factors. The stochastic MO-MF-MC problem that can be framed from two perspectives. These perspectives, which are discussed in this section are the Constraint Based (CB) and Pareto Optimal Vector (POV) perspectives. Note that in this paper, these perspectives are QoS-based or QoS-focused. The modeling for each of these perspectives involves six steps, which are the formulation of: 1) QoS metrics as objective functions, 2) QoS bounds, 3) one liner connectivity, 4) compact max connectivity, 5) connectivity indicator (for one liner and compact max form) and 6) the optimal connectivity. The notations used in the formulation, and their definitions are outlined in Table I.

TABLE I. FORMULATION NOTATIONS AND DEFINITIONS

NOTATION	DEFINITION
i	Index of a QoS metric
k	Total number of QoS metrics
x	Decision variable
b	QoS bounds
f(x)	Objective function
g(x)	Inequality constraint
h(x)	Equality constraint
$g_j(x,\omega)$ $h_t(x,\omega)$	Stochastic constraint, (device/environmental)
ω	Randomness/uncertainty
∂	Weight
β	set of all x that fulfills the QoS bounds for POV perspective
P	Probability of occurrence
α	Probabilistic threshold/reliability level
σ	POV directions. σ_i=+1: metric is minimized and -1: metric is maximized
1	Binary indicator

1) Constraint Based (CB) perspective

The CB perspectives for optimal connectivity, their formulation steps and how they are interrelated are presented in this subsection. CB perspective treats connectivity as a feasibility question on a strict binary bound or a chance bound.

The outcome for connectivity is either binary (feasible or not) or based on the chance of achieving a given probabilistic threshold. The former case is termed constraint-based binary (CBB) while the latter is constraint based stochastic (CBS). From CBB perspective, connectivity exists if the specified QoS targets are satisfied; otherwise, it does not. CBS states that connectivity exists when the QoS metric bounds are met with a probability.

Moreover, the binary outcome in CBB can be specified as being deterministic (CBB-D) or as stochastic (CBB-S). In CBB-D, QoS objective functions are set to be achieved in a deterministic "ideal" environmental condition in which there no uncertainty or randomness. The objective function takes the form f(x) in equation (2). In addition, with reference to equation (1), these functions may be constrained with equality or inequality and expressed as fi(x)=bi, fi(x)<bi or fi(x)>bi, if no randomness exists. Equations (2) – (7) express the formulation steps for CBB-D.

STEP 1: QoS metrics	$f_i(x), \quad i = 1 \ldots . k$	(2)
STEP 2: QoS bounds	$f_i(x) \leq b_i \quad i = 1 \ldots . k \quad \text{and} \quad f_i(x) \geq b_i \quad i = 1 \ldots . k$	(3)
STEP 3: One liner Connectivity	$Connectivity \Leftrightarrow \forall i,\ f_i(x) \leq b_i$	(4)
STEP 4: Compact max form	$Connectivity \Leftrightarrow \max\ (f_i(x) - b_i), \leq 0$	(5)
STEP 5: Connectivity Indicator	*One liner:* $\Phi_{det}(x) = \mathbf{1}\{\forall i:\ f_i(x) \leq b_i\} \in \{0,1\}$	(6a)
	Compact max form: $\Phi_{det}(x) = \mathbf{1}\left\{\forall i:\ \max_{i=1\ldots k}\ (f_i(x) - b_i), \leq 0\right\} \in \{0,1\}$	(6b)
STEP 6: Optimal Connectivity	$Optimal\ Connectivity \Leftrightarrow \Phi_{det}(x) = 1$	(7)

The QoS objective functions in CBB-S are set with "realistic conditions", that reflects the existence of randomness within the network. The functions take the form f(x,ω), in equation (8) where ω indicates the uncertainty influencing the QoS objective. If randomness exists, the objective functions may also be constrained with equality or inequality and can expressed as fi(x,ω)=bi, fi(x,ω)<bi or fi(x,ω)>bi). Equations (8) – (13) express the formulation steps for CBB-S.

STEP 1: QoS metrics	$f_i(x,\omega), \quad i = 1 \ldots . k$	(8)
STEP 2: QoS bounds	$f_i(x,\omega) \leq b_i \quad i = 1 \ldots . k$	(9a)
	$f_i(x,\omega) \geq b_i \quad i = 1 \ldots . k$	(9b)
STEP 3: One liner Connectivity	$Connectivity \Leftrightarrow \forall i,\ f_i(x,\omega) \leq b_i$	(10)
STEP 4: Compact max form	$Connectivity \Leftrightarrow \max_{i=1\ldots k}\ (f_i(x,\omega) - b_i), \leq 0$	(11)
STEP 5: Connectivity Indicator	*One liner:* $\Phi_{stoch}(x,\omega) = \mathbf{1}\{\forall i:\ f_i(x,\omega) \leq b_i\} \in \{0,1\} \quad for\ each\ \omega$	
	Compact max form: $\Phi_{stoch}(x,\omega) = \mathbf{1}\left\{\forall i:\ \max_{i=1\ldots k}\ (f_i(x,\omega) - b_i), \leq 0\right\} \in \{0,1\}\ for\ each\ \omega$	(12)
STEP 6: Optimal Connectivity	$Optimal\ Connectivity \Leftrightarrow \Phi_{stoch}(x,\omega) = 1$	(13)

In CBS, connectivity is stochastic, and objective functions are chance (stochastically) constrained and takes the form P(fi(x, ω) $\leqslant$bi) $\in$ [α 1]). Objectives must satisfy at least a target probability threshold. Connectivity holds when the probabilistic QoS requirements modeled by the objective functions are all satisfied in other words, each QoS bound, bi is met with probability of at least a target α. (i.e. within threshold [α 1] and under bounded device and environmental constraints

A QoS metric is satisfied if there is a probability of $\geqslant \alpha$ of its value being within the required bound. CBS builds upon the CBB-S notion by requiring that connectivity is established at a probability of at least some target level (e.g., 95%). Connectivity is defined in terms of the reliability of the network given uncertain conditions. This means that the network is "connected" when the QoS bounds are achieved with at least the likelihood threshold that is specified, thus reflecting randomness and variation in channel and traffic conditions. This captures real-world variability in the NFW environment while still being constraint-based. The requirements are stated in probabilistic terms. So, CBS is a decision-level formalization that uses the CBB-S and then controls it via a probabilistic threshold, in order to gauge the networks' reliability in the presence of uncertainty A reliability threshold is a common way to certify connectivity under uncertainty in wireless QoS contexts [21][22].

Generally, from CB perspective, optimal connectivity exists if all QoS objective function are simultaneously met within acceptable bound, if one bound cannot be guaranteed, then connectivity does not exist. The connectivity feasibility indicator is either {0,1} or it is feasible with a probability $\alpha \in$ [α 1]. All quantities f(x, ω) and constraints g(x,ω) or h(x,ω) have fixed performance expectations, which may be deterministic or stochastic. Equations (14) – (19) give the formulation steps for CBS.

STEP 1: QoS metrics	$f_i(x,\omega), \quad i = 1 \ldots . k$	(14)
STEP 2: QoS bounds	$f_i(x,\omega) \leq b_i \quad i = 1 \ldots . k$	
	$f_i(x,\omega) \geq b_i \quad i = 1 \ldots . k$	(15)
STEP 3: One liner Connectivity	$Connectivity \Leftrightarrow P(\forall i,\ f_i(x,\omega) \leq b_i)$	(16)
STEP 4: Compact max form	$Connectivity \Leftrightarrow P(\max_{i=1\ldots k}\ (f_i(x,\omega) - b_i), \leq 0$	(17)
STEP 5: Connectivity Indicator	$\Phi_{stoch}(x,\omega) = \mathbf{1}\{\forall i:\ f_i(x,\omega) \leq b_i\} \in \{0,1\}$	
	$\Phi_{stoch}(x,\omega) = \mathbf{1}\left\{\forall i:\ \max_{i=1\ldots k}\ (f_i(x,\omega) - b_i), \leq 0\right\} \in \{0,1\}$	
	For both online and compact max, probability of being connected	
	$\Phi_{CBS}(x,\omega) = \mathbf{1}\ (P(\Phi_{stoch}(x,\omega) = 1)) \in [0,1]$	(18)
STEP 6: Optimal Connectivity	*Chance constraint declaration of optimal connectivity at level alpha*	
	$Optimal\ Connectivity \Leftrightarrow \Phi_{CBS}(x,\omega)\ \geq \alpha$	(19)

2) Pareto Optimal Vector (POV) perspective

The constraint-based perspective strictly defines a binary feasible region which shows that all QoS metric bounds are being met. This does not require Pareto optimization because connectivity is bound by hard constraints. However, in optimization practice, especially under stochastic and resource-constrained environments, it is rarely possible to meet all objectives' strict thresholds simultaneously. Thus, Pareto optimality becomes important. Instead of absolute satisfaction, connectivity can be interpreted as being Pareto efficient, which means that no objective (e.g., latency) can be improved without worsening another (e.g., data rate). Thus, the connectivity indicator can be expressed as belonging to the Pareto frontier of feasible solutions. Pareto-based modeling is highly suitable for HIoT D2D where traffic flows are of MC-BAMS types, the environment is NFW with stochastic conditions and devices introduce constraints. The QoS performance objectives take the form σi fi(x, ω) ⩽bi where σi ∈{+1, −1} encodes direction (σi = +1 for "minimize," σi = −1 for "maximize," so all objectives are cast as ⩽). The QoS bounds bi are hardbound ceilings for QoS performance of cost type metrics, where upper limits is b^{max} and hardbound floors for QoS performance of benefit-type metrics with lower limits b^{min}. A Pareto efficient point (PEP) is any feasible decision, which no other feasible decision dominates under the Pareto dominance condition. A Pareto efficient vector (PEV) is the space vector of the objectives induced by a set of PEPs that reflects the combination of QoS metrics to be met simultaneously. The set of all PEVs form the Pareto front and naturally exhibits trade-offs among metrics. In Pareto perspective, connectivity means there exists at least one feasible PEV; which is expressed with the usual one-liner feasibility condition. Alternatively, the compact max form generates PEPs and thus PEVs using weighted sum scalarization. Pareto Optimal Vectors (POVs) denote the subset of PEVs that satisfy the floors/ceilings bound for the objective functions. Thus, the connectivity indicator specifies the binary existence of at least one POV. Optimal connectivity exists if there is at least one POV that can provide an acceptable optimal operational trade-off for the network QoS performance required by an application. The parameters of that POV are then used to configure D2D links. A PEV is the objective-space performance vector on the Pareto front while a POV is a PEV that satisfies the specified floors/ceilings. Equations (20) – (25) gives the optimal connectivity formulation steps for the POV perspectives.

STEP 1: QoS metrics
$$\mathbf{f}(x,\omega) = (\sigma_i\, f_i(x,\omega),\ i = 1 \ldots\ldots k) \tag{20}$$

STEP 2: QoS bounds
$$\mathbf{f}(x,\omega) \le \mathbf{b}^{min}\ , \mathbf{f}(x,\omega) \ge \mathbf{b}^{max} \tag{21}$$

STEP 3: One liner Connectivity

Given the Pareto dominance condition.:

$$\forall i\ f_i(x',\omega) \le\ f_i(x,\omega)\ \textit{and}\ \exists j:\ f_j(x',\omega) <\ f_j(x,\omega).$$

$$Connectivity \Leftrightarrow \{(x,\omega)\ \in\ PEV := \{\ \nexists x':\ \forall i\ f_i(x',\omega) \le\ f_i(x,\omega)\ \textit{and}\ \exists j:\ f_j(x',\omega) < f_j(x,\omega)\}\ \} \tag{22}$$

STEP 4; Compact max form
$$Connectivity \Leftrightarrow \{(x,\omega)\ \in\ PEV := \{\ \exists x: \min_{z} \textstyle\sum_{i=1}^{k} \delta_i\ (f_i(x,\omega)\,, \delta_i \ge,\ \ \textstyle\sum_i \delta_i = 1\}\} \tag{23}$$

STEP 5; Connectivity Indicator
$$\Phi_{POV}(x,\omega) = \mathbf{1}\{x \in \beta\colon (x,\omega)\ \in\ PEV\} \in \{0,1\} \tag{24}$$

STE 6: Optimal Connectivity
$$Optimal\ Connectivity \Leftrightarrow \{x *\in : \Phi_{POV}(x,\omega) = \{x \in \beta\colon (x,\omega) \in PEV\} = 1$$
$$\beta = \{x\colon \mathbf{f}(x,\omega) \le \mathbf{b}^{min}\}\text{---------- bound condition} \tag{25}$$

E. Justification for POV

In real practice, multi-objective trade-offs are unavoidable. Though strict feasibility defines and models the ideal connectivity, Pareto vectors define and model the realistic operating points where optimal trade-offs are achieved. If strict thresholds are non-negotiable, then connectivity is treated with a hard feasibility. If trade-offs are possible, then connectivity is represented as a Pareto vector solution space. However, when conflicting objectives exist, such as minimizing latency, maximizing throughput, the network does not have a single feasible optimum, instead there is a set of solutions that form the Pareto optimal vectors, which are the feasible regions of connectivity. A set of Pareto optimal vectors indicate connectivity. In addition, the ability to visualize trade-offs through Pareto fronts makes stochastic MO-MF-MC optimization effective in HIoT D2D networks. For instance, a Pareto front might reveal that slightly higher latency can significantly extend device battery life, which is an acceptable trade-off for routine monitoring, but unacceptable in emergency care. Such nuanced decision support is vital for adaptive, real-time systems, where conditions shift unpredictably and human lives may depend on microsecond-level performance [21].

V. CONCLUSION AND FUTURE WORK

This paper studied connectivity in HIoT D2D networks operating in NFW environments and under strict resource limits, which makes connectivity fundamentally stochastic. Therefore, optimal connectivity in HIoT D2D networks has been modelled as a stochastic MO-MF-MC optimization problem, where the network must meet diverse traffic demands while operating within strict device and environmental limitations. The paper identified the unique characteristics of traffic flow in HIoT D2D as MC-BAMS.

Optimal connectivity was formulated from the CB and POV perspectives. A justification was made for the POV perspective. The constraint-based view defines optimal connectivity as either deterministic or chance constrained for reliability targets, while a Pareto view reveals the trade-off frontier where no QoS metric improves without another worsening within limits. POV is supported because it explicitly manages trade-offs among competing QoS metrics (e.g., latency, jitter, loss, data rate) while respecting device and environmental constraints. This perspective provides a practical foundation for scalable adaptive HIoT device to device systems in dynamic clinical settings. Future work includes a lightweight protocol that instantiates the chance constrained Pareto framework on constrained devices, online learning to tune priorities thresholds and schedules in real time, energy aware orchestration that couples power budgeting harvesting and thermal safety with QoS guarantees, privacy and safety co design aligned with clinical risk, hardware in the loop validation in ICU and home care testbeds with 5G and 6G URLLC, and open benchmarks to support reproducible progress on dependable HIoT connectivity.

ACKNOWLEDGMENT

The authors would like to thank Cape Peninsula University of Technology, Cape Town, South Africa for funding this research.

REFERENCES

[1] T. Islam and C. Kwon, “Survey on the state-of-the-art in device-to-device communication: A resource allocation perspective,” Ad Hoc Networks, vol. 136, p. 102978, Nov. 2022, doi: 10.1016/j.adhoc.2022.102978.

[2] S. E. El-deep, A. A. Abohany, K. M. Sallam, and A. A. Abd El-Mageed, “A comprehensive survey on impact of applying various technologies on the Internet of Medical Things,” Artificial Intelligence Review, vol. 58, art. 86, 2025, doi: 10.1007/s10462-024-11063-z.

[3] A. Iqbal, T. Khurshaid, A. Nauman, and S.-B. Rhee, “Energy-Aware Ultra-Reliable Low-Latency Communication for Healthcare IoT in Beyond 5G and 6G Networks,” Sensors, vol. 25, no. 11, p. 3474, May 2025, doi: 10.3390/s25113474.

[4] O. Bello and S. Zeadally, “Intelligent Device-to-Device Communication in the Internet of Things,” IEEE Systems Journal, 2015.

[5] A. E. Alattar and S. Mohsen, “A Survey on Smart Wearable Devices for Healthcare Applications,” Wireless Personal Communications, vol. 132, pp. 775–783, 2023, doi: 10.1007/s11277-023-10639-2.

[6] O. Bello and S. Zeadally, “Communication Issues in the Internet of Things (IoT),” in Next-Generation Wireless Technologies, Computer Communications and Networks (CCN), vol. 3016, pp. 189–219, 2013.

[7] M. Chen, J. Xu, and Y. Zhang, “Intelligent intensive care unit: Current and future trends,” Intensive Care Research, 2023. SpringerLink

[8] E.S. Dahiya, A.M. Kalra, A.Lowe and G. Anand., “Wearable technology for monitoring electrocardiograms in adults: A scoping review,” Sensors, vol. 24, no. 4, 2024. MDPI

[9] Y. Zhai et al., “5G-network-enabled smart ambulance: Architecture, application, and deployment,” Univ. of Sheffield White Rose Research, 2021. White Rose Research Online.

[10] A. K. Mishra, A. Behera, and A. K. Turuk, “Toward QoS monitoring in IoT edge devices driven healthcare—A survey,” Sensors, vol. 23, no. 21, 2023.

[11] A. Al Rashdi et al., “IoT-driven wearable devices enhancing healthcare: ECG classification at the edge,” Digital Health, 2024. ScienceDirect.

[12] A. K. Jha et al., “A survey of Internet of Medical Things: technology, application and challenges,” Internet of Things and Cyber-Physical Systems, 2024.

[13] Z. Guo, J. Li, and X. Zhang, “Internet of Medical Things: A systematic review,” Neurocomputing, 2023.

[14] Z. S. Lipcsey, E. O. Effanga, and C. A. Obikwere “Single and Multi-Objective Optimization – A Comparative Analysis,” J. Math. Comput. Sci., 2019.

[15] Z. Fei, B. Li, S. Yang, C. Xing, H. Chen, and L. Hanzo, “A Survey of Multi-Objective Optimization in Wireless Sensor Networks: Metrics, Algorithms and Open Problems,” arXiv preprint arXiv:1609.04069, 2016.

[16] N. Francesco, P. Gabriele and F. Emiliano, “Optimized Scheduling of IoT Devices in Healthcare Facilities,” Applied Sciences, vol. 15, no. 8, 4456, 2024.

[17] N. Mehran, D. Kimovski, and R. Prodan, “MAPO: A Multi-Objective Model for IoT Application Placement in a Fog Environment,” arXiv preprint arXiv:1908.01153, 2019.

[18] X. Zhang, R. Cheng, and Y.T. Y. Jin “Emerging Trends in Evolutionary Multi-Objective Optimization for IoT-Based Large-Scale Healthcare Applications,” IOS Press, 2022.

[19] C. She et al., "A tutorial on ultrareliable and low-latency communications in 6G: Integrating domain knowledge into deep learning." Proceedings of the IEEE 109.3 (2021): 204-246.

[20] O. L. A. López, et al., “Statistical Tools and Methodologies for URLLC—A Tutorial,” arXiv preprint arXiv:2212.03292, 2022. arXiv

[21] A. Iqbal, T. Khurshaid, A. Nauman, and S.-B. Rhee, “Energy-Aware Ultra-Reliable Low-Latency Communication for Healthcare IoT in Beyond 5G and 6G Networks,” Sensors, vol. 25, no. 11, p. 3474, May 2025, doi: 10.3390/s25113474. MDPI

[22] G. Ghatak, S. R. Khosravirad, and A. De Domenico, “Stochastic Geometry Framework for Ultrareliable Cooperative Communications With Random Blockages,” IEEE Internet of Things Journal, vol. 9, no. 7, pp. 5150–5161, Apr. 2022, doi: 10.1109/JIOT.2021.3108955

On the Pseudo-Bayesian Broadcast Control Algorithm for Slotted ALOHA in Multi Packet Reception and under Impaired Channel Conditions

Vicente Casares-Giner* and Frank Y. Li‡
*Departamento de Comunicaciones, Universitat Politècnica de València (UPV), València 46022, Spain
‡Dept. of Information and Communication Technology, University of Agder (UiA), N-4898 Grimstad, Norway
Email: vcasares@upv.es; frank.li@uia.no

***Abstract*—The basic concept of slotted ALOHA as a Random Access Protocol (RAP) is commonly implemented for ubiquitous access in many wireless networks. In this paper, we study the generalization of the *network control by Bayesian broadcast* to the environment of M-MRP Multiple Packet Reception (MPR) when channel impairments are considered. In our M-MPR model, up to M data packets transmitted in the same time-slot can be correctly decoded by using capture effect and some advanced signal processing techniques such as Successive Interference Cancellation (SIC) combined with Multiple-Input Multiple-Output (MIMO). We show that the broadcast or permission probability that maximizes the throughput (packets per slot successfully transmitted) is sensitive to channel characteristics. While with ideal channel conditions of maximum capacity a binary feedback – collision versus non-collision – is required, and in the more realistic channel conditions, $M + 1$ feedback is needed.**

Keywords-Pseudo-Bayesian Control, Multiple Packet Reception.

I. INTRODUCTION

For ubiquitous multi-access in wireless networks, a single channel is shared by a population of devices or users. In order to share this common transmission medium among users, a Medium Access Control (MAC) protocol must be properly designed. When users act in an independent manner, i.e., with minimum coordination between them, we need a suitable Random Access Protocol (RAP). The area of RAPs started with the seminal work by N. Abramson in 1970 [1], where the ALOHA protocol was proposed. Later in 1972 [2], Roberts adds to ALOHA the additional feature of slot synchronization, so the S-ALOHA was proposed as a substantial improvements of its throughput, increasing from the $1/2e \approx 0.1839$ channel utility for pure ALOHA to $1/e \approx 0.3679$ packets/slot for S-ALOHA. Since then, many RAPs based on the ALOHA principles have been proposed for wired Local Area Networks (LANs) and wireless (cellular, Wireless Fidelity (WiFi), etc.) communication systems. The main advantage of ALOHA protocol is its easy and simple implementation. Unlike Carrier Sense Multiple Access (CSMA) protocol, in ALOHA no sensing functioning needs to be performed. Furthermore, the hidden terminal effect that can significantly deteriorate the CSMA performance does not affect the operation of the ALOHA protocol. A basic background on this matter can be found in [3] [5] [7].

ALOHA alike protocols are inherently located at the MAC layer. The improvement of ALOHA protocols can be achieved when combining with other physical layer techniques such as Multi-User Detection (MUD), Multiple-Input Multiple-Output (MIMO) or a combination of both techniques (MU-MIMO). In MUD, a single receiver is able to decode the intended signals from interference and noise. MUD techniques include Maximum-Likelihood (ML), Parallel Interference Cancellation (PIC), Successive Interference Cancellation (SIC), etc. In MIMO technique, more than one antenna at transmitter and at the receiver part are installed to get improvements in parameters such as throughput and channel robustness. For more details, interested readers are referred to [11]. At the physical layer, the use of MIMO, MUD and SIC will benefit the Multiple Packet Reception (MPR) reception technique. So, a cross layer cooperation based on the use of at the physical layer and the S-ALOHA protocol at the MAC layer will bring benefits in the throughput of wireless access system. Thanks to this cooperation, we can enjoy the M-MPR capability.

Additional contributions in the M-MPR area can be found in [6] [10]; where the number of packets that can be received and decoded simultaneously is M, and the stability analysis of MPR is studied in a deep way. In [12], the authors study the M-MPR using the principle of MUD at the Base Station (BS). The authors adopt the adaptive interference canceler employing the Recursive Least Square Maximum Likelihood Sequence Estimation (RLS-MLSE) scheme. Through computer simulation and field trial under a realistic scenario, it is shown that up to three ($M = 3$) simultaneously transmitted packets can be detected, even though they limit their study to $M = 2$. That is, for $M = 2$, very reliable of real time applications, the maximum throughput can exceed 0.7, which is a significant improvement compared to the convention S-ALOHA of $1/e \approx 0.3679$.

In [14], a finite number of devices access to a common wireless channel using S-ALOHA, where the M-MPR scheme with the *all-or-nothing* philosophy is assumed. Devices operate in saturation conditions (there are always packets to be transmitted) and the permission or transmission probability is constant. Their analysis lacks of dynamic adaptation of the transmission probability. In [15], the authors provide an in-depth analysis of the M-MPR protocol for ALOHA and CSMA random access algorithms. However, with regard to ALOHA protocol, the analysis does not take into account the arrival process that could joint backlogged data packets.

In order to avoid total loss of packets to collisions, several strategies supporting power transmission have been proposed for ALOHA packets [17] [18]. Hence, in [18], the authors study the non-orthogonal random access technique for 5th Generation (5G) networks in which due to the different level

of the received power at the BS, it enables the BS to decode two packets simultaneously using SIC. The analysis is carried out in terms of access delay, throughput, and energy efficiency.

The capture effect can happen so allowing the decoding of a number of packets lower or equal to the number of packets that simultaneously coincide in the same time slot. The authors of [12] provide an analysis quite parallel to our work but the novelty of our work is the Bayesian estimation of the number of users in contention in a framed-slotted ALOHA environment, as an enhancement to the work by [4]. In [19], the distribution of new plus backlogged packets are assumed to follow a Poisson distribution.

In all these previous studies, the main assumption is that the channel is ideal, i.e., neither fading nor interference happens. All of them consider that the channel capacity is M and, when the number of data packets in one slot is not greater than M all packets can be successfully decoded, otherwise the slot is considered as collision (garbled).

In this work, we assume a general model where $\alpha_{m,k}$ for $0 \leq k \leq m \leq M$ denotes the conditional probability to detect correctly k packets assuming that m packets were transmitted. The aim of this paper is to extend the pseudo-Bayesian broadcast control algorithm of Rivest [4] developed to Single-Packet Reception (SPR) to the case of MPR. Then, first we deal with a finite number of active devices and second we follow with an arbitrary number of active devices. The closest approach to our work or the most related work with our paper is the one presented in [16], but they use the *all-or-nothing* model defined below.

The rest of this paper is organized as follows. In Sec. II, we describe the model of the system under study. In Sec. III, the optimal permission probability for a given number of active devices is derived. Sec. IV deals with the estimation of the number of active devices, so with the updating permission probability based on the Bayesian rules. In Sec. V, we introduce the common assumption of Poisson distribution for the number of active devices and the Pseudo Bayesian procedure is described. In Sec. VI, some particular cases are studied. The paper ends with conclusions in Sec. VII.

II. System model

Consider a time-slotted channel. A finite number of active devices, sufficiently large enough, transmit their packets uplink towards an Access Point (AP) or a 5G BS, i.e., nest generation Node B (gNB). A given device becomes active when it has a packet ready to transmit. Packets are of constant length that fits with the length of the time-slot.

Devices follow the Immediate First Transmission (IFT) principle instead of the Delayed First Transmission (DFT) principle. That is, as soon as a given device becomes active the corresponding packet joints the set of backlogged packets and follows the RAP's rules. In the RAP, all active devices (new or backlogged) transmit with the same broadcast or permission probability provided by the gNB instantly in the beginning of a time-slot. In other words, new and backlogged packets are treated in the same way. The permission probability is updated by the gNB in a slot-by-slot basis according to the observed results in each time slot and according to the expected number of new active devices (the arrival process).

In the M-MPR model, the channel for transmission-reception is represented by a set of conditional probabilities, time invariant, given in the following the MPR stochastic matrix [6],

$$\mathbf{A} = \left[\begin{array}{cccc|c} \alpha_{0,0} & 0 & \dots & 0 & \alpha_{0,c} \\ \alpha_{1,0} & \alpha_{1,1} & \dots & 0 & \alpha_{1,c} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \alpha_{M,0} & \alpha_{M,1} & \dots & \alpha_{M,M} & \alpha_{M,c} \\ \hline \alpha_{M_>,0}=0 & \alpha_{M_>,1}=0 & \dots & \alpha_{M_>,M}=0 & 1 \end{array}\right] \quad (1)$$

In (1), $M_>$ is identified as *greater than* M and $\alpha_{i,c}$ denotes the probability the receiver interprets as collision when i packets are transmitted. The set $\{\alpha_{i,j}\}$ contains the conditional probabilities that characterize the transmission-reception characteristics of the wireless channel. Each probability $\alpha_{i,j}$ is interpreted as follows. For an arbitrary time slot, first, we assume that no packets are transmitted. Then, with probability $\alpha_{0,0}$, the slot is correctly interpreted by the gNB, i.e., as a hole, and with probability $\alpha_{0,c} = 1 - \alpha_{0,0}$, the empty slot may be seen as a garbled or collision time-slot, for instance, due to the interference and noise of the channel. Second, we assume that a single packet has been transmitted, the second row of the MPR matrix. Then, the gNB interprets, with probability $\alpha_{1,0}$, as an empty slot (the transmitted packet might vanish due to channel fading conditions), with probability $\alpha_{1,1}$, the packet is correctly decoded and with probability $\alpha_{1,c} = 1 - \alpha_{1,1} - \alpha_{1,0}$, the slot is observed as a garbled time slot (collision). Third, we assume that two packets are simultaneously transmitted, the third row of the MPR matrix. Then, with probability $\alpha_{2,0}$, the slot is observed as empty; with probability $\alpha_{2,1}$, one of the two packets is correctly decoded while the other one is lost (the capture effect [8]); with probability $\alpha_{2,2}$, both packets are correctly decoded (using SIC techniques [18]), and with probability $\alpha_{2,c} = 1 - \alpha_{2,0} - \alpha_{2,1} - \alpha_{2,2}$, the observed slot is seen as garbled, as a collision slot. And so on. Finally, when in the same observed time slot more than M packets are transmitted, with probability 1, the gNB interprets as a collision slot, i.e., for $i > M$ we have $\alpha_{i,j} = 0$ and $\alpha_{i,c} = 1$.

In the M-MPR model, the *all-or-nothing* scheme has often been considered. Accordingly, the receiving station is able to successfully decode m simultaneous transmissions with probability one if and only if $m \leq M$ and no decoding can be achieved when $m > M$, which in turns means that $\mathbf{A} = \mathbf{I}$, the identity matrix. This is the typical assumption in many papers such as [15], [16], [19]. Our study generalizes this particular case. For some particular cases, in the same way as in [9], we consider the case where the set of probabilities $\{\alpha_{i,j}\}$ being system feature, are known *a priori* or a good estimation of them is known.

III. Broadcast or Permission Probabilities

We consider a number of active devices N_t, each one with a single packet ready to be transmitted at time-slot t. The idea

is to use the optimal broadcast or permission probability that maximizes some relevant function, such as the throughput, defined as the mean number of packets successfully transmitted in time-slot t. Here, we obtain the optimum permission probability, first when the number of active devices is finite and second when this number follows a given distribution.

A. For a Fixed Number of Active Devices

We assume a fixed number of active devices, $N_t = n$, each one with one packet ready to be transmitted at time-slot t. We consider that $n > M$. N_t needs to be estimated, but initially we assume that the gNB has perfect knowledge of it. Each active device will transmit with the probability of permission $b_{M,t}$ and will wait for the next slot with the probability $w_{M,t} = 1 - b_{M,t}$ (the IFT principle). Then, the following events are considered, empty slot (hole), slot with m successes (success=m), with $0 \leq m \leq M$, and slot with collision; i.e., the probability of observing a hole,

$$\begin{aligned} &Pr(hole/(N_t = n, b_{M,t})) = \\ &H_{b_{M,t}}(n) = \textstyle\sum_{k=0}^{M} B_k^n(b_{M,t})\alpha_{k,0}, \end{aligned} \tag{2}$$

where $B_k^n(b_{M,t})$ denotes the binomial distribution,

$$B_k^n(b_{M,t}) = \left(\begin{array}{c} n \\ k \end{array} \right) b_{M,t}^k . w_{M,t}^{n-k}, \quad 0 \leq k \leq n.$$

The probability of observing m successes,

$$\begin{aligned} &Pr(success = m/(N_t = n, b_{M,t})) = \\ &S_{m,b_{M,t}}(n) = \textstyle\sum_{k=m}^{M} B_k^n(b_{M,t})\alpha_{k,m}, \end{aligned} \tag{3}$$

and, the probability to observe a collision,

$$\begin{aligned} &Pr(collision/(N_t = n, b_{M,t})) = \\ &C_{b_{M,t}}(n) = 1 - H_{b_{M,t}}(n) - \textstyle\sum_{m=1}^{M} S_{m,b_{M,t}}(n) = \\ &1 - \textstyle\sum_{m=0}^{M} S_{m,b_{M,t}}(n); \quad (S_{0,b_{M,t}}(n) = H_{b_{M,t}}(n)). \end{aligned} \tag{4}$$

Observe that the event hole can be regarded as the event m = 0 success, i.e., $H_{b_{M,t}}(n) = S_{0,b_{M,t}}(n)$, and this explains the last equality in (4). The mean value of the number of packets successfully transmitted is given, after some simple rearrangement of terms, by

$$\begin{aligned} &E(\#successes/(N_t = n, b_{M,t})) = \\ &\textstyle\sum_{m=1}^{M} m Pr(success = m/(N_t = n, b_{M,t})) \\ &\quad = \textstyle\sum_{m=1}^{M} m \sum_{k=m}^{M} B_k^n(b_{M,t})\alpha_{k,m} = \sum_{m=1}^{M} B_m^n(b_{M,t})\overline{\alpha}_m, \end{aligned} \tag{5}$$

where $\overline{\alpha}_m = \sum_{k=1}^{m} k\alpha_{m,k}$ $(0 < m \leq M)$ is the expected number of correctly decoded packets when m packets are transmitted simultaneously in the same time-slot [13]. The maximum of (5) can be computed by differentiating and root finding. Then, from (5), we have found for $\hat{b}_{M,t}$ (also, see (5) in [15] where $\overline{\alpha}_m = m$),

$$\hat{b}_{M,t} = \frac{\sum_{m=1}^{\min(n,M)} m\overline{\alpha}_m B_m^n(\hat{b}_{M,t})}{n \sum_{m=1}^{\min(n,M)} \overline{\alpha}_m B_m^n(\hat{b}_{M,t})} = h_{M,\alpha}(\hat{b}_{M,t}) \tag{6}$$

with $\alpha = [\overline{\alpha}_1, \overline{\alpha}_2, \ldots, \overline{\alpha}_M]$. In (6) we can apply the fixed point iteration method, i.e., $\hat{b}_{M,t}^{(i+1)} = h_{M,\alpha}(\hat{b}_{M,t}^{(i)})$, $i = 1, 2, \ldots$, with $\hat{b}_{M,t}^{(0)} \in [0, 1]$ and the optimum permission probability $\hat{b}_{M,t} = \hat{b}_{M,t}^{(\infty)}$ is obtained, i.e., the iteration always converges to the unique solution. Moreover, explicit expressions can be found for $M = 1, 2$, and 3. When $M = 1$, we have $\alpha_{\geq 2,c} = 1$, equivalent to $\overline{\alpha}_{\geq 2} = 0$ and trivially we obtain $\hat{b}_{1,t} = K_1/n = 1/n$. When $M = 2$, $\alpha_{\geq 3,c} = 1$ (equivalent to $\overline{\alpha}_{\geq 3} = 0$), the optimum value of the permission probability, $\hat{b}_{2,t}$ is,

$$\hat{b}_{2,t} = \frac{(n-1)\overline{\alpha}_2 - (n+1)\overline{\alpha}_1 + \sqrt{\Delta}}{n\left[(n-1)\overline{\alpha}_2 - 2\overline{\alpha}_1\right]} > \frac{1}{n}; \quad , n = 3, 4, \ldots$$

with $\Delta = (n-1)\left[(n-1)\left(\overline{\alpha}_1^2 + \overline{\alpha}_2^2\right) - 2\overline{\alpha}_1\overline{\alpha}_2\right] = (n-1)\left[n\left(\overline{\alpha}_1^2 + \overline{\alpha}_2^2\right) - \left(\overline{\alpha}_1 + \overline{\alpha}_2\right)^2\right]$. For large values of n we can write,

$$\hat{b}_{2,t} \approx \frac{\overline{\alpha}_2 - \overline{\alpha}_1 + \sqrt{\overline{\alpha}_1^2 + \overline{\alpha}_2^2}}{n\overline{\alpha}_2} = \frac{K_2}{n} > \frac{1}{n}; \quad , n = 3, 4, \ldots$$

with

$$\begin{aligned} &K_2 = 1 + \frac{\sqrt{1 + \left(\overline{\alpha}_2/\overline{\alpha}_1\right)^2} - 1}{\overline{\alpha}_2/\overline{\alpha}_1} = 1 + \frac{\sqrt{1+x^2}-1}{x} = \\ &1 + \frac{1}{2}x - \frac{1}{2^2.2!}x^3 + \frac{1.3}{2^3.3!}x^5 - \frac{1.3.5}{2^4.4!}x^7 + \frac{1.3.5.7}{2^5.5!}x^9 \ldots \\ &\quad \text{and} \quad \mathrm{x} = \frac{\alpha_{2,1} + 2\alpha_{2,2}}{\alpha_{1,1}} = \frac{\overline{\alpha}_2}{\overline{\alpha}_1}. \end{aligned} \tag{7}$$

Note that when $\overline{\alpha}_2 \to 0$, i.e., $x \to 0$, the evaluation of (7) using the closed form (the expression with a square root) may lead to some imprecise calculation. In this case we could use the approximation given by the Taylor expansion.

Due to the page limit, we omit the exact analytical expression for $M = 3$. In general, for any M, and for large values of n, $\hat{b}_{M,t}$ can be expressed as $\hat{b}_{M,t} \approx K_M/n$. In fact, (5) can be approximated by

$$\begin{aligned} &E(\#successes/(N_t = n, b_{M,t})) = \\ &\approx \textstyle\sum_{m=1}^{M} \frac{(nb_{M,t})^m}{m!}\overline{\alpha}_m e^{-nb_{M,t}}. \end{aligned} \tag{8}$$

B. For a Random Number of Active Devices.

Now, we assume that N_t follows a discrete probability distribution, $p_{n,t}$, with Generator Function (GF), given by, respectively

$$Pr(N_t = n) = p_{n,t}; \quad P_t^*(z) = \sum_{n=0}^{\infty} p_{n,t} z^n. \tag{9}$$

Furthermore, we assume that the gNB has a perfect knowledge of $p_{n,t}$. Therefore, unconditioning (5) with $p_{n,t}$, the throughput, which defined as the expected number of successes at time-slot t, is given by, after some algebra,

$$T_{M,\alpha}(b_{M,t}) = E(Pr(\text{success at slot t})/b_{M,t}) = \sum_{m=1}^{M} \frac{b_{M,t}^m}{m!} \frac{d^m P_t^*(w_{M,t})}{dw_{M,t}^m} \overline{\alpha}_m. \quad (10)$$

The optimum permission probability $\hat{b}_{M,t} = 1 - \hat{w}_{M,t}$ that maximizes (10), a polynomial in the unknown variable $b_{M,t}$, can be computed by differentiating and root finding. In a practical sense, the computation required to obtain $b_{M,t}$, would be time consuming. This can be avoided by using the approximation $\hat{b}_{M,t} \approx K_M/E(N_t)$. However, the pseudo-Bayesian broadcast algorithm described in the next section appears to be an excellent approach [4].

IV. Estimating Number of Active Devices

In an M-MPR channel, the permission probability $\hat{b}_{M,t}$ to be used in time-slot t is evaluated according to the procedure described in previous section. $\hat{b}_{M,t}$ is updated on a slot-by-slot basis. The update procedure is based on the outcomes in time-slot t observed by the gNB and on the arrival process of new packets, i.e., on the number of devices that become active during time-slot t. For the first item, we apply Baye's rule, as suggested in [4]. For the second item we consider a general distribution $\{a_{n,t}\}$ with GF, $A^*(z) = \sum_{n=0}^{\infty} a_{n,t} z^n$. Furthermore, the arrival process is assumed to be independent of the RAP.

A. Bayesian Updating of the Probability Vector

Assume that the procedure to estimate the probability vector N_t, $\overline{p}_t = [p_{0,t}, p_{1,t}, p_{2,t}, \ldots]$, is reasonably good. Now, we describe how the gNB updates this probability vector of N_t, given that slot t was a hole, a success-m, or a collision. Denote E=Evidence (hole, success-m, collision) and H=Hypothesis ($N_t = n$ data packets). The Bayes' rule tells us,

$$Pr(H/E) = \frac{Pr(E/H)Pr(H)}{Pr(E)}. \quad (11)$$

Then, the gNB will use the evidence available up to time-slot t to update $\{p_{n,t}\}$, given the available evidence. This is the so called Bayesian broadcast procedure, since it relies on Bayesian reasoning to estimate $\overline{p}_t = [p_{0,t}, p_{1,t}, p_{2,t}, \ldots]$ according to (11).

Let $p'_{n,t}$ denote the final probability $Pr(N_t = n/E_t)$ where E_t is the slot t evidence (hole, success-m, or collision), i.e., $\overline{p}'_t = [p'_{0,t}, p'_{1,t}, p'_{2,t}, \ldots]$. The probabilities $p'_{n,t}$ $(Pr(H/E))$ are easily obtained using Bayes' rule by multiplying each initial probability $p_{n,t}$ $(Pr(H))$ by the appropriate likelihood $H_{b_{M;t}}(n)$, $S_{m,b_{M;t}}(n)$ or $C_{b_{M;t}}(n)$ $(Pr(E/H))$ (see (2), (3) and (4)), according to whether a hole, success-m, or collision was observed, and then normalizing so that the $p'_{n,t}$ add up to one. Then, the numerator of (11) is evaluated as follows,

If the gNB observes a hole,

$$\overline{p}'_t = \frac{[p_{0,t}H_{b_{M;t}}(0), p_{1,t}H_{b_{M;t}}(1), \ldots]}{Ch_t}. \quad (12)$$

If the gNB observes a success-m event, for $m = 1, 2, \ldots, M$,

$$\overline{p}'_t = \frac{[p_{0,t}S_{m,b_{M;t}}(0), p_{1,t}S_{m,b_{M;t}}(1), \ldots]}{Cs_{m,t}}. \quad (13)$$

Finally, if the gNB observes a collision event,

$$\overline{p}'_t = \frac{[p_{0,t}C_{b_{M;t}}(0), p_{1,t}C_{b_{M;t}}(1), \ldots]}{Cc_t}. \quad (14)$$

where $Ch_t = \sum_{n=0}^{\infty} p_{n,t}H_{b_{M;t}}(n)$, $Cs_{m,t} = \sum_{n=0}^{\infty} p_{n,t}S_{m,b_{M;t}}(n)$ and $Cc_t = \sum_{n=0}^{\infty} p_{n,t}C_{b_{M;t}}(n)$ are the respective normalization constants. Note that the case hole can be regarded as a particular case of success-m when $m = 0$, i.e., $H_{b_{M;t}}(k) = S_{0,b_{M;t}}(k)$ and $Ch_t = Cs_{0,t}$.

B. Modeling Successful Packet Transmission

When the gNB observes the evidence S_m $(m = 1, 2, \ldots, M)$, the number of packets pending to be transmitted is m less than the estimated number before the access action. For the evidences H and C the number of packets that are pending to gain the access in the next time slot $t+1$ is the same as the one we have at time slot t. Therefore, considering the observations, hole, success-m $(m = 1, 2, \ldots, M)$ or collision, we have, including the GF of the probability vector,

If a hole is observed,

$$p''_{n,t} = p'_{n,t} \Rightarrow \quad P_t^{''*}(z) = P_t^{'*}(z). \quad (15)$$

If a success-m is observed,

$$p''_{n,t} = p'_{n+m,t} \Rightarrow \quad P_t^{''*}(z) = P_t^{'*}(z)z^{-m}. \quad (16)$$

If a collision is observed,

$$p''_{n,t} = p'_{n,t} \Rightarrow \quad P_t^{''*}(z) = P_t^{'*}(z). \quad (17)$$

C. Modeling the Arrivals of New Packets

Let us assume that new packets arrive independently of the the contention process. Assuming a memoryless arrival process on a slot basis, we define $a_{n,t}$ the probability that n packets are generated in time slot t with GF $A_t^*(z) = \sum_{n=0}^{\infty} a_{n,t} z^n$. Furthermore, we also assume that $\hat{a}_{n,t}$, $(\hat{A}_t^*(z))$, the estimation of $a_{n,t}$, $(A_t^*(z))$, it can be done with sufficient accuracy.

D. The Probability Vector at Time Slot t+1

Since the arrival process is independent of the RAP, the GF of probability vector at time-slot $t+1$ is the product of the two related generating functions, i.e.,

$$P_{t+1}^*(z) = \sum_{n=0}^{\infty} p_{n,t+1} z^n = P_t^{''*}(z)\hat{A}_t^*(z) = \begin{cases} P_t^{'*}(z)\hat{A}_t^*(z), & \text{hole} \\ P_t^{'*}(z)z^{-m}\hat{A}_t^*(z), & \text{success}-\text{m} \\ P_t^{'*}(z)\hat{A}_t^*(z), & \text{collision} \end{cases} \quad (18)$$

and the optimum broadcast probability $\hat{b}_{M,t+1}$ for the next slot $t+1$ is derived using the vector probability given by (18) in (10) and the root finding procedure. With this last step, the cycle is completed.

V. The Probability Vector: Poisson Arrivals

The previous procedure can be simplified by assuming that, in the same way as in many other works, [4], [5], [6], [12], [16], [18], [19], the vector of probabilities $p_{n,t}$ at time-slot t, (new arrivals + backlogged packets) can be approximated reasonably, by a Poisson distribution with rate ν_t. In this case, (9) turns as,

$$p_{n,t} = \frac{(\nu_t)^n}{n!}e^{-\nu_t}, \quad P_t^*(z) = e^{\nu_t(z-1)}. \tag{19}$$

Observe that now, the slot-by-slot updating procedure for the probabilities $p_{n,t}$ is simplified to the task of updating the rate $\nu_t = E(N_t)$, i.e., the single parameter that defines the Poisson distribution. We recall that ν_t is the average number of active devices at the beginning of time-slot t and it must be estimated. Then, inserting (19) into (10) and with the notation $x = \nu_t b_{M,t}$, after some algebra,

$$\begin{aligned} &T_{M,\alpha}(x) = E(Pr(\text{success at slott})/b_{M,t}) = \\ &= \sum_{k=1}^{M} \frac{(\nu_t b_{M,t})^k}{k!}\overline{\alpha}_k e^{-\nu_t b_{M,t}} = \sum_{k=1}^{M} \frac{x^k}{k!}\overline{\alpha}_k e^{-x}. \end{aligned} \tag{20}$$

Notice that the throughput is a function of the product $x = \nu_t b_{M,t}$. Let $\hat{x}_M = K_M$ be the value that maximizes $T_{M,\alpha}(x)$. Then, setting to zero the first derivative of (20), we have,

$$dT_{M,\alpha}(x)dx = \sum_{k=1}^{M}\left(\frac{x^{k-1}}{(k-1)!}\overline{\alpha}_k - \frac{x^k}{k!}\overline{\alpha}_k\right)e^{-x} = 0 \tag{21}$$

Leaving aside the exponential factor e^{-x}, the condition in (21) can be expressed in the following form,

$$x = \frac{x\sum_{k=1}^{M}\frac{x^{k-1}}{(k-1)!}\overline{\alpha}_k}{\sum_{k=1}^{M}\frac{x^k}{k!}\overline{\alpha}_k} = h_{M,\alpha}(x) \tag{22}$$

where we have defined the function $h_{M,\alpha}(x)$.

In addition, it is trivial to check that $h_{M,\alpha}(x) < h_{M+1,\alpha}(x)$ for $0 < x$, we can assert that, $\ldots \hat{x}_{M-1} < \hat{x}_M < \hat{x}_{M+1} \ldots$. Therefore, additional computing time savings can be achieved by choosing as the initial estimation for $\hat{x}_{M+1}$ the previous value, i.e., $x_{M+1}^{(0)} = \hat{x}_M = K_M$.

For $M = 1, 2, 3$, closed form expressions are obtained for $\hat{x}_M = K_M$; but, in general numerical computation to find K_M is required.

Since N_t is randomly distributed and since $\hat{b}_{M,t}$ is a probability where $\hat{x}_M = \nu_t \hat{b}_{M,t}$ we finally set,

$$\hat{b}_{M,t} = 1 - \hat{w}_{M,t} = min\left(\frac{K_M}{\nu_t}, 1\right). \tag{23}$$

Clearly, ν_t in (23) is unknown so it needs to be estimated and adapted in a slot-by-slot manner. Let $\hat{\nu}_t$ denote the estimation of ν_t at the beginning of time-slot t (in (23) $\hat{\nu}_t$ will be used instead of ν_t). Then, as we have discussed before, $\hat{\nu}_{t+1}$, the estimation of ν_{t+1}, is supported by two items. First, by the outcomes of slot t observed by the gNB. Second, by the arrival process of new packets that joint the backlogged packets and follow the common RAP. Remember, the algorithm is supported by the IFT principle.

A. Bayesian Updating of the Probability Vector

If the gNB observes a hole, (12) becomes, after the normalization step,

$$p'_{n,t} = \frac{\sum_{k=0}^{M} B_k^n(b_{M,t})\alpha_{k,0}\frac{\nu_t^n}{n!}e^{-\nu_t w_t}}{\sum_{k=0}^{M}\frac{(\nu_t b_{M,t})^k}{k!}\alpha_{k,0}}; \quad n \geq 0. \tag{24}$$

with GF,

$$P_t'^*(z) = \frac{\sum_{k=0}^{M}\frac{(\nu_t b_{M,t}z)^k}{k!}\alpha_{k,0}}{\sum_{k=0}^{M}\frac{(\nu_t b_{M,t})^k}{k!}\alpha_{k,0}}e^{\nu_t w_{M,t}(z-1)}. \tag{25}$$

We observe that (25) is a weighted sum of $M+1$ Poisson distributions, where each distribution is obtained by shifting k positions to the right ($k = 0, 1, \ldots, M$) the distribution $e^{\nu_t w_{M,t}(z-1)}$. Consequently, we could reconsider the initial hypothesis of Poisson distribution for $p_{n,t}$ and to inquire about a linear combination of $M+1$ Poisson distributions as a better distribution for $p_{n,t}$. However, to derive this possibility is beyond the scope of this paper.

The first derivative of $P_t'^*(z)$ evaluated at $z = 1$ is,

$$\begin{aligned} &\text{mean value}_{E=H} = \nu_t w_{M,t} + \frac{\sum_{k=0}^{M} k\frac{(\nu_t b_{M,t})^k}{k!}\alpha_{k,0}}{\sum_{k=0}^{M}\frac{(\nu_t b_{M,t})^k}{k!}\alpha_{k,0}} = \\ &\nu_t - x\frac{\sum_{k=0}^{M}\frac{x^k}{k!}(\alpha_{k,0}-\alpha_{k+1,0})}{\sum_{k=0}^{M}\frac{x^k}{k!}\alpha_{k,0}} \end{aligned} \tag{26}$$

with $x = \nu_t b_{M,t}$ and $\alpha_{M+1,0} = 0$, see channel characteristics in (1). Observe that, according to (15), (16), (17), we identify $P_t''^*(z) = P_t'^*(z)$.

Note that if $\alpha_{k,0} = \delta_{k,0}$ (Kronecker delta) then $meanvalue_{E=H} = \nu_t w_{M,t} = \max(\nu_t - K_M, 0)$. In other words, if $b_{M,t} = 1$, we are certain that the number of data packets ready for transmission was zero. Otherwise, this case cannot be confirmed when $b_{M,t} < 1$.

If the gNB observes the success-m event (13), i.e., m packets are successfully decoded, including the normalization step, we have,

$$p'_{n,t} = \begin{cases} 0; n = 0, 1, \ldots, m-1; \\ \frac{\sum_{k=m}^{M} B_k^n(b_{M,t})\alpha_{k,m}\frac{\nu_t^n}{n!}e^{-\nu_t w_{M,t}}}{\sum_{k=m}^{M}\frac{(\nu_t b_{M,t})^k}{k!}\alpha_{k,m}}; \quad n \geq m. \end{cases} \tag{27}$$

with a generating function,

$$\begin{aligned} &P_t'^*(z) = \sum_{n=0}^{\infty} p'_{n,t}z^n = \\ &\frac{\sum_{k=m}^{M}\frac{(\nu_t b_{M,t}z)^k}{k!}\alpha_{k,m}}{\sum_{k=m}^{M}\frac{(\nu_t b_{M,t})^k}{k!}\alpha_{k,m}}e^{\nu_t w_{M,t}(z-1)}. \end{aligned} \tag{28}$$

As in (25), we observe that (28) is a weighted sum of $M-m+1$ Poisson distributions, where each distribution is obtained by shifting the same distribution $e^{\nu_t w_{M,t}(z-1)}$ k positions to

the right ($k = m, m+1, \ldots, M$). The first derivative of (28) evaluated at $z = 1$, gives us, after some simple algebra,

$$\begin{aligned} &\text{mean value}_{E=S_m} = \\ &= \nu_t + m - x + \frac{\sum_{k=m}^{M} (k-m) \frac{x^k}{k!} \alpha_{k,m}}{\sum_{k=m}^{M} \frac{x^k}{k!} \alpha_{k,m}} \end{aligned} \tag{29}$$

with $x = \nu_t b_{M,t}$.

As soon as at least one of the parameters $\alpha_{k,m}$ ($k = m, m+1, \ldots, M$) is greater than zero (column m of matrix **A**, (1)), the first fraction in (29) is greater than or equal to m and it is a non-decreasing function for $x = \nu_t b_{M,t} \geq 0$. First, it is trivial to see that, for $x \to 0$ the fraction approach to m (L'Hopital's rule). To check the non-decreasing property, we proceed in a similar manner to the checking procedure we use for $h_{M,\alpha}(x)$ in (22). Then, the interpretation of (29) is that when the event *success-m* is observed by the gNB at least m packets, those that successfully pursue medium access, were transmitted in the observed time slot. We add further discussions when dealing with two particular cases in Sec. VI.

Then, from (16), the construction of $P_t^{''*}(z)$ implies that, after the observation *success-m*, the distribution of $p'_{n,t}$ must be shifted m positions to the left. We do this action with the term z^{-m}, i.e., $P_t^{''*}(z) = P_t^{'*}(z) z^{-m}$. Also, we remark the fact that the event "hole", (25), (26), can be seen as a particular case of the event *success-m*, (28), (29), for $m = 0$.

When considering the *all-or-nothing* channel model, i.e., when $\alpha_{k,m} = \delta_{k,m}$ for $0 \leq m \leq M$, then $\text{mean value}_{E=S_m} = \nu_t w_{M,t} = \max(\nu_t - K_M, 0)$. In other words, if $b_{M,t}$ was one, we are certain that the number of data packets ready for transmission was m. If $b_{M,t} < 1$, some uncertainty exists about such an assumption.

When the gNB interprets as collision, i.e., one garbled slot is observed, (14) becomes, including the normalization step,

$$\begin{aligned} &p'_{n,t} = \\ &= \frac{(\nu_t)^n}{n!} e^{-\nu_t w_{M,t}} \frac{1 - \sum_{k=0}^{M} \binom{n}{k} b_{M,t}^k w_t^{n-k} (1-\alpha_{k,c})}{e^{\nu b_{M,t}} - \sum_{k=0}^{M} \binom{n}{k} b_{M,t}^k w_t^{n-k} (1-\alpha_{k,c})} \end{aligned} \tag{30}$$

where $\alpha_{k,c} = 1 - \sum_{l=0}^{k} \alpha_{k,l}$ for $k \leq M$ and $\alpha_{k,c} = 1$ for $k > M$. Its generating function is,

$$\begin{aligned} &P_t^{'*}(z) = P_t^{''*}(z) = \\ &\frac{e^{\nu_t b_{M,t} z} - \sum_{k=0}^{M} (1-\alpha_{k,c}) \frac{(\nu_t b_{M,t})^k}{k!} z^k}{e^{\nu_t b_{M,t}} - \sum_{k=0}^{M} (1-\alpha_{k,c}) \frac{(\nu_t b_{M,t})^k}{k!}} e^{\nu_t w_{M,t} (z-1)}. \end{aligned} \tag{31}$$

where the first equality in (31) comes from (17). Notice that, in opposite way to (25) and to (28), (31) is not represented by a linear combination of Poisson distributions.

The first derivative of (31) in z = 1 gives us, using the notation of $x = \nu_t b_{M,t}$

$$\text{mean value}_{E=C} = \nu_t + x \frac{\sum_{k=0}^{M} (\alpha_{k+1,c} - \alpha_{k,c}) \frac{x^k}{k!}}{e^x - \sum_{k=0}^{M} (1-\alpha_{k,c}) \frac{x^k}{k!}}. \tag{32}$$

Note that it is reasonable to assume that the fraction of (32) is positive for $x > 0$. In fact, obviously the denominator is always positive since $e^x > \sum_{k=0}^{M} (1-\alpha_{k,c}) x^k/k!$. Also, the numerator is always positive as we admit the common sense assumption that $\alpha_{k+1,c} \geq \alpha_{k,c}$, meaning that the probability of observing a collision with $k+1$ packets is not less than the probability of observing a collision with k packets.

B. Modelling Successful Packet Transmission

For arrival, we also simplify the Poisson process with rate λ_t. Then, inserting (25), (28) and (31) into (18), we observe that, in general the resulting estimated probability vector for time-slot $t+1$ is no longer Poisson, i.e., $P_t^{''*}(z) e^{\hat{\lambda}_t (z-1)} \neq e^{\hat{\nu}_{t+1}(z-1)}$. Nevertheless we can approach the resulting distribution of $P_t^{''*}(z) e^{\hat{\lambda}_t (z-1)}$ by one of Poisson for $p_{n,t+1}$ with mean value $\hat{\nu}_{t+1}$ equal to the mean value of the computed vector probability $P_t^{''*}(z) e^{\hat{\lambda}_t (z-1)}$. In other words, we obtain, by using $x = \hat{\nu}_t b_{M,t}$, that

For a hole,

$$\hat{\nu}_{t+1} = \hat{\lambda}_t + \nu_t - x + \frac{\sum_{k=0}^{M} k \frac{x^k}{k!} \alpha_{k,0}}{\sum_{k=0}^{M} \frac{x^k}{k!} \alpha_{k,0}}; \tag{33}$$

For a success-m,

$$\hat{\nu}_{t+1} = \hat{\lambda}_t + \hat{\nu}_t - x + \frac{\sum_{k=m}^{M} (k-m) \frac{x^k}{k!} \alpha_{k,m}}{\sum_{k=m}^{M} \frac{x^k}{k!} \alpha_{k,m}}; \tag{34}$$

For a collision,

$$\hat{\nu}_{t+1} = \hat{\lambda}_t + \hat{\nu}_t + x \frac{\sum_{k=0}^{M} (\alpha_{k+1,c} - \alpha_{k,c}) \frac{x^k}{k!}}{e^x - \sum_{k=0}^{M} (1-\alpha_{k,c}) \frac{x^k}{k!}}; \tag{35}$$

Then, the deriving cycle is completed.

C. The Pseudo Bayesian Procedure

Here we summarize how the procedure works. At the end of time-slot $t-1$, the gNB estimates the number of devices (new arrivals + backlogged), $\hat{\nu}_t$, that will be active in the next time-slot t. Based on (33), (34) and (35), the gNB needs to,

- inform about the permission probability, $b_{M,t} = \min(K_M/\hat{\nu}_t, 1)$, for time-slot t used by all active devices.
- if the gNB observes a success-m ($m = 0$ is a hole, while $0 < m \leq M$ indicates a success with multiplicity m) decrement the actual estimation $\hat{\nu}_t$ as,

$$\hat{\omega}_t = \hat{\nu}_t - \left(K_M - \frac{\sum_{k=m}^{M} (k-m) \frac{K_M^k}{k!} \alpha_{k,m}}{\sum_{k=m}^{M} \frac{(\hat{\nu}_t b_{M,t})^k}{k!} \alpha_{k,m}} \right); \tag{36}$$

- if the gNB observes a collision increment the actual estimation $\hat{\nu}_t$ as,

$$\hat{\omega}_t = \hat{\nu}_t + K_M \frac{\sum_{k=0}^{M} (\alpha_{k+1,c} - \alpha_{k,c}) \frac{K_M^k}{k!}}{e^{K_M} - \sum_{k=0}^{M} (1-\alpha_{k,c}) \frac{K_M^k}{k!}}; \tag{37}$$

the gNB configures,

TABLE I. M-MPR: OPTIMAL THROUGHPUT $T_{M,\alpha}(\hat{x}_M)$ WITH $\hat{x}_M = K_M$ FOR A CHANNEL WITH MAXIMUM CAPACITY; $\overline{\alpha}_m = m$, $m = 1, 2, \ldots M$; I.E. MATRIX $\mathbf{A} = \mathbf{I}$, SEE (1).

$M \rightarrow$	1	2	3	4
$T_{M,\alpha}(\hat{x}_M)$	0.36879	0.83996	1.37110	1.94238
$\hat{x}_M = K_M$	1.00000	1.61803	2.26953	2.94518

$$\hat{\nu}_{t+1} = \hat{\omega}_t + \hat{\lambda}_t \tag{38}$$

where the estimation value $\hat{\lambda}_t$ can be set equal to the number of successful packet transmitted in time-slot t.

VI. SOME PARTICULAR CASES

As illustrative examples, we discuss in this section the obtained results for two cases in M-MPR. First, the *all-or-nothing* model and second the non-perfect capture model.

A. The All-or-Nothing Model

In this case, the channel is characterized by the identity matrix of suitable dimensions, i.e., $\mathbf{A} = \mathbf{I}$, which means that $\alpha_{m,m} = 1$, i.e., $\overline{\alpha}_m = m$ for all $0 \leq m \leq M$ and $\alpha_{m,c} = 1$ for all $m > M$. These are the transmission-reception characteristics used in [16]. Then, the throughput, $T_{M,\alpha}(\nu_t, b_{M,t})$ is given by,

$$T_{M,\alpha}(\nu_t, b_{M,t}) = \sum_{m=1}^{M} \frac{(\nu_t b_{M,t})^m}{(m-1)!} e^{-\nu_t b_{M,t}} = \begin{cases} \sum_{m=1}^{M} \frac{\nu_t^m}{(m-1)!} e^{-\nu_t}; & \nu_t \leq K_M \rightarrow b_{M,t} = 1; \\ \sum_{m=1}^{M} \frac{K_M^m}{(m-1)!} e^{-K_M}; & \nu_t > K_M \rightarrow b_{M,t} < 1. \end{cases} \tag{39}$$

For $M = 1$, the SPR case, $K_1 = 1$ regardless the value of $\overline{\alpha}_1$. The maximum achievable throughput is $e^{-1}\overline{\alpha}_1 \approx 0.3679\overline{\alpha}_1$, then equal to e^{-1} (S-ALOHA) when $\overline{\alpha}_1 = 1$.

For $M = 2$, $K_2 = (1+\sqrt{5})/2 \approx 1.618034$, see (7), and the maximum throughput is ≈ 0.839962 (coincident with [12]).

For $M = 3$, $K_3 = (S_1 + S_2 + 1)/3 \approx 2.26953084$ where $S_{1,2} = \sqrt[3]{37 \pm 3\sqrt{114}}$, and the maximum achievable throughput is ≈ 1.37110.

For $M > 3$ we do not find a closed form expression, so we resort to numerical calculation as has described above.

Table I shows the maximum throughput $T_{M,\alpha}(x)$ for the *all-or-nothing* model in M-MPR for several values of M, in coincidence with the values obtained in [12].

About the Pseudo Bayesian procedure in this case we notice that in case of hole or success-m, (36) becomes (the same action for all those events),

$$\hat{\omega}_t = \hat{\nu}_t - \hat{\nu}_t b_{M,t} = \hat{\nu}_t - \min(K_M, \hat{\nu}_t) = \max(\hat{\nu}_t - K_M, 0); \tag{40}$$

and in case of collision, we have from (37)

$$\hat{\omega}_t = \hat{\nu}_t + \frac{\frac{K_M^{M+1}}{M!}}{e^{K_M} - \sum_{k=0}^{M} \frac{K_M^k}{k!}}; \tag{41}$$

The final step is achieved when (41) is inserted into (38).

TABLE II. M-MPR: $\Delta\hat{\nu}_t$ FOR A "ALL-OR-NOTHING" CHANNEL; $\overline{\alpha}_m = m$, $m = 1, 2, \ldots M$ (MATRIX $\mathbf{A} = \mathbf{I}$, SEE (1)).

$M \rightarrow$	1	2	3	4
hole, $m = 0$	-1.00000	-1.61803	-2.26953	-2.94518
success, $m = 1$	-1.00000	-1.61803	-2.26953	-2.94518
", $m = 2$	-	-1.61803	-2.26953	-2.94518
", $m = 3$	-	-	-2.26953	-2.94518
", $m = 4$	-	-	-	-2.94518
", $m = 5$	-	-	-	-
", $m = 6$	-	-	-	-
collision, $m > M$	1.39221	1.89876	2.34994	2.76516

The fraction in (41) is the bias or error of the *a priori* estimate of $\hat{\nu}_t$ evaluated at the beginning of time-slot t. At the end of this time-slot t, after the observation of the event *collision* has been taken into account, $\hat{\omega}_t$ reflects the *a posteriori* estimate of the number of packets involved in that collision. In other words, $\hat{\omega}_t$ is the corrected estimate of $\hat{\nu}_t$. Notice that for $\hat{\nu}_t \rightarrow 0$ the bias approaches to $M + 1$, i.e., $\hat{\omega}_t \rightarrow M + 1$ as expected. That is, since the system is an M-MPR with the *all-or-nothing* capability, $M+1$ is the minimum number of packets involved in one collision, very close to this value for very low traffic. Although, surprisingly, the bias in (41) decreases when $\hat{\nu}_t$ increases from zero up to $\hat{\nu}_t = K_M$ (in this interval the probability $b_{M,t}$ keeps constant equal to one) the net effect is that the *a posterior* estimate $\hat{\omega}_t$ increases when $\hat{\nu}_t$ increases, as common sense dictates. Note that it is straightforward to check that the first derivative of the bias is negative for any value of $x = \nu_t b_{M,t}$. However, it is also surprising that the bias remains constant, for values of $\hat{\nu}_t > K_M$ (in this case $b_{M,t} < 1$). That is, when $\hat{\nu}_t > K_M$ ($b_{M,t} < 1$) the bias keeps constant, equal to $1.39221, 1.89876, 2.34994, \ldots$ respectively for $M = 1, 2, 3, \ldots$. Those values are reflected in the row *collision* of Table II and are the positive bias we use for the Bayesian estimation of the number of packets involved in one collision.

Moreover, it is worth mentioning that the maximum achievable throughput per slot, $T_{M,\alpha}(\hat{x}_M)$, increases with M, as expected, i.e., starting with ≈ 0.3679 for SPR, i.e., $M = 1$, then to ≈ 0.839962 for $M = 2$, then to ≈ 1.37110 for $M = 3$, and so on. In fact, it is a linear increasing form.

Then, we conclude that it is trivial to compute the updated broadcast or permission probability $\hat{b}_{M,t}$ as has been summarized in Sec. V-C. We remark that the gNB acts according to a binary feedback, i.e., *non-collision* versus *collision*, as observed in Table II.

B. The Non-Perfect Capture Effect Model

With this model, the gNB has the chance to correctly decode one packet despite the presence of other packets in the same time slot. In general, the probability that one packet is decoded successfully depends on the number of packets involved in the collision [8]. Here we study the simple case of *non-perfect capture*, i.e., according to a noiseless channel based on [3],

$$\alpha_{0,0}=1;\quad \alpha_{m,1}=\begin{cases}1; & m=1,\\ q^m; & m=2,\dots,M\\ 0; & m>M.\end{cases};$$
$$\alpha_{m,c}=\begin{cases}0; & m=0,1.\\ 1-q^m; & m=2,\dots,M\\ 1; & m>M.\end{cases}; \tag{42}$$

so, $\overline{\alpha}_0 = 0$, $\overline{\alpha}_1 = 1$ and $\overline{\alpha}_m = q^m$ for $m = 2,\dots,M$. Equivalently, in matrix form,

$$\mathbf{A}=\left[\begin{array}{cccc|c} 1 & 0 & \dots & 0 & 0\\ 0 & 1 & \dots & 0 & 0\\ \vdots & \vdots & \ddots & \vdots & \vdots\\ 0 & q^M & \dots & 0 & 1-q^M\\ \hline \alpha_{>M,0}=0 & \alpha_{>M,1}=0 & \dots & \alpha_{>M,M}=0 & 1 \end{array}\right] \tag{43}$$

Clearly, we have the perfect capture case when $q = 1$. On the other hand, when $q \to 0$, the model degenerates to the SPR model in which $M = 1$, i.e., no capture effect. In general, a greater capture capability is obtained with large values of q (to deal with how the value of q could be estimated is out of the scope of this paper). Then, the events observed by the gNB are: *hole*, *success*-1, and *collision*, and the corresponding actions associated to (36) and (37) become as,

If a hole is observed

$$\hat{\omega}_t = \max(\hat{\nu}_t - K_M, 0); \tag{44}$$

If a success-1 is observed ($m = 1$)

$$\hat{\omega}_t = \hat{\nu}_t - \left(\hat{\nu}_t b_{M,t} - \frac{\sum_{k=2}^{M}(k-1)\frac{(q\hat{\nu}_t b_{M,t})^k}{k!}}{\hat{\nu}_t b_{M,t} + \sum k = 2M\frac{(q\hat{\nu}_t b_{M,t})^k}{k!}}\right); \tag{45}$$

If a collision is observed,

$$\hat{\omega}_t = \hat{\nu}_t + \hat{\nu}_t b_{M,t}.$$
$$\cdot\frac{(1-q)\hat{\nu}_t b_{M,t}+\sum_{k=2}^{M} q^k(1-q)\frac{(\hat{\nu}_t b_{M,t})^k}{k!}+q^M\frac{(\hat{\nu}_t b_{M,t})^M}{M!}}{e^{\nu_t b_{M,t}}-1-\hat{\nu}_t b_{M,t}-\sum_{k=2}^{M}\frac{(q\hat{\nu}_t b_{M,t})^k}{k!}}. \tag{46}$$

From previous expressions, we have ternary feedback in the non-perfect capture effect. The optimal throughput has been evaluated for several values of the parameter q, see (20);

$$T_{M,\alpha}(\hat{x}_M) = \sum_{m=1}^{M}\frac{\hat{x}_M^k}{k!}\overline{\alpha}_k e^{-\hat{x}_M}. \tag{47}$$

The results are reported in Table III.

TABLE III. M-MPR: OPTIMAL THROUGHPUT $T_{M,\alpha}(\hat{x}_M = K_M)$ FOR A CHANNEL WITH NON-PERFECT CAPTURE EFFECT ACCORDING TO (43).

$q\downarrow$	$M\rightarrow$	1	2	3	4
0.0	$T_{M,\alpha}(\hat{x}_M)$	0.36787	-	-	-
	$\hat{x}_M = K_M$	1.00000	-	-	-
0.1	$T_{M,\alpha}(\hat{x}_M)$	0.36787	0.36972	0.36978	0.36978
	$\hat{x}_M = K_M$	1.00000	1.00500	1.00533	1.00534
0.3	$T_{M,\alpha}(\hat{x}_M)$	0.36787	0.38480	0.38662	0.38676
	$\hat{x}_M = K_M$	1.00000	1.04490	1.05460	1.05578
0.5	$T_{M,\alpha}(\hat{x}_M)$	0.36787	0.41659	0.42659	0.42814
	$\hat{x}_M = K_M$	1.00000	1.12310	1.17364	1.18606
0.7	$T_{M,\alpha}(\hat{x}_M)$	0.36787	0.46786	0.50236	0.51222
	$\hat{x}_M = K_M$	1.00000	1.23183	1.38575	1.45989
0.9	$T_{M,\alpha}(\hat{x}_M)$	0.36787	0.54132	0.63352	0.68095
	$\hat{x}_M = K_M$	1.00000	1.35419	1.67287	1.94236
1.0	$T_{M,\alpha}(\hat{x}_M)$	0.36787	0.58693	0.72603	0.81671
	$\hat{x}_M = K_M$	1.00000	1.41421	1.81712	2.21336

VII. CONCLUSIONS

In this paper, we generalize the pseudo-Bayesian broadcast control algorithm when the communication system works in the environment of M-MPR in a time slot-based scheme. Up to M packets that are simultaneously are transmitted in the same time slot can be received and perfectly decoded. To that purpose, the use of capture effect, SIC, and MIMO techniques are essential to increase throughput.

ACKNOWLEDGMENT

This work is supported through Grant PID2021-123168NB-I00, funded by MCIN/AEI, Spain/10.13039/50 1100011033 and the European Union, A way of making Europe/ERDF.

REFERENCES

[1] N. Abramson, "The ALOHA system – Another alternative for computer communications," in *Proc. AFIPS Fall Joint Comput. Conf.*, vol. 37, pp. 281-285, 1970.

[2] L- G. Roberts, "Extensions of packet communication technology to a hand held personal device," in *Proc. AFIPS Spring Joint Comput. Conf.*, vol. 40, pp. 295-298, 1972.

[3] D. H. Davis and S. A. Gronemeyer, "Performance of slotted ALOHA random access with delay capture and randomized time of arrival," *IEEE Trans. Commun.*, vol. COM-28, pp. 703-710, May 1980.

[4] R. L. Rivest, "Network control by Bayesian broadcast," *IEEE Trans. Inf. Theory*, vol. IT-33, no. 3, pp. 323-328, May 1987.

[5] J. C. Arnback and W. van. Blitterswijk, "Capacity of slotted ALOHA in Rayleigh-fading channels," *IEEE J. Sel. Area Commnun.*, vol. SAC-5, no. 2, pp. 261-269, Feb. 1987.

[6] S. Ghez, S. Verdú, and S. Schwartz, "Stability properties of slotted ALOHA with multipacket reception capability," *IEEE Trans. Autom. Control*, vol. 33, No. 7, pp 640-649, Jul. 1988.

[7] R. Rom and M. Sidi, *Multiple Access Protocols*, Springer-Verlag, 1989.

[8] J. Weiselthier, A. Ephremides, and L. A. Michaels, "An exact analysis and performance evaluation of framed ALOHA with capture," *IEEE Trans. Commun.*, vol. 37, no. 2, pp. 125-137, Feb. 1989.

[9] M. Paterakis and P. Papantoni-Kazakos, "A simple window random access algorithm with advantageous properties," *IEEE Trans. Inf. theory*, vol.35, no. 5, pp. 1124-1130, Sep. 1989.

[10] S. Ghez, S. Verdú, and S. Schwartz, "Optimal decentralized control in the random access multipacket channel," *IEEE Trans. Autom. Control*, vol. 34, no. 11, pp 1553-1563, Nov. 1989.

[11] S. Verdú-Lucas, *Multiuser Detection*, Cambridge University Press, 1998, ISBN 0-521-59373-5.

[12] T. Shinomiya and H. Suzuki, "Slotted ALOHA mobile packet communication systems with multiuser Detection in a base station," *IEEE Trans. Veh. Technol.*, vol. 49, no. 3, pp. 948-955, May 2000.

[13] Q. Zhao and L Tong, "A Multiqueue service room MAC protocol for wireless networks with multipacket reception," *IEEE/ACM Trans. Netw.*, vol. 11, no. 1, pp. 125-37, Feb. 2003.

[14] R.-H. Gau, "Performance analysis of slotted ALOHA in interference dominating wireless ad-hoc networks," *IEEE Commun. Lett.*, vol. 10, no. 5, pp 402-404, May 2006.

[15] Y. H. Bae, B. D. Choi, and A. S. Alfa, "Achieving maximum throughput in random access protocols with multipacket reception," *IEEE Trans. Mobile Comput.*, vol. 13 no. 3, pp. 497-511, Mar. 2014.

[16] J. Goseling, C. Stefanovic, and P. Popovski, "A pseudo-Bayesian approach to sign-compute-resolve slotted ALOHA," in *Proc. IEEE Int. Conf. Commun. (ICC-Workshops)*, Jun. 2015, pp. 2092-2096.

[17] J.-B. Seo, H. Jin, and B. C. Jung "Non-orthogonal random access with channel inversion for 5G networks," in *Proc. IEEE Int. Conf. Inf. Commun. Technol. Convergence*, pp. 117-119, Oct. 2017.

[18] J.-B. Seo, B. C. Jung, and H. Jin, "Nonorthogonal random access for 5G mobile communication systems", *IEEE Trans. Veh. Technol.*, vol. 67, no.8 pp. 7876-7880, Aug. 2018.

[19] A. Baiocci and F. Ricciato, "Analysis of pure and slotted ALOHA with multi-packet reception and variable packet size," *IEEE Commun. Lett.*, vol. 22, no.7 pp. 1482-1485, Jul. 2018.

Low-Power Distributed Acoustic Sensor Network for Autonomous Wildlife Monitoring Using LoRa and AI for Digital Twin

Gonzalo de Miguel, Miguel Zaragoza-Esquerdo, Alberto Ivars-Palomares, Sandra Sendra, Jaime Lloret
Instituto de Investigación para la Gestión Integrada de Zonas Costeras, Universitat Politècnica de València
C/Paranimf, 1, 46730 Grao de Gandia, Spain
email: gdemig1@posgrado.upv.es, mizaes2@epsg.upv.es, aivapal@epsg.upv.es, sansenco@upv.es, jlloret@dcom.upv.es

***Abstract*—Biodiversity loss driven by climate change, habitat degradation, and anthropogenic pressures demands efficient wildlife monitoring solutions. Conventional methods are often costly, invasive, and limited in spatial or temporal coverage. Acoustic monitoring provides a non-intrusive alternative but faces challenges related to high data volumes, limited power availability, and restricted communication bandwidth in remote deployments. This paper presents a low-power distributed acoustic sensor network for autonomous wildlife monitoring, with emphasis on bird species. Each node combines an ESP32 microcontroller, a high-sensitivity digital microphone, and a Long Range (LoRa) transceiver to capture and transmit event-triggered audio. Real-time Fast Fourier Transform (FFT) analysis detects relevant acoustic activity, triggering Adaptive Differential Pulse Modulation (ADPCM) compression and LoRa-based transmission to a central receiver. The backend decodes the audio, applies the BirdNET Artificial Intelligence (AI) model for species identification, and stores results in a MongoDB database with web-based visualization. Experimental validation demonstrates high detection reliability for species with distinctive calls, confirming the system's scalability, energy efficiency, and suitability for long-term biodiversity monitoring in remote environments without continuous connectivity.**

Keywords-. LoRa, acoustic sensors, wildlife monitoring, bioacoustic, low-power IoT, BirdNET, FFT, ADPCM compression, environmental sensing, edge computing.

I. INTRODUCTION

In recent decades, biodiversity conservation has become a global priority. Impacts resulting from climate change, urbanization, intensive agricultural expansion, and noise pollution are causing a drastic decline in many animal species, such as birds and insects. These species groups are essential to ecological balance due to their role in pollination, biological pest control, and forest regeneration [1].

Therefore, passive and noninvasive wildlife monitoring has become a fundamental research and environmental management tool. Traditional wildlife monitoring methods, such as camera trapping or manual censuses, present significant limitations regarding coverage, cost, impact, and dependence on the human factor. Faced with these restrictions, distributed acoustic sensors have proven to be an effective alternative for detecting animal presence through their vocalizations or sounds associated with their activity [2]. Acoustic technology allows species to be detected even in low visibility conditions or at night, greatly expanding observation time windows.

However, one of the main challenges facing acoustic monitoring systems is data processing and transmission. Continuous audio recording generates large volumes of information, which algorithms for artificial intelligence must efficiently manage for storage, transmission, or processing. Furthermore, these systems must operate in remote areas without electrical infrastructure or conventional connectivity.

In this scenario, Low-Power Wide-Area Network (LPWAN) technologies, such as LoRaWAN have emerged as promising solutions. LoRaWAN enables data transmission over long distances (up to several kilometers) with minimal power consumption, utilizing Europe's free 868 MHz spectrum. Unlike other alternatives such as Sigfox or Narrowband Internet of Things (NB-IoT), LoRaWAN stands out for its flexibility, low cost, and open ecosystem, which facilitates its adoption in academic and industrial settings [3].

This paper proposes designing and implementing a distributed network of autonomous acoustic sensors for wildlife detection, focusing primarily on birds. The system comprises nodes based on ESP32 microcontrollers and high-sensitivity digital microphones (such as the INMP441), capable of performing real-time spectral analysis using FFT. Only in the presence of acoustic activity within the expected frequency ranges does the system trigger audio recording and compression, which is then transmitted in fragments via LoRa to another node, which then transmits to its web server. On the server, the Python backend is responsible for receiving and assembling the audio fragments, decoding them, and analyzing them using the locally running BirdNET tool. BirdNET has demonstrated high accuracy in species classification using convolutional neural networks trained with millions of acoustic recordings from birds worldwide. After species identification, the data is stored in a MongoDB database, from which interactive dashboards are generated for visualization and temporal and geographic analysis. The added value of this project lies in the combination of four key elements:

- real-time acoustic detection.
- energy efficiency.

- optimized LoRa communication.
- intelligent local processing.

This architecture allows the system to be deployed in rural or protected areas without constant maintenance or connection to mobile or Wi-Fi networks. Furthermore, its modular and open design facilitates scalability and adaptation to different ecological contexts.

The need for this type of solution is evident in the face of ecosystem management and protection challenges. In European countries like Spain, the decline of endemic species, such as the lesser grey shrike (Lanius minor) and the black stork (Ciconia nigra), requires new monitoring tools that allow for rapid and precise action. Ultimately, this project responds to a real need to improve environmental monitoring systems through low-cost, highly efficient, and minimally intrusive technologies. It provides a viable, replicable, and sustainable solution for researchers and administrators.

The main objective of this work is to design, implement, and validate a low-power distributed acoustic sensor network capable of autonomously detecting and identifying wildlife species—particularly birds—in remote environments without continuous connectivity, and to integrate the collected data into a digital twin for environmental monitoring. The proposed system combines energy-efficient hardware, optimized communication via LoRa, and artificial intelligence-based bioacoustic analysis to provide a scalable, low-cost, and minimally intrusive solution that feeds real-time information into a virtual replica of the monitored ecosystem.

The paper is structured as follows. Section II reviews related work in acoustic monitoring and low-power communication technologies. Section III details the proposed system architecture and operation. Section IV presents and discusses the experimental results. Finally, Section V provides the conclusions and outlines directions for future work.

II. RELATED WORK

Numerous solutions have been developed in recent years for environmental monitoring and bioacoustics detection of wildlife, leveraging Low-Power Wide-Area Networks (LPWAN) such as LoRaWAN due to their low consumption and broad coverage.

A notable reference is the work by FentonSigla [4], which presents a distributed acoustic monitoring system characterized by energy efficiency and flexibility. Based on ESP32 nodes and INMP441 microphones, its architecture shares similarities with this project. However, it only extracts derived acoustic parameters (such as Sound Pressure Level (SPL) or direction of arrival) instead of transmitting audio fragments for detailed analysis.

Other contributions explore complementary approaches, such as the Internet of Things (IoT) architecture by Mohandass et al. [5] for animal health monitoring and intrusion detection, or the work of Ojo et al. [6], which experimentally evaluates LoRa propagation in forest environments. Likewise, Martínez Rach et al. [7] designed a ZigBee-based bioacoustics sensor to detect the red palm weevil, focusing on pest monitoring with high accuracy in acoustic recognition.

Beyond connectivity and hardware, several methods have been proposed for acoustic activity detection and wildlife classification. Traditional threshold-based triggering [8] and more advanced spectral analysis using FFT [10] allow event-driven audio capture, although both are susceptible to false activations under noisy conditions. Recent works have also incorporated Machine Learning at the edge (TinyML), such as Tinybird-ML [9], capable of performing syllable-level bird song analysis with low-power consumption. Similarly, call density estimation methods [10] directly model the occurrence of vocalizations without relying solely on threshold events, increasing robustness in complex soundscapes. Another line of research uses animal-borne soundscapes loggers [11], enabling classification and transmission directly from tags attached to animals, particularly for underwater soundscapes.

One of the most widely adopted tools in classification models is BirdNET [13], an Artificial Intelligence (AI) based system trained with millions of recordings worldwide. BirdNET applies convolutional neural networks to spectrograms for species identification and has demonstrated high accuracy even in noisy conditions. Its ability to operate locally, without dependence on cloud services, makes it especially suitable for autonomous monitoring projects such as the one presented here. Alongside BirdNET, lightweight embedded classifiers based on spectrograms and Mel-Frequency Cepstral Coefficients (MFCCs) [10] have also been explored. However, they remain limited by the computational and memory constraints of low-power devices.

Overall, prior work highlights both feasibility and the challenges of combining low-power communication with bioacoustics analysis. The present study builds upon these contributions by integrating real-time acoustic activity detection, efficient LoRa transmission, and BirdNET-based classification into a distributed sensor network designed for long-term wildlife monitoring.

III. SYSTEM PROPOSAL

This section describes the design and implementation of the proposed distributed acoustic sensor network for autonomous wildlife monitoring. It begins with an overview of the system's architecture, detailing the hardware components, communication modules, and operating principles. Then, it explains the end-to-end workflow, from audio capture and activation strategies to compression, segmentation, and LoRa-based transmission. Subsequent subsections address the reception, decoding, and artificial intelligence-based bioacoustic analysis, followed by the storage and visualization of results.

A. System Overview

The proposed solution consists of the design and implementation of a distributed network of energy-efficient wireless acoustic sensors capable of capturing sounds emitted by wildlife, identifying relevant events in real time, and transmitting the audio fragments to a LoRa-based remote processing infrastructure. The information is processed with

artificial intelligence tools to determine the detected species and is stored in a cloud-based MongoDB database, allowing for subsequent visualization and analysis through interactive dashboards.

The developed system is based on a low-power, low-cost architecture, designed to operate autonomously in natural environments. Each sensor node comprises three main elements: an ESP32 microcontroller, an INMP441 digital microphone, and a LoRa communication module with its corresponding antenna. Their technical characteristics and function within the system are described below.

This section describes the overall operation of the system, from audio capture to results visualization. The process is divided into three main blocks: the transmitter, the receiver, and the backend, as shown in Figure 1.

The transmitting node captures audio in short windows and applies real-time FFT spectral analysis to detect acoustic activity in the target band. When a significant event is detected, the recording of the entire fragment is triggered, which is then compressed using the ADPCM algorithm. Once compressed, the file is fragmented and transmitted to the receiving node via LoRa.

At the receiver, the fragments are reassembled to reconstruct the original file. Once completed, the file is temporarily stored and automatically sent via WiFi to a web server for processing.

In the backend, the ADPCM file is decoded into WAV format. The audio is then analyzed using BirdNET to identify animal species based on their vocalizations. The results obtained and the original fragment are stored in a MongoDB database. Finally, all this information is accessible through a web panel that allows users to view, filter, and query the detected acoustic events in a structured manner.

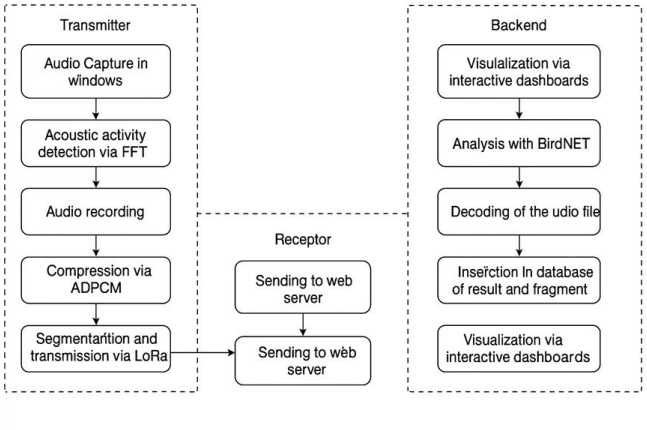

Figure 1. Block diagram of the complete system operation, from audio capture to its analysis and insertion into the database.

B. Acoustic activity detection and audio capture

The first key component of the proposed system is the process of capturing ambient audio by the sensor nodes. Since audio fragments contain critical information for detecting animal species, their recording must be selective, energy-efficient, and accurate enough to ensure its subsequent use in bioacoustic analysis processes. To this end, the nodes implement an intelligent activation strategy, meaning they are not continuously recorded but are activated only when they detect relevant acoustic activity. This decision was made after comparing activation methodologies.

After evaluating these approaches, FFT spectral analysis was selected as the activation strategy because it offers an appropriate balance between accuracy, energy efficiency, and feasibility of implementation on an ESP32-based platform. Table 1 below shows an estimated comparison between the different acoustic activation methodologies [12].

TABLE 1. QUANTITATIVE COMPARISON OF METHODOLOGIES

Method	Precision	Consumption (mA)	Latency
Sound threshold	20–40%	1–5	1 ms
FFT	60–80%	10–20	50–100 ms
TinyML	85–95%	50–100	200–500 ms
Multiple sensors	70–90%	5–15	10–50 ms

Once an acoustic event is detected in the band of interest, the node begins recording an audio fragment. To do this, an INMP441 digital microphone is connected to the ESP32 via the Inter-IC Sound (I2S) interface, allowing high-quality sampling at 16 kHz with 24-bit resolution. The recording duration is set to 20 seconds.

Once the capture is complete, the audio fragment is saved to the ESP32's flash memory using the SPI Flash File System (SPIFFS). This non-volatile memory, accessible like a small virtual disk, allows files to be preserved even after reboots or power losses. Throughout the process, the node continues monitoring the environment to verify the persistence of acoustic activity, thus avoiding storing empty or redundant fragments.

The resulting file represents the basic information unit of the system, which will subsequently be compressed and transmitted using LoRa technology for remote analysis.

C. Audio compression and segmentation

Since LoRa technology presents strict limitations regarding bandwidth and maximum packet size (for example, 51 bytes per packet in the European band with SF12), it is essential to apply data compression techniques to reduce the amount of information before transmission. In this project, the ADPCM algorithm was chosen, widely used in embedded applications due to its low computational cost and good compromise between compression and fidelity. ADPCM is a differential coding technique that predicts the value of the next audio sample based on the previous one and transmits only the quantized difference. This difference is represented with fewer bits than a full sample. In this project, a 4-bit-per-

sample encoding is used, which reduces the file size by half compared to an 8-bit linear Pulse-Code Modulation (PCM), and up to 4 times compared to a 16-bit recording.

The algorithm is efficient enough to run in real time on an ESP32 without the need for additional coprocessors, and the resulting quality has proven sufficient for bioacoustics analysis tasks such as BirdNET classification, especially in environments without excessive noise.

Once the node has captured a 20-second fragment of digital audio, the buffer is processed by the ADPCM algorithm. The result is a compressed binary file that occupies approximately 160KB.

The main advantage of ADPCM in this context is its low CPU and RAM requirements, allowing for efficient real-time implementation without compromising system autonomy. Furthermore, its simple structure facilitates encoding and decoding both on the node and in the backend.

However, it also has some limitations. Its compression is not as efficient as that of codecs such as MP3 or Opus, and it is more sensitive to noise in signals with abrupt changes. Even so, it has been experimentally verified that files compressed with ADPCM maintain sufficient fidelity for BirdNET to correctly identify characteristic vocalizations of wild species.

This file is temporarily stored in memory and later segmented into blocks compatible with LoRa payload limitations. Segmentation is performed by ensuring that each packet contains a header with minimal information such as fragment number, node ID, and end-of-transmission flag. This allows the complete file to be reconstructed on the destination server even if packets are received out of order.

D. Transmission of compressed audio via LoRa

A point-to-point wireless communication system based on LoRa technology was implemented to transmit compressed audio fragments, using two Heltec WiFi LoRa 32 V2 boards. These boards operate on the 433 MHz band, which allows for more flexible experimental use as they are not subject to the duty cycle restrictions inherent to LoRaWAN. The modulation was configured with a Spreading Factor of 7, a bandwidth of 250 kHz, and a coding rate 4/5, optimizing the balance between transmission speed and channel robustness.

The compressed file, approximately 160 kB for 20 seconds of audio, is fragmented into blocks of 220 bytes of data plus a 2-byte header. Each packet includes a sequence identifier and an end-of-transmission indicator, allowing for orderly reconstruction at the receiver. A sliding window of size three is used to improve efficiency, allowing multiple packets to be kept in flight without saturating the channel.

Based on a Heltec board, the receiver reconstructs the file over SPIFFS and acknowledges each packet using ACKs. If a packet is not acknowledged, the transmitter automatically resends it after a delay. Once all the fragments have been received, the receiving node verifies the file's integrity using a Secure Hash Algorithm (SHA-256) hash function and, if everything is correct, sends the file over WiFi to a web server for analysis.

This scheme has proven effective and robust in a laboratory environment, enabling reliable transmission without perceptible loss of quality and the need for LoRaWAN infrastructure [13].

E. Receiving, reassembling and decoding the file

Once all the compressed audio file fragments have been transmitted via LoRa, the receiving node stores them locally and reconstructs the complete file in. ADCPM format [14]. This reconstruction is based on the indices' order in each packet header, allowing the content to be assembled accurately even if the fragments arrive out of order or with an unavoidable delay.

When the End-Of-Transmission (EOF) packet is detected, the file is considered complete and is saved in the receiving node's SPIFFS file system. At that point, the file is automatically sent to a web server via WiFi, where it is decoded.

The backend, developed in Python, converts the ADCPM file into an audio file in WAV format format. To achieve this, a decoder is implemented that reverses the ADPCM compression process, reconstructing a 16-bit, 16kHz linear PCM signal. This transformation is essential to ensure compatibility with acoustic analysis tools such as BirdNET, Audacity, or Sonic Visualizer.

The decoding process is fully automated and is part of the system's continuous processing flow. This integration ensures that each recording transmitted via LoRa can be reliably stored and analyzed, maintaining the fidelity necessary for subsequent acoustic classification based on artificial intelligence.

F. Bioacoustic analysis

Once the audio file has been reconstructed and converted to the appropriate format, the next step is automatically identifying the species in the recording. To do this, BirdNET [15] was used, an artificial intelligence tool developed by the Center for Conservation Bioacoustics at Cornell University, in collaboration with the Technical University of Chemnitz. This platform is specifically designed to recognize bird vocalizations, although it can also detect other types of fauna in more advanced versions.

BirdNET works by converting the audio into spectrograms, which visually represent how the signal's energy is distributed over time and at different frequencies. From this representation, a convolutional neural network model, pre-trained with millions of recordings, can identify characteristic patterns associated with different species.

One of BirdNET's most significant advantages is that it can operate locally, without relying on cloud services. This makes it especially useful in projects like this one, which seek to maintain system autonomy and minimize the need for a permanent connection. For this work, BirdNET-Analyzer was used, a version optimized for execution on personal computers that is easily integrated into automated analysis flows.

BirdNET was chosen for several reasons. On the one hand, it is a tool widely validated in scientific work, with excellent results even in noisy environments or low-quality recordings. On the other hand, it is specifically oriented toward the acoustic analysis of wildlife, which fits perfectly with the objective of this project. Unlike other, more generic audio classifiers, BirdNET returns very detailed information: the common and scientific names of the species, the exact time it was detected, and the confidence level of the prediction.

Furthermore, as an open-source project with clear documentation and an active community, its integration has been relatively simple and offers room for improvement for future versions. However, for it to function correctly, the input files must meet certain format conditions, which have been considered from the early design stages, both in audio capture and compression and decoding.

G. *Storing and displaying results*

Once the audio file has been reconstructed and converted to the appropriate format, the next step is automatically identifying species. For this purpose, the system integrates BirdNET-Analyzer [14], executed locally on the backend server.

The tool processes the decoded audio fragments by converting them into spectrograms and applying a convolutional neural network inference. The backend records the species name, confidence score, and time stamps for each detected vocalization, storing these results with the original audio fragment in the database.

This integration allows the proposed architecture to benefit from a widely validated AI model while maintaining local autonomy, without needing cloud-based services. Furthermore, the modular design of the backend enables future integration of alternative classifiers (e.g., TinyML models or call density estimation methods) to complement BirdNET or extend recognition to other taxa.

IV. RESULTS

Throughout the development of the system, multiple tests have been performed to validate the correct operation of each module and to assess the performance of different activation methods for acoustic event detection. These tests, described in the corresponding sections, focus primarily on the evaluation of FFT-based activation and the subsequent classification of bird species using BirdNET under control conditions. The validation process employed recordings from the Xeno-Canto platform [16], played back near the microphone to simulate realistic field scenarios. Tests were conducted with three bird species—Eurasian Nightjar (Caprimulgus europaeus), Eurasian Blackbird (Turdus merula), and Mallard (Anas platyrhynchos)—using 10 or 11 audio fragments per species. These species were selected to represent different levels of vocal distinctiveness: the Eurasian Nightjar has a highly characteristic and continuous call, the Eurasian Blackbird produces more common and melodically variable songs, and the Mallard emits short, low-frequency quacks that can be like other waterfowl sounds. This diversity allows the evaluation of the system under varying degrees of classification difficulty.

Next, we will discuss the accuracy and confidence results. For the Eurasian Nightjar, the top 1 classification accuracy was 72.7 % (8 out of 11 recordings correctly identified). Confidence scores for correct detections were generally high but not uniformly near 1.0, with some variability across Figure 2 recordings. This indicates that, even after compression, segmentation, transmission, and reconstruction, the call retains enough spectral fidelity for reliable recognition in most cases, though a fraction of recordings still leads to misclassification.

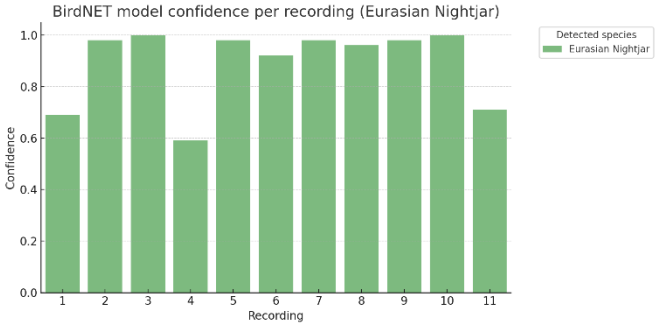

Figure 2. Species detected by recording and confidence level for Eurasian Nightjar.

In contrast, species with more common or less distinctive vocalizations, such as the Eurasian Blackbird or the Mallard, show a greater dispersion in confidence levels and, in some cases, lower results, as seen from Figures 3 and 4. This is consistent with the difficulty of automatically identifying sounds overlapping with many other species.

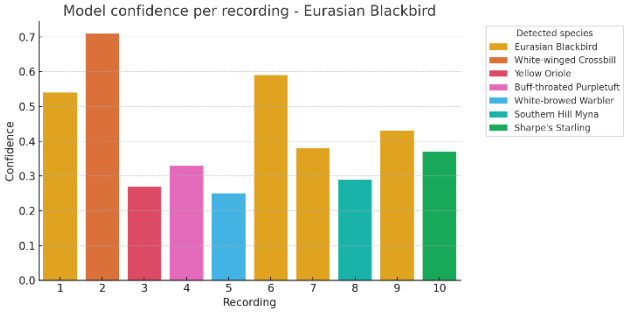

Figure 3. Species detected by recording and confidence level for Eurasian Blackbird.

Figures 3 and 4 show the results of the BirdNET model in ten analyses performed on recordings of the Eurasian Blackbird and Mallard, respectively. Each bar represents a species detected by the model in a specific recording, with its corresponding confidence level. Unlike the Eurasian Nightjar, these recordings show greater dispersion of results, with several species identified as possible candidates. In many recordings, the Eurasian Blackbird, like the Mallard, Anas platyrhynchos, appears with medium or low confidence, while in others it is outperformed by acoustic similar or commonly

occurring species. This behavior highlights the system's sensitivity to song characteristics and the fidelity of the transmission process, especially in species with less distinctive vocalizations.

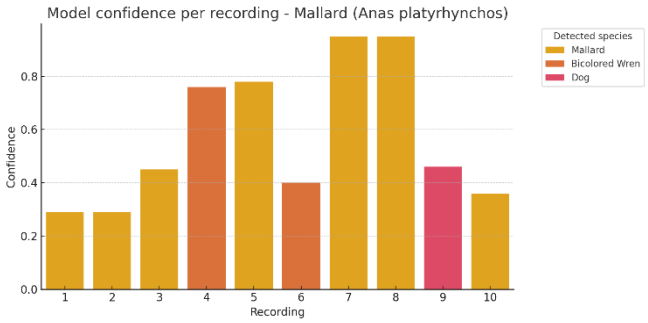

Figure 4. Species detected by recording and confidence level for Mallard Anas Platyrhynchos.

To gain a broader view of the system's performance, a comparative graph was created representing the BirdNET model's reliability for multiple bird species common in urban and natural environments in Spain. This comparison is shown in Figure 5. The graph uses violin diagrams to represent the complete distribution of confidence values obtained by the BirdNET model in the different recordings analyzed for each species. This type of representation allows us to observe the median confidence, the variability, and the density of values. The wider the curve in each area, the greater the concentration of detections in that confidence range. For the Eurasian Nightjar and Common Nightingale species, both with very distinctive and melodic vocalizations, the system presents consistently high confidence values, close to 1.0.

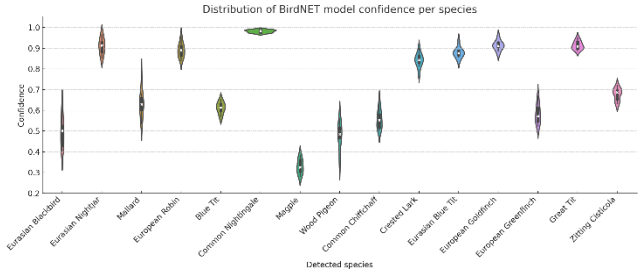

Figure 5. Distribution of BirdNET model confidence by species.

In contrast, species such as the Eurasian Blackbird, the Magpie, or the Wood Pigeon, whose songs are more common, less distinctive, or louder, present more variable and generally lower confidence levels. One of the conclusions drawn from the tests is that part of the loss in detection reliability, especially in species with less distinctive calls, is due to the implemented audio compression. To reduce the size of the transmitted fragments and adapt to the limitations of the LoRa channel and the need for higher transmission speeds, a compression scheme based on the ADPCM codec was chosen. While maintaining reasonable quality for human vocal frequencies and simple song patterns, this introduces degradations that affect the spectral integrity of certain birds' songs. It has been observed that, in complementary tests conducted with the same audio fragments but without applying prior compression, the system's reliability increases slightly, confirming that compression, although necessary for channel efficiency, sometimes negatively impacts the identification capacity of the artificial intelligence model. Furthermore, it should be noted that the system relies on pre-trained AI models (BirdNET), whose extensive database does not always offer uniform performance for all species. It is possible that some of the birds used in the tests are not sufficiently represented in the model's training set, contributing to lower reliability in certain circumstances.

These factors, combined with the acoustic characteristics of each species, explain the differences observed in the quality of the detections and should be considered when interpreting the system's results represented in Figure 6.

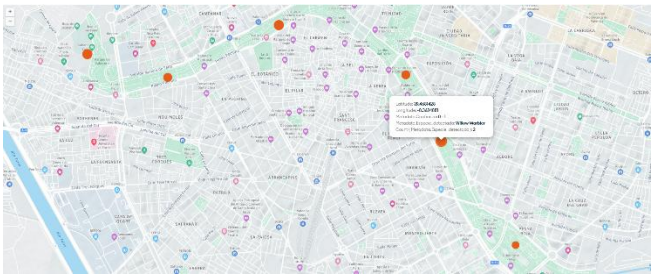

Figure 6. Location map of the detected species.

V. CONCLUSIONS AND FUTURE WORK

This work has presented the design, implementation, and validation of a low power distributed acoustic sensor system that automatically detects wildlife, focusing on birds. The integration of accessible technologies such as ESP32, digital microphones, and LoRa communication, combined with advanced artificial intelligence models (BirdNET), has enabled the development of an efficient and autonomous environmental monitoring solution. Additionally, the system features a modular design that facilitates expansion, integration with additional sensors, and advanced analysis through web platforms.

The results indicate that the system can detect and identify species with distinctive calls under real conditions, maintaining acceptable performance despite limitations imposed by ADPCM compression and the constraints of the LoRa channel. Compression, necessary to optimize transmission, introduces degradations that affect the detection of less distinctive vocalizations, representing a challenge to be addressed.

It has been shown that the AI models and dataset used do not offer uniform coverage for all species, affecting reliability in some instances.

Future directions include optimizing compression algorithms, incorporating edge AI inference in sensor nodes to further reduce data transmission, and deploying the system in natural environments powered by renewable energy to evaluate autonomy and robustness. Additionally, future work should focus on a more comprehensive evaluation of the

entire system beyond model accuracy. This includes experiments under real deployment conditions with a network of low-cost, energy-efficient sensors, reporting key performance metrics such as LoRa packet loss rates, battery lifetime (closely tied to local processing and transmission loads), and overall system reliability in the field.

In summary, this project represents a significant advance toward accessible, scalable, and automated biodiversity conservation and monitoring systems, providing innovative tools that could be integrated into large-scale environmental programs.

ACKNOWLEDGMENT

This work has been funded by the Ministry of Science, Innovation and Universities through the State Research Agency (AEI) with the projects PID2020-114467RR-C33/AEI/10.13039/501100011033, TED2021-131040B-C31.

REFERENCES

[1] A. Farina and S. James, "The ecoacoustic event: A conceptual framework for the ecological role of sounds," *Biological Conservation*, vol. 195, pp. 80–87, 2016. [Online]. Available: https://www.sciencedirect.com/science/article/abs/pii/S0006320716300118

[2] E. Vidaña-Vila, J. Navarro, C. Borda-Fortuny, D. Stowell, and R. M. Alsina Pagès, "Low-cost distributed acoustic sensor network for real-time urban sound monitoring," *Electronics*, vol. 9, no. 12, p. 2119, 2020. [Online]. Available: https://www.mdpi.com/2079-9292/9/12/2119

[3] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi, "Long-range communications in unlicensed bands: The rising stars in the IoT and smart city scenarios," *IEEE Wireless Communications*, 2016. [Online]. Available: https://arxiv.org/abs/1510.00620

[4] S. Fenton, "An ultra-low energy solution for large-scale distributed audio monitoring," *IEEE Access*, vol. 11, pp. 2156–2168, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10005289

[5] S. Mohandass, S. Sridevi, and R. Sathyabama, "Animal health monitoring and intrusion detection system based on LoRaWAN," *Turkish Journal of Computer and Mathematics Education*, vol. 12, no. 2, pp. 2397–2403, 2021. [Online]. Available: https://www.proquest.com/openview/81b1732640fc8e67432d8dc6f5cd6bf2/1?pq-origsite=gscholar&cbl=2045096

[6] M. O. Ojo, D. Adami, and S. Giordano, "Experimental evaluation of a LoRa wildlife monitoring network in a forest vegetation area," *Future Internet*, vol. 13, no. 5, p. 115, 2021. [Online]. Available: https://www.mdpi.com/1999-5903/13/5/115

[7] M. Martínez Rach, H. Migallón Gomis, O. López Granado, M. Pérez Malumbres, A. Martí Campoy, and J. J. Serrano Martín, "On the design of a bioacoustic sensor for the early detection of the red palm weevil," *Sensors*, vol. 13, no. 2, pp. 1706–1729, 2013. [Online]. Available: https://www.mdpi.com/1424-8220/13/2/1706

[8] J. Juodakis and S. Marsland, "Wind-robust sound event detection and denoising for bioacoustics," *arXiv preprint*, arXiv:2110.05632, 2021. [Online]. Available: https://arxiv.org/abs/2110.05632

[9] L. Schulthess, S. Marty, M. Dirodi, M. D. Rocha, L. Rüttimann, R. H. R. Hahnloser, and M. Magno, "Tinybird-ML: An ultra-low power smart sensor node for bird vocalization analysis and syllable classification," *arXiv preprint*, arXiv:2407.21486, 2024. [Online]. Available: https://arxiv.org/abs/2407.21486

[10] J. Liang, I. Nolasco, B. Ghani, H. Phan, E. Benetos, and D. Stowell, "All thresholds barred: Direct estimation of call density in bioacoustic sound event detection," *Frontiers in Bird Science*, vol. 4, p. 1380636, 2024. [Online]. Available: https://www.frontiersin.org/articles/10.3389/fbirs.2024.1380636/full

[11] T. Noda et al., "Animal-borne soundscape logger as a system for edge classification of sound sources and data transmission for monitoring near-real-time underwater soundscape," *Scientific Reports*, vol. 14, p. 6394, 2024. [Online]. Available: https://doi.org/10.1038/s41598-024-56439-x

[12] J. D. Rojas and P. A. Dayton, "Optimizing acoustic activation of phase change contrast agents with the activation pressure matching method: a review," *IEEE Transactions on Ultrasonics, Ferroelectrics, and Frequency Control*, vol. 64, no. 1, pp. 264–272, Jan. 2017.

[13] M. Zaragoza-Esquerdo, L. Parra, S. Sendra, and J. Lloret, "LoRa video streaming in rural wireless multimedia sensor networks," in *Proc. 2024 19th Int. Symp. Wireless Communication Systems (ISWCS)*, Jul. 2024, pp. 1–6.

[14] T. Nishitani, I. Kuroda, M. Satoh, T. Katoh, and Y. Aoki, "A CCITT standard 32 kbit/s ADPCM LSI codec," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 35, no. 2, pp. 219–225, Feb. 1987.

[15] BirdNET Team, "BirdNET-Analyzer: A tool for bird species identification from audio recordings," GitHub, 2023. [Online]. Available: https://github.com/birdnet-team/BirdNET-Analyzer

[16] Xeno-Canto Foundation, "Xeno-Canto: Sharing bird sounds from around the world," 2024. [Online]. Available: https://www.xeno-canto.org.