# Radio Access for Ultra-Reliable Communication in 5G Systems and Beyond

Kotaba, Radoslaw

*DOI (link to publication from Publisher):*
10.54337/aau473637196

*Publication date:*
2022

*Document Version*
Publisher's PDF, also known as Version of record

Link to publication from Aalborg University

*Citation for published version (APA):*
Kotaba, R. (2022). *Radio Access for Ultra-Reliable Communication in 5G Systems and Beyond*. Aalborg Universitetsforlag. https://doi.org/10.54337/aau473637196

# RADIO ACCESS FOR ULTRA-RELIABLE COMMUNICATION IN 5G SYSTEMS AND BEYOND

BY
RADOSŁAW KOTABA

DISSERTATION SUBMITTED 2022

AALBORG UNIVERSITY
DENMARK

# Radio Access for Ultra-Reliable Communication in 5G Systems and Beyond

Ph.D. Dissertation

Radosław Kotaba

Dissertation submitted March, 2022

| | |
|---|---|
| Dissertation submitted: | March, 2022 |
| PhD supervisor: | Prof. Petar Popovski<br>Aalborg University |
| Assistant PhD supervisor: | Assoc. Prof. Carles Navarro Manchón<br>Aalborg University |
| PhD committee: | Professor Mads Græsbøll Christensen (chairman)<br>Aalborg University, Denmark<br>Professor Gustavo de Veciana<br>University of Texas, USA<br>Principal Research Scientist Jakob Hoydis<br>NVIDIA, France |
| PhD Series: | Technical Faculty of IT and Design, Aalborg University |
| Department: | Department of Electronic Systems |

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-933-2

Published by:
Aalborg University Press
Kroghstræde 3
DK – 9220 Aalborg Ø
Phone: +45 99407140
aauf@forlag.aau.dk
forlag.aau.dk

Printed in Denmark by Rosendahls, 2022

# Curriculum Vitae

## Radosław Kotaba

Radosław Kotaba received dual M.Sc. degrees from the ENST Télécom Paris (Institut Eurecom) and the AGH University of Science and Technology, Krakow, in 2016. He is currently pursuing the Ph.D. degree in wireless communications with Aalborg University. From 2016 to 2019, he held an industrial Ph.D. position at Intel Mobile Communications, Denmark. His research interests include ultra-reliable low-latency communications with the focus on random access protocols and their design, non-orthogonal multiple access techniques, and device-to-device communications.

# Abstract

Ultra-reliable low-latency communication (URLLC) and massive machine-type communication (mMTC) are two flagship use cases of the fifth generation (5G) cellular networks. Yet, finding efficient communication schemes that would enable those services is challenging due to their respective requirements. In mMTC, the challenge is due to the massive number of devices which are limited by their battery-life and processing capabilities. On the other hand, in URLLC not only every packet has to be delivered with virtually 100% success rate, but also within an extremely short period of time. Achieving all of those goals is particularly challenging in the uplink which in normal circumstances needs to be fully coordinated by the base station and requires more auxiliary procedures than downlink. These procedures might not always be feasible due to the latency involved (URLLC) and prohibitively high overhead (mMTC).

This thesis proposes several novel access techniques addressing the challenges of the aforementioned use cases. The primary focus is on developing new diversity schemes for URLLC, that jointly optimize the traffic of many users. Two variants are investigated – grant-free access without feedback and scheduled access with feedback and retransmissions – which are intended for scenarios with varied level of stringency in terms of latency requirements. To address the shortcomings of the current state-of-the-art grant-free schemes, this thesis proposes the usage of specifically designed access patterns, which allow to introduce certain amount of coordination and thus provide reliability guarantees. For the scheduled access, we propose a novel technique that leverages non-orthogonal multiple access (NOMA), advanced receiver processing and power optimization to achieve highly reliable and efficient communication. Lastly, we tackle the problem of reliable massive access by considering an optimized feedback scheme which jointly encodes the acknowledgements of all active users, as well as propose a method to introduce identification an authentication capabilities to the existing concept of unsourced random access.

# Resumé

Ultra-pålidelig lavlatens-kommunikation (URLLC) og massiv maskintype-kommunikation (mMTC) er to flagskibsanvendelser af den femte generations (5G) cellulære netværk. På trods af det, er det grundet deres respektive krav en udfordring at finde effektive kommunikationsmetoder, der muliggør disse anvendelser. I mMTC skyldes udfordringen det enorme antal enheder, som er begrænset af deres batterilevetid og processeringsegenskaber. På den anden side, i URLLC skal hver pakke leveres ikke kun med praktisk talt 100% pålidelighed, men også inden for en ekstremt kort periode. At nå alle disse mål er særligt udfordrende i uplink, som under normale omstændigheder skal koordineres fuldt ud af basestationen og kræver flere hjælpeprocedurer end downlink. Disse procedurer er ikke altid praktiske på grund af den involverede latenstid (i URLLC) og uoverkommelig høj overhead (i mMTC).

Denne afhandling foreslår flere nye teknikker, der adresserer udfordringerne ved de førnævnte anvendelser. Det primære fokus er at udvikle nye metoder til at opnå diversitet for URLLC, der optimerer trafikken fra mange samtidige brugere. To varianter undersøges – bevillingsfri (”grant-free”) kommunikation uden feedback, og skemalagt kommunikation med feedback og retransmissioner – som er beregnet til scenarier med varieret niveau af stringens med hensyn til applikationernes latenstidskrav. For at afhjælpe manglerne ved de nuværende state-of-the-art bevillingsfri metoder, foreslås der i denne afhandling brugen af specifikt designede kommunikationsmønstre, som gør det muligt at indføre en vis mængde koordinering og dermed give pålidelighedsgarantier. Til den skemalagte kommunikation foreslår vi en ny teknik, der udnytter ikke-ortogonal fleradgang (NOMA), avanceret modtagerprocessering og strømoptimering for at opnå yderst pålidelig og effektiv kommunikation. Til sidst tackler vi problemet med pålidelig mMTC ved at studere en optimeret feedback-metode, som koder modtagelsesbekræftelser fra alle aktive brugere i samme besked, samt foreslår en metode til at introducere identifikation og autentificeringsfunktioner til kommunikationsmetoder baseret på oprindelsesfri kommunikation.

# Contents

# Acknowledgements

I would like to begin by expressing my gratitude to my supervisors – Professors Petar Popovski and Carles Navarro Manchón. Your guidance was absolutely vital and formative for me as a researcher, your comments and criticism – invaluable, and your support – appreciated every step of the way.

I am also deeply grateful to the rest of our research group – the Connectivity section at Aalborg University, as well as its administrative personnel, for all the help and support they provided. Special thanks to Anders, who has been my office-mate the longest and with whom we even managed to produce a few papers. I appreciate all our discussions, both the serious and the ones less so.

Next, I would like to thank people from my time as industrial PhD student at Intel Mobile Communications, particularly my manager and supervisor Tommaso, co-supervisor Nuno, and Xavier – it was a pleasure to work with you and, who knows, maybe we will have a chance to do so again.

I would also like to express my gratitude to Gustavo de Veciana, Jakob Hoydis and Mads Græsbøll Christensen for accepting to be the members of my PhD assessment committee and for dedicating their time to evaluate this thesis.

My infinite gratitude goes to my parents and my brother who have not stopped supporting me for one second, and were ready to give me a push whenever I needed it. And I needed it quite a few times. My parents don't know English so they will not be able to read it, hence, simply – dziękuję.

Then, there is another group of people, special one, in that they never really had to tolerate me or spend time with me, but, for reasons unknown, chose and keeping choosing to do so – friends.

I would like to thank Jan and Basia for letting me be a part of their life and for being one of the reasons I always look forward to coming back to Kraków. Then, the other members of our circle of friends – Artur, Klaudia, Kamil, Ola and Natalia, with whom I thoroughly enjoy spending the time – our trips, parties and festivals always turn out to be the highlight of my year. Finally, this list wouldn't be complete without recognizing my oldest friends – Bartek and Iza, whom I've known since high school and who, despite being

scattered around the world, still like to stay in touch.

And there it is.

I have to admit, it was a strange and at times difficult several years. Longer and lonelier than I had imagined too. It was not always easy to see the finish line, but eventually, little by little, I managed to get there. A few weeks before finalizing this thesis, I remembered a poem written by Robert Frost – "The Road Not Taken", fragment of which, at the risk of being cheesy, I quote below. It has been stuck in my head ever since and it seemed fitting to use it to conclude this thesis (and indeed, a significant chapter in my life).

„I shall be telling this with a sigh
Somewhere ages and ages hence:
Two roads diverged in a wood, and I—
I took the one less traveled by,
And that has made all the difference."

For better or worse.

Radosław Kotaba
Aalborg University, March 7, 2022

# Part I

# Introduction

# Chapter 1

# Introduction and Motivation

It is hard to imagine today's world without all the services and technologies that were and are being enabled by the constantly evolving communication systems. At this point, almost everyone is "connected" and, even more importantly, is used to being so virtually 24/7. This is particularly striking when we consider how fast this radical shift occurred. The mobile telephony that allowed us to call and exchange text messages while on the move has been with us (at least on a universal, accessible scale) for a bit more than two decades. Meanwhile, the broadband access through our personal devices (which materialized in the form of 3.5G networks) has become a norm over just the previous 10 years. Within that short time, the ease of access to the cellular, and more generally wireless, connectivity has changed the way we live, work and think. And if it wasn't obvious enough already how critical the digital infrastructure is to our society, certainly, the global pandemic of COVID-19 removed any remaining doubt.

## 1 5th generation of cellular networks

Currently, we are witnessing another major shift taking place as we transition to the next, fifth generation of wireless systems – 5G. It has been acknowledged early on, that 5G would bring more than just an incremental increase in the data rates — a tendency that the evolution from 2G to 4G followed. While this trend was expected to continue in a form of enhanced Mobile Broadband (eMBB), it has also been envisioned that 5G would cater to two other exciting services with radically different requirements: massive Machine-Type Communications (mMTC) and Ultra-Reliable and Low-Latency Communications (URLLC). Furthermore, because of these diverse requirements, the consensus was that a "one-size-fits-all" type of system design is no longer the right approach. Instead, more emphasis was placed on the flexibility and modularity,

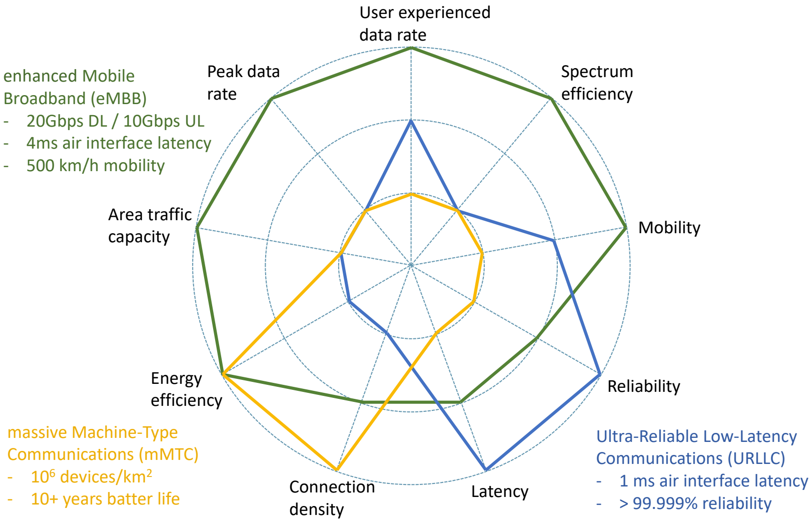


**Fig. 1.1:** The high-level overview of the 5G use cases and their requirements, inspired by the METIS II project [3].

with the intent that 5G system should eventually be able to support multitude of services, different traffic loads and various business models. In that regard, it became common among researchers and the industry to depict 5G as a multi-faceted system, similarly to the way it is shown in Fig. 1.1.

For me, a researcher who actively worked on the 5G subject, it was particularly interesting to observe how this topic was evolving, then became standardized and eventually started to materialize in the form of first deployments. At the time of writing this thesis, 5G networks are rapidly being adopted and gradually become available in more and more locations [1, 2]. According to the latter report, in some European countries like Denmark and Germany the coverage already exceeds 90% of the population. On a global scale, as per [1], it is expected that by 2027 nearly half of the mobile subscriptions will be for 5G services.

## 1.1 Ultra-Reliable Low-Latency Communication

With the evolution of the Internet of Things (IoT) (which itself was a consequence of the growing popularity of machine-type communication), new, unprecedented use cases started to attract attention of industries and consumers. To name a few:

- Factory/Industrial automation – which deals with monitoring and automated control of processes within a factory, including closed-loop control applications and robotics. Traditionally based on wired networks which limit the flexibility, these use cases could particularly benefit from enabling control over radio links.

- Electricity distribution and grid control – including applications such as control and protection of the power distribution grid, fault diagnosis, fault isolation and system restoration (FISR).
- Tactile interaction – covers use cases that involve human-system interaction, e.g. wireless control of real or virtual objects that provide a tactile feedback in a form of audio, visuals or sensation of touch. Examples include remote healthcare, manufacturing and gaming.
- Autonomous driving – focuses on automated and self driving, road safety applications and traffic management services. One of the visions is to enable fully connected cars, which will be able to jointly react to various road situations through cooperation.

The MTC nature of these use cases makes their traffic characteristics significantly different from those of more traditional, primarily human-oriented applications. Consequently, a significant effort was devoted early on to identify the specific requirements of these new use cases. For example, the METIS project [4], which run between 2012 and 2015 has determined that some of them might require latencies as low as 1 ms and reliability in the order of 99.999% (five nines) up to even 99.9999999% (nine nines)[1]. As such, URLLC became an umbrella term for the family of use cases, as well as a mode of operation within 5G, which is characterized by the transmission of relatively short (in some cases also intermittent) packets that are subject to extremely high reliability and low latency requirements. It also became clear that these requirements either cannot be fulfilled, or doing so would be very inefficient with the existing (at that time) radio access technologies. The reason was due to their design, i.e. they were created with human-centric applications in mind and as such with the goal of maximizing the throughput, rather than focusing on latency and reliability.

## 1.2 Massive Machine-Type Communication

Massive machine type communication has been created to allow a robust and cost-effective connection of *massive*, reaching millions per km$^2$, number of low-power devices, generating a huge volume of short data packets. Unlike URLLC, mMTC is not a completely new service and a limited support

[1] To give it a context, according to the current rules of the Powerball lottery in USA, the chance of hitting a jackpot is around $\sim 1 : 300{,}000{,}000$, which means that in the most demanding URLLC applications observing a failure should be 3 times less likely than winning the main prize.

for it has been introduced already in the Release 13 of 4G in the form of two standards: narrowband IoT (NB-IoT) and LTE-MTC (or eMTC) [5, 6]. However, these earlier solutions did not provide support for sufficient connection densities, which are expected to be 10x higher in 5G.

The primary use case for mMTC is the deployment of large sensor networks in various settings and environments. One example is the emergence of smart cities, where multitude of sensors are used to monitor gas, water and electricity consumption and automatically report them to the supplier. Another example is waste management where the devices could be used to detect which bins are close to being full and send the appropriate notifications, thus reducing the operating costs of waste collection teams. In agriculture, sensors can be used to report weather, moisture levels and soil analysis allowing farmers to optimize their irrigation and fertilization frequency. Finally, mMTC can have application in the traffic management on a broader, city-wide scale by providing input from a large number of cameras and sensors. This data can be then analyzed and used to help decongest busiest areas by redirecting the traffic. To enable all these use cases, radio access technologies in 5G will not only need to support high connection densities, but also provide better coverage and be energy efficient to make uninterrupted operation of batter-operated devices possible.

## 1.3 Going Beyond 5G

Despite the progress 5G networks have made towards more ubiquitous and flexible connectivity that reaches beyond simple pursuit of higher rates, at the time of writing this thesis not all of 5G's (admittedly ambitious) promises have been fulfilled. It isn't completely surprising either. Even though it is already in deployment phase, the development of 5G technology didn't come to a standstill and is constantly being improved and refined through future Releases 17 and 18 (5G-Advanced Evolution).

Lastly, it is hard to say whether it is the new use case that appear due to the technology being available, or rather it is the technology that is trying to catch up with the demand and new ideas. As with most things in life, the answer is perhaps somewhere in the middle. Regardless, after the inception of URLLC and mMTC, it didn't take long before both researchers and enthusiasts (sometimes being the same person) started to wonder how the future of wireless connectivity is going to look like. In the literature, there are already talks of mobile broadband reliable low-latency communication (MBRLLC) and massive URLLC (mURLLC), with the ambition to enable them in 6G networks [7].

# 2 Structure of the Thesis

With the above context in mind, this thesis aims at identifying and solving the key engineering challenges posed by URLLC and mMTC, as well as their potential future extensions. In particular, the focus is on enabling reliable and spectrally-efficient access in the uplink.

The rest of the thesis is structured as follows. In Chapter 2 a general problem is posed, followed by a motivating example and the definition of specific research questions guiding this thesis. The Chapter 3 is dedicated to URLLC. First, the more general background is provided including description of the scenarios and their requirements. Then, two sections follow that focus on complementary approaches – grant-free access without feedback and grant-based access with feedback and retransmissions. For each, a detailed description is given including review of the state-of-the-art, as well as summary of the contributions made in this thesis. In Chapter 4 the issue of reliable communication in massive access is treated. Again, a more in-depth background of the problem is given, followed by the review of the state-of-the-art and finally the discussion of contributions. The Chapter 5 concludes this extended introduction by summarizing all the contributions in the context of the specific research questions, as well as provide an outlook for the future research. Finally, in Part II a compilation of the publications that make up this thesis can be found.

# Chapter 2

# Problem Statement

After providing a broad background and setting the scene in the previous chapter, it is time to formally state the central problem that this thesis is considering.

*In connection with the arrival of 5G, there is a need to support new, ambitious use cases. In particular, growing popularity of mission-critical applications, requires enabling communication with extreme reliability and tight latency guarantees. Furthermore, a huge increase in the density of devices due to the emergence of MTC, calls for support of massive random access. Fulfilling these requirements with the existing radio access technologies, which were designed with traditional, primarily human-oriented applications in mind, either cannot be done, or would very inefficient. Therefore, there is a need for new solutions for the air interface that are spectrally efficient and optimized towards the specific 5G use cases.*

Before proceeding, perhaps we should first ask why exactly is that a problem? In other words, what makes it so challenging? After all, one could argue that the current communication technologies are already very efficient and close to the theoretical capacity [8, 9]. Reliability, in terms of overall low error probability, is ensured by setting up multiple stages of error correction and detection on different levels of the layered communication model and combining it with retransmission mechanism. Low latency communications, while not explicitly supported through a dedicated mode of operation, could be achieved with so-called one-shot transmissions [10]. Lastly, the solutions addressing the massive access started to appear already in LTE in its Release 13 (as discussed in more detail in the previous chapter).

The challenge, however, emerges when trying to meet a few of those objectives simultaneously and is related to the fundamental trade-offs between spectral efficiency (throughput), latency and reliability.

To better explain it, let us analyze step by step the following example of the uplink access. Clearly, in terms of latency, the best solution for a device would be to perform a single, one-shot transmission over a dedicated frequency channel as soon as it has data to send. To make such scheme reliable, however, either the power would have to be extremely high or the coding rate very low (thus making it inefficient) in order to ensure that the packet is decoded at the receiver even in the presence of unfavourable channel conditions such as shadowing or deep fades. Instead, one possibility would be to transmit the same packet multiple times over different channels (subcarriers) to obtain *diversity*. Nevertheless, most of the time single packet would prove to be enough, thus making this approach wasteful in terms of spectral efficiency. Furthermore, the device might be inactive majority of the time and not even use the pre-allocated resources. The idea then, would be to assign these resources to a group of devices to improve the utilization of the channel. At that point we introduce the probability of collisions which inevitably brings the reliability of communication down.

An alternative is to rely on retransmissions. In the simplest mechanism, the device would transmit a packet and then wait for a simple, 1-bit feedback from the receiver informing it whether decoding was successful or retransmission is required. That way, the packet is transmitted multiple times only when necessary. Clearly, such scheme is optimized towards spectral efficiency, but due to the inherent delays involved in the two-way communication - not particularly low-latency. Additionally, a single bit of feedback only allows to communicate that decoding was unsuccessful, but not how far from decoding the receiver is. As such, in response to the negative feedback another full packet would be sent, which might not always be necessary. Nevertheless, with some modifications and by restricting the maximum number of retransmissions, this mechanism could still have merit in URLLC.

Another set of challenges appear as the number of devices becomes *massive*. In that case, allocation of orthogonal resources is not just inefficient, but simply impossible. Instead, to transmit their data, devices need to contend for shared medium by relying on random access. However, due to the number of transmitters, this will inevitably cause many collisions and as a result, it may take many attempts before data is successfully delivered. On the other hand, if the devices are also battery-operated and, as such, additionally constrained in terms of energy (as is typically the case in massive access scenarios), the approach based on continued repetitions might not be feasible.

The problem of fundamental trade-offs between latency, reliability and spectral efficiency (throughput), have been captured very well in [11]. Using models based on LTE the authors provide an insight into the interplay between the aforementioned metrics in a cellular network. The main takeaway is that it is not possible to enhance all three KPIs simultaneously. In fact, if we were to improve any two of them indefinitely, the third one would inevitably go to zero. Nevertheless, using their semi-analytical framework, the authors were able to identify the main bottlenecks and give an indication of what is achievable in 5G. The solutions and enablers they propose in their paper happen to coincide with the vision of the author of this thesis and include:

- Optimizing radio access in terms of interference footprint in conjunction with using more advanced receivers (capable of interference cancellation),
- Improvements to the retransmission schemes,
- More flexible link adaptation techniques and QoS-aware scheduling taking into account reliability requirements.

## 1 Research Questions

Based on the discussion in the preceding section and having in mind the earlier example, it is time to formulate the more specific research questions.

1. How to achieve the reliability targets in a communication system with minimal signaling while avoiding excessive resource preallocation? Can the overall performance be improved by organizing better the access of many users?
2. How to jointly optimize the retransmissions of many users so that resources are used efficiently?
3. Assuming rich/extended feedback can be introduced, how to leverage it to ensure reliable and spectrally efficient communication?
4. As the original 5G use cases evolve towards future applications, can reliable communication be provided in a scenario with massive number of devices? What kind of radio access technology should be used, keeping in mind specific challenges of the massive access?

## 2 Methodology

In the development of individual contributions that make up this thesis several tools have been used, that can be broadly divided into numerical and analytical types. The first broad class includes simulations and numerical methods, which have all been implemented in MATLAB. The simulations proved to be a crucial tool throughout all of the research stages and their results can be found in virtually all of the contributions in Part II. Importantly, they served two purposes: to a) verify and confirm the correctness of the developed analytical solutions, and b) to provide complementary results in scenarios where due to the complexity of the problem no closed-form expressions (or analytical in general for that matter) could be found. The simulations relied primarily on the Monte-Carlo approach to capture the randomness in fading channel realizations, activation of users and their decoding processes. Additionally, some of the contributions involved defining and solving an optimization problem. Except for the simplest cases where the solution could be found directly based on derivatives, an optimization toolbox have been used. The particular numerical methods chosen were interior-point and sequential quadratic programming.

In terms of analytical methods, the primary focus was on the development of relevant bounds and approximations. The latter are particularly valuable in the context of URLLC, where simulations can be often challenging due to the large number of samples required. Consequently, any approximation that allows to simplify computations, or, ideally, avoid simulations altogether is a powerful tool and a meaningful contribution in itself. Furthermore, analytical expressions, bounds and approximations are extremely valuable for the understanding of the dominant mechanisms and limitations to the system performance, as well as providing theoretical guidelines for new designs and concepts.

In each paper relevant KPIs and performance measures are defined, and the results are assessed and compared to relevant benchmarks / prior solutions in order to evaluate the quality and usefulness of the proposed schemes.

This is also the right place to mention the challenges that were present throughout this study. The fundamental one stems from the very nature of the topic being researched. The notion of ultra-reliability entails that failures are rare events, which makes observing them, especially in large numbers, problematic. Consequently, in many cases where simulations were the only available tool[1], obtaining meaningful results and plots that are smooth required tens of millions of samples and days of computations. In those cases, writing a code that is efficient and which can be parallelized was an important constraint.

Another challenge related to ultra-reliability is that its primary metric of interest – outage probability – requires the knowledge of the underlying SINR distribution in order to be able to compute it. Meanwhile, distributions are inherently difficult to work with, since very often even simple operations such as sums or products of random variables do not have closed forms. This is in contrast to other metrics such as rate and spectral/power efficiency which can be captured very well using only the mean and/or variance.

[1] As already mentioned earlier, even when the analytical results could be obtained, some simulations were still necessary to first confirm their validity.

# Chapter 3

# Ultra-Reliable Low-Latency Communication

## 1 Background

The first general category of use cases that this thesis is focusing on is URLLC, which is characterized by the transmission of relatively small packets with high reliability and low latency. Before proceeding with the discussion about URLLC it is necessary to define its most relevant metrics and KPIs. Although the terms such as latency, reliability, etc. have already been used in the previous chapters and their meaning is intuitively clear, their formal definition has been missing:

- **End-to-end (E2E) latency** – the time it takes to transfer a certain amount of data from one communication interface to the other, counting from the moment it is transmitted by the source until it is successfully received at the destination. In some cases, the definition can further specify certain packet size, e.g. 32 bits.

- **Air interface latency** – a component of the E2E latency, which is restricted to the physical layer procedures. It includes queuing delay, processing time (encoding/decoding, modulation/demodulation, channel estimation etc.), propagation latency, retransmission delay.

- **Reliability** – the percentage of successfully delivered packets among all the packets that were transmitted. To be considered successful, the packet must be decoded within a deadline, i.e. adhere to the latency constraint.

- **Communication service availability** – the percentage of time communication service with specified QoS (i.e. given latency, reliability and throughput) is provided, out of the total time it is expected to be provided. In the literature, availability is also sometimes defined in terms of percentage of the area rather than time.

- **Rate** – the raw information rate, i.e. excluding overhead and redundancy, at which data is transmitted measured in bits per symbol or bits per channel use.
- **Throughput** – average amount of successfully delivered data measured in bits per second.

In this thesis, as well as papers that constitute it, the above definitions apply[1].

When speaking about URLLC, it is important to realize that it encompasses many different scenarios (some of which have already been mentioned in Chapter 1). While the figures such as 1 ms latency and 99.999% (five nines) reliability became almost synonymous with URLLC at this point, the actual requirements are much more diverse. Furthermore, it should be stressed that the 1 ms latency typically refers to the air interface, not E2E latency. The Table 3.1, which has been compiled based on [10, 12, 13], is an overview of the most relevant use cases foreseen in 5G and lists their requirements[2]. As can be observed, many of the use cases do not actually require single-digit latencies (although it should be stressed that these correspond to the E2E delay, not air interface one which would have to be lower). In fact, some of the scenarios, like fault location in a factory or UAV control, can tolerate up to $\sim$ 100 ms, which technically is achievable even in 4G albeit not with sufficient reliability [14]. Nevertheless, these are still beyond the capabilities of 4G networks when taking into account other requirements which have to be simultaneously fulfilled. A particularly interesting application is the replacement of the wired connections by wireless links. This is of primary interest to the industry and factories of the future, where physical connections make full automation and deployment of autonomous robots more challenging. However, based on the requirements, which involve extremely low latency and unusually high (for URLLC) data rates, this use case is more in line with the 6G concept of MBRLLC.

[1] The distinction between reliability and availability can be sometimes confusing as different authors tend to provide different definitions. For example, in some 3GPP documents the following description is given [12]: "reliability covers the communication-related aspects between two nodes (here: end nodes), while communication service availability addresses the communication-related aspects between two communication service interfaces.". In other words, according to this definition reliability focuses on the connectivity between the lower layers (PHY/MAC), while availability refers to the service/application.

[2] We note the difference in the way reliability is denoted in the Table 3.1. 3GPP tends to express it in the unit of the mean time between failures, while other sources prefer to use the more traditional definition, i.e. percentage of the successful packets.

**Table 3.1:** URLLC use case requirements

| Use case | Max. E2E latency | Reliability | Availability | User experienced throughput | Connection density |
|---|---|---|---|---|---|
| Control of Automated Guided Vehicles | 5 ms | 99.999% | | 100 kbps DL control,<br>3-8 Mbps UL video | |
| Robotic tooling - motion control | 1-2 ms | ~ 10 yrs | $10^{-6}$ | | |
| Wired to wireless link replacement | < 1 ms | ~ 10 yrs | $10^{-6}$ – $10^{-8}$ | 50-500 Mbps | |
| Closed-loop control in process automation | 10 ms | ~ 1 yr | $10^{-6}$ – $10^{-8}$ | | |
| Differential protection | 5-15 ms | 99.999% | $10^{-5}$ | 2.4 Mbps | 10-100/km$^2$ |
| Fault location identification | 140 ms | 99.9999% | | 100 Mbps | 10/km$^2$ |
| Electricity distribution - high voltage | 5 ms | 99.999% | $10^{-6}$ | 10 Mbps | 1000/km$^2$ |
| Electricity distribution - medium voltage | 40 ms | 99.9% | $10^{-3}$ | 10 Mbps | 1000/km$^2$ |
| UAV command and control | < 100 ms | 99.999% | | | |
| Augmented reality | < 10 ms | ~ 1 month | $10^{-5}$ | | |
| Intelligent transport systems – infrastructure backhaul | 30 ms | 99.999% | $10^{-6}$ | 10 Mbps | 1000/km$^2$ |

## 1.1 URLLC Enablers

As URLLC have been gaining attention, a number of researchers as well as standardization bodies started to look into different areas where the improvements could be made.

One of the most prominent factors in terms of latency reduction was the introduction of new numerology [15, 16] and flexible frame structure including mini-slots [17]. The former introduces additional modes (values) for the subcarrier spacing, i.e. 30, 60, 120, 240 kHz, in addition to the 15 kHz used in LTE. This alone allows to shorten the duration of the slots by a factor of 2, 4, 8 and 16 respectively. Meanwhile, the new frame structure, and the concept of mini slots in particular, provides additional degree of flexibility as it allows to use slots that are shorter than 14 OFDM symbols (unlike in LTE, where it is fixed). This has a direct impact on the latency, but also facilitates the scheduling, which can be performed with higher granularity. The two techniques combined play a key role in reducing the overall latency of the system and bringing it closer to the URLLC targets [17]. The second important work item on the road towards 5G was the simplification of the protocols. Since most of the auxiliary procedures rely on some form of the handshake between UEs and the BS, there have been many efforts to streamline this step and reduce the amount of signalling. As a result, new, more lightweight protocols for the handovers [18], random access [19] and grant acquisition [20] have been proposed. In that context, particularly important is the introduction of the new transmission mode called grant-free (GF) access [21], which enables the devices to send their packets virtually at will and with no extra delay. It is meant to extend the semi-persistent scheduling [22] functionality introduced in LTE by providing users with even more flexibility and freedom. Transmission techniques utilizing grant-free access are one of the primary topics of this thesis and a subject of the dedicated Section 2 below. Lastly, there are research efforts that seek latency improvements at the higher layers, such as optimization of the scheduling algorithms [23], mobile edge computing and offloading [24], and network edge caching [25] to name a few.

Complementary to the latency reduction techniques, there are a number of methods that focus on improving the reliability, which is typically achieved by providing some form of diversity. The first category are the spatial diversity schemes. Among them, the most popular and this point indispensable technology, is the usage of multiple antennas, i.e. MIMO [26], and its extension Massive MIMO, which leverages the effect of channel hardening [27]. Another technique (that can also be viewed as a generalization of MIMO) is the multipoint transmission where multiple access points or base stations may transmit/receive the same packet from different locations in a coordinated manner [28]. Another category are the diversity schemes which strive to overcome the negative effects of fading channels and achieve reliability through packet repetition in time and/or frequency. They are motivated by the fact that single-shot transmissions have been shown to be largely inefficient (DL [29], UL [30]), although possible if the BS is equipped with many antennas exhibiting low correlation [10]. As such, all three diversity techniques – spatial, time and frequency – play an important role in 5G and are expected to remain vital in 6G.

The transmit diversity category can be further divided into two broad groups: feedback-less and feedback-based schemes. The fundamental difference between the two is that in the latter the additional transmissions are performed only after it has been explicitly signalled, with the so called negative acknowledgement (NACK), that the decoding was not successful. Among the feedback-based diversity techniques the most widely used is the hybrid automatic repeat request (HARQ) [31], which combines forward correction coding (FEC) and simpler ARQ. In HARQ, each transmission is encoded with certain amount of redundancy that allows to recover some of the erroneous bits (FEC). Should that prove insufficient, retransmission can be requested, but, unlike in legacy ARQ, the prior (unsuccessful) transmission is not discarded and instead soft combined with the new one to improve the chance of decoding. Depending on the content of the retransmission and, consequently, combining method, we distinguish Chase Combining (CC) and Incremental Redundancy (IR) HARQ. In the former, each transmission contains the same symbols and combining is performed in the power domain, while in the latter, the packets provide different sets of coded bits, or redundancy versions (RVs), which are then concatenated[3]. Clearly, due to the inherent two-way communication, HARQ (and grant-based schemes in general) are characterized by a higher latency, which clashes with the URLLC requirements. Nevertheless, as remarked at the beginning of this chapter, some URLLC use cases have more relaxed latency constraints. Furthermore, 5G NR provides the tools to reduce the latency through other means, i.e. numerology and frame structure enhancements. Meanwhile, the fact that additional transmissions in HARQ are performed only upon request make the approach spectrally efficient and thus, attractive from the system design perspective and worth considering also in URLLC. It should also be pointed out that feedback-less schemes are especially suitable for grant-free access, while feedback-based retransmissions like HARQ, combine naturally with grant-based type of operation. The HARQ schemes are another main topic of this thesis and are further discussed in Section 3.

Last important enabler are the advanced signal processing techniques at the receiver, particularly the successive interference cancellation (SIC) [32]. The ability to iteratively remove strong signals as they are being decoded significantly increases the performance of the systems based on non-orthogonal access. In the case of grant-free schemes, SIC contributes to improved collision resolution capabilities and thus allows to support higher traffic loads. On the other hand, in the non-orthogonal grant-based access it decreases the latency by reducing the need for retransmissions.

## 2 Diversity Transmission Schemes in Grant-Free Access

When dealing with applications that call for the most stringent air interface latency constraints such as $0.5 - 2$ ms, there is simply no time to perform

[3] While the distinction between CC and IR have been originally proposed in the context of HARQ, it should be noted that the same can be applied in the feedback-less schemes. I.e., the multiple diversity transmissions can be either exact copies or contain different redundancy versions.

any of the auxiliary procedures, i.e. transmission of the scheduling request, reception of grant with resource allocation followed by potentially multiple rounds of retransmissions. In those cases, the GF access based methods are the only feasible solution.

In its essence, grant-free is a type of random access that originates from one of the most prominent communication protocols, namely the ALOHA system [33], and especially its slotted version [34]. In its most basic form, the solution relies on dividing the communication channel into transmission opportunities (slots), which are used by the population of randomly active devices to send their data. Furthermore, only the singular transmissions can be decoded, i.e. the collisions result in a loss of all the packets involved. Such a scheme, while conceptually simple and elegant, does not provide a very high spectral efficiency, hence there have been many efforts to extend and improve it. Among the most prominent are the Contention Resolution Diversity Slotted ALOHA (CRDSA) [35] and Irregular Repetition Slotted ALOHA (IRSA) [36]. Both of them rely on the same mechanism, namely, each packet is transmitted multiple times to a) provide diversity and b) (even more importantly) to enable successive interference cancellation. The difference is that in CRDSA the number of repetitions is fixed, while in IRSA it is a random variable governed by parameter called degree distribution. In addition to the aforementioned, which both rely on repetition of the packet replicas, i.e. exact copies like in CC-HARQ, there is a variant called Coded Slotted ALOHA (CSA) [37] where transmissions are instead distinct coded versions similar to IR-HARQ.

In all of the above techniques, the choice of the slots used for transmissions is fully random. Differently to this, some authors (including the author of this thesis) postulate the use of specifically designed access patterns that are pre-assigned to the users. Such a solution allows to introduce some degree of coordination between users and thus provide certain performance guarantees (in terms of reliability, throughput or latency). One of the first contributions that have investigated the pattern design and assignment in the context of GF access is [38]. The combinatorial approach presented therein have inspired future works such as [39–41]. In particular, the idea of using Steiner systems in [40], i.e. a specific code design that controls the amount of collisions and interference, have been further explored within the scope of this thesis. Alternative approaches, such as the one presented in [41], draw the inspiration from low density parity check (LDPC) codes.

The capability of various grant-free based schemes to support URLLC in 5G constitute a large body of research on its own. In [42] the authors investigate the impact of collisions on the achievable latency and reliability performance in a simple repetition-based GF access. Interestingly, they show that such a scheme struggles to fulfill the URLLC requirements when the traffic is aperiodic and validate their findings using a system level simulator that models the operation of 5G NR. Another extensive work on the topic heavily leveraging system level simulations is [43], where the author evaluates a range of grant-free (primarily repetition based) techniques with an emphasis on accurate modelling of authentic scenarios and use cases. Worth mentioning is also [44], which jointly compares different transmission techniques in the context of URLLC bridging the gap between the grant-free and grant-based concepts. Namely, the authors consider the baseline repetition coding as well as its proactive and reactive variant. In the proactive scheme, the device also transmits the replicas a preconfigured number of times in a GF manner, however the BS has the option to send "early termination" signal if it manages to decode the packet before the last replica is transmitted. On the other hand, in the reactive scheme only the first copy of the signal is transmitted without prior scheduling, while the rest follow a typical HARQ process. The contribution show that with strict latency constraints, the schemes using less signalling (i.e. repetition coding and proactive) are generally superior.

It should be noted, that the GF access methods implemented in 5G are so far limited to simple repetition schemes. Nevertheless, GF as an actual mode of operation in mobile networks is relatively new and it is expected that more advanced solutions, for example utilizing special pattern design described earlier, might be implemented in future releases as well as Beyond 5G.

Lastly, some practical comments. As already mentioned, in GF access the devices are allowed to transmit their packets virtually at will. There are, however, some caveats.

Firstly, it is a common assumption that GF access is limited to a designated, typically small portion of bandwidth, rather than be allowed to occur anywhere. Indeed, if this was not the case, the completely unpredictable GF packets could prove really disruptive to the rest of the system and its services, such as eMBB. Moreover, the fact that GF signals could appear anywhere at any time would entail prohibitively high complexity for the receiver, which would be forced to scan all of its operating bandwidth[4].

Secondly, it would be unreasonable to assume that the pool of resources is fixed. Over time, the number of URLLC devices is likely to fluctuate, thus requiring adjustments to the size and location of the pool.

Lastly, there is a question of admission control and the fact that BS, who in the end is the one responsible for providing ultra-reliability, needs to be aware of the number of devices using the GF pool and their traffic characteristics. The reliability guarantees are not absolute, i.e. they apply only under certain conditions. For example, for a given pool size the error probability below

[4] Nevertheless, the concept of coexistence of the eMBB and URLLC have also been explored in the literature, where the high priority URLLC traffic is either superposed on top of the eMBB packets or puncturing (preemption) is used. Both downlink [45] and uplink [46] scenarios have been considered, however the former is generally much easier to implement and coordinate.

$10^{-5}$ can be ensured only when the mean number of transmitting (active) devices is below certain threshold.

This already implies that even in the case of GF, there needs to be some level of coordination and signalling between the UE and the BS that have to occur prior to the transmission. In particular, to the best of the author's knowledge, there is no dedicated signalling or specific field in the messages broadcasted by the cell such as master information block (MIB) or system information block (SIB), that would allow the device to locate the pool of GF resources and immediately start using them without prior handshake. All of this works in favor of the pre-assigned patterns described earlier, since their assignment can be incorporated into the already existing auxiliary procedures without posing additional overhead.

## 2.1 Summary of Contributions

Within this PhD project, the following papers on the topic of diversity transmission schemes in grant-free access have been written:

**Paper A:** Radosław Kotaba, Carles Navarro Manchón, Tommaso Balercia and Petar Popovski, "Uplink Transmissions in URLLC Systems with Shared Diversity Resources", IEEE Wireless Communications Letters, Vol. 7, No. 4, pp. 590–593, 2018. (*published*)

**Paper B:** Radosław Kotaba, Carles Navarro Manchón and Petar Popovski, "Enhancing Performance of Uplink URLLC Systems via Shared Diversity Transmissions and Multiple Antenna Processing", 53*rd* Asilomar Conference on Signals, Systems, and Computers, pp. 1409–1415, 2019. (*published*)

**Paper C:** Christopher Boyd, Radosław Kotaba, Olav Tirkkonen and Petar Popovski, "Non-Orthogonal Contention-Based Access for URLLC Devices with Frequency Diversity", IEEE 20*th* International Workshop on Signal Processing Advances in Wireless Communications (SPAWC), 2019. (*published*)

**Paper D:** Radosław Kotaba, Roope Vehkalahti, Čedomir Stefanović, Olav Tirkkonen and Petar Popovski, "Deterministic Patterns for Multiple Access in Latency-Constrained Ultra-Reliable Communications", IEEE Transactions on Communications. (*submitted*)

The papers that comprise this line of work are closely related, however each explores different aspects of the grant-free access.

Chronologically the first in this group, Paper A coins the term *transmissions with shared diversity resources* (TSDR) and represents certain guiding principle in the design of the GF access of many users. The idea, which materializes in different forms in contributions A-D, is to introduce some coordination among users by imposing a structure on their access patterns. The aim is to provide certain reliability guarantees through time/frequency diversity while controlling the interference. In Paper A this is done specif-

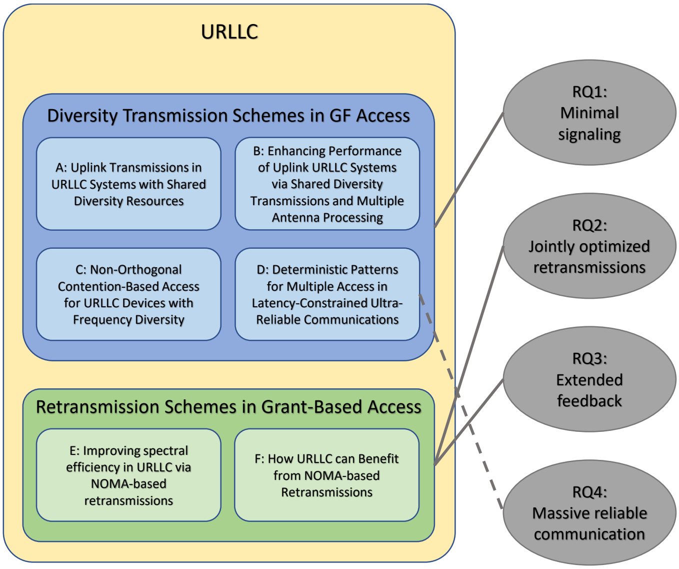

URLLC
Diversity Transmission Schemes in GF Access
A: Uplink Transmissions in URLLC Systems with Shared Diversity Resources
B: Enhancing Performance of Uplink URLLC Systems via Shared Diversity Transmissions and Multiple Antenna Processing
C: Non-Orthogonal Contention-Based Access for URLLC Devices with Frequency Diversity
D: Deterministic Patterns for Multiple Access in Latency-Constrained Ultra-Reliable Communications
Retransmission Schemes in Grant-Based Access
E: Improving spectral efficiency in URLLC via NOMA-based retransmissions
F: How URLLC can Benefit from NOMA-based Retransmissions
RQ1: Minimal signaling
RQ2: Jointly optimized retransmissions
RQ3: Extended feedback
RQ4: Massive reliable communication

**Fig. 3.1:** Schematic representation of the contributions in the area of URLLC and their relation to research questions

ically by dividing the access frame (composed of $M$ slots) into two parts – one with $N$ resources and one with the remaining $M - N$. In the first part, each of the $N$ devices have a dedicated, orthogonal slot, while the additional *diversity transmissions* are performed over the non-orthogonal second part. The rationale behind this approach was to ensure that the receiver obtains at least one uninterfered copy of the signal that would facilitate the usage of SIC. The Paper A was also meaningful to the author in another way, as it provided him with an important framework that would be reused in future contributions. Namely, by treating different slots used for transmission as *virtual* antennas, the problem of decoding the signals can be perceived as a form of MU-MIMO, which in turn gives access to a vast range of available analytical and semi-analytical tools. In particular, based on [47, 48] it was possible to investigate the performance of the TSDR scheme in the presence of channel estimation errors (with MMSE and ZF receiver respectively) and their impact on the SIC.

Paper B extends the prior work in several important ways. Firstly, it generalizes the concept of TSDR by dividing the available channel resources into low and high contention portions (that way, the arrangement from Paper A where each device has one orthogonal slot and multiple diversity transmissions becomes a special case). Secondly, as a consequence of the proposed division into low and high contention slots, we investigate the idea of using unequal transmit powers. Lastly, we extend the model by considering mul-

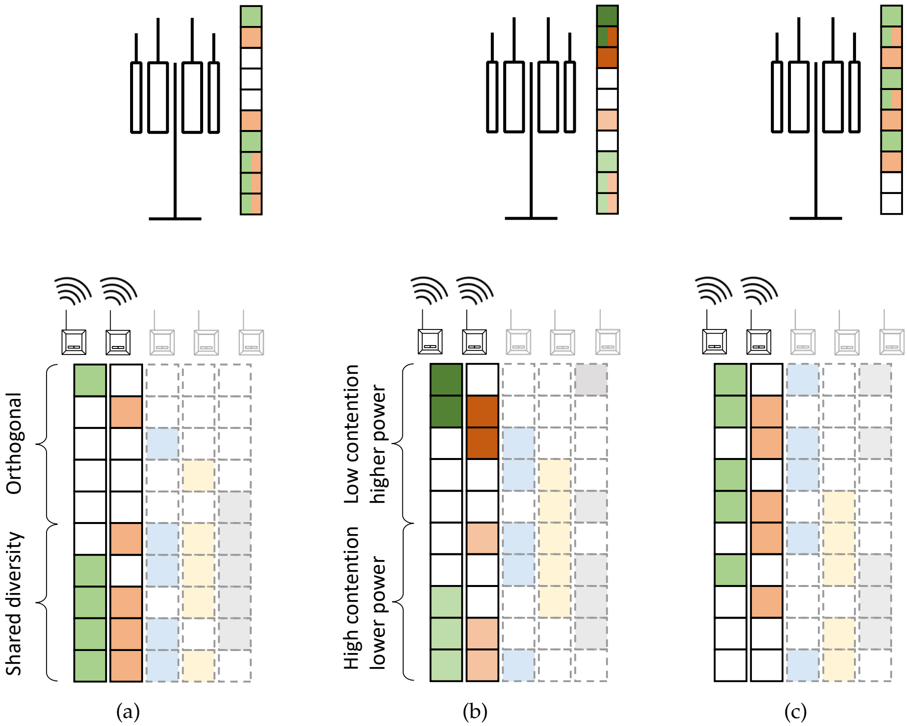


**Fig. 3.2:** Comparison of three access pattern design approaches a) TSDR scheme, b) Generalized TSDR with low and high contention and unequal power allocation, c) $(3, 5, 10)$-Steiner system.

tiple receive antennas at the BS. One of the most important contributions of this work involve developing a relatively simple, semi-analytical expression for the outage probability and its Chernoff bound.

Lastly, the Paper C and its journal extension Paper D focus more on the specific access pattern design, which was inspired by the combinatorial [38] and interference cancelling code design [39]. The approach relies on $(t, K, M)$ Steiner system which is a constant weight $K$-out-of-$M$ code having a property that any two codewords (in this case equivalent to access patterns) share at most $t-1$ positions, thus limiting the number of collisions and interference. The receiver considered in that work was based on maximum ratio combining (MRC) with and without SIC, for which we were able to derive very accurate analytical approximations and bounds. Overall, the Steiner system has been shown to significantly outperform an approach based on random selection in terms of raw performance (outage probability and resource efficiency), in addition to simplifying the system design. Specifically, the introduction of the access patterns means that the maximum contention in each slot is know. This in turn allows to generate sufficient number of orthogonal pilot sequences and pre-allocate them to the users in a way that ensures there will be no collisions among pilots. Such property provides another compelling argument in favor of implementing GF access based on pre-assigned patterns.

To visualize the difference between strategies considered in papers A-D, in Fig. 3.2 a toy example is shown with 2 active users (out of 5 in total) accessing a frame consisting of 10 slots and employing 5-repetition coding.

**Future work:** A natural extension that could be considered in future work would be to jointly compare the three schemes.

# 3 Retransmission Schemes in Grant-Based Access

As already established, not all URLLC use cases are characterized by latency constraints that are so strict as to completely preclude more traditional grant-based access with retransmissions. One embodiment of such retransmission schemes is HARQ, which was first described in early 1980's [31] and has been a part of the standards since its introduction in 3G's high speed packet access (HSPA) [49]. To this day, HARQ mechanism remains one of the most spectrally efficient ways of providing reliability through diversity. Due to this, there have been many efforts to accommodate it in URLLC, with standardization bodies primarily focusing on enhancements at the physical layer and without modifying the protocol itself. These include: new numerology, shortening of the frames, as well as imposing more aggressive timing budgets for the processing done by the UEs and BS. As a result of these enhancements, it is expected that even in URLLC (with the exception of the most stringent use cases) it will be feasible to perform a small number of 1 to 2 retransmissions[5]. A good overview of the timing issues and an example of HARQ scheme in URLLC can be found in [50].

Considering all the advantages of retransmissions and the flexibility they bring in terms of system design, it is not surprising that the topic continues to attract the attention of the researchers. Consequently, there are many contributions among the state-of-the-art that are worth mentioning. Here, I would like to focus on some of the earlier publications on the HARQ topic which have inspired contributions that were developed within this PhD project. Although most of the works below treat about general HARQ enhancements not restricted to URLLC, they are still very useful and the concepts presented therein continue to be relevant.

- Early NACK [51] — this term corresponds to the family of solutions that try to predict (in this particular case based on the soft inputs and

[5] For comparison, in LTE the default number of allowed retransmissions is 8.

log-likelihood ratios of individual bits) with high accuracy the success/failure of the decoding before the final result is known. Based on such prediction the retransmission request can be sent early, i.e. without waiting for the result of full turbo-decoding procedure, thus potentially speeding up the process by the amount of necessary processing time.

- Enriched feedback [52] — by replacing binary success/failure type of feedback with an enriched one that consists of multiple bits, it is possible to inform the transmitter how far the decoder was from the success. This, in principle, allows to adapt the code rate for the following retransmission so that just the right amount of redundancy is provided. With such approach, it is possible to increase the throughput of the system, and by combining several retransmissions in one packet, also decrease the system-level latency.

- Backtrack Retransmission (BRQ) [53] — HARQ implementation that allows to increase the throughput by sending only the necessary amount of redundancy bits and chaining the decoding process of subsequent packets, at the cost of higher latency. This contribution considers practical aspects of enriched feedback, which has a finite length and is used to report SINRs of previous packets via vector quantization.

- Root Cause Aware HARQ [54] — another interesting way of using enriched feedback is to indicate explicitly the cause of the decoding failure. The authors propose an algorithm in which the receiver sends as a negative acknowledgement (NACK) the ID of the dominant interferer. Then, the network uses this information to mute the interfering cell by forcing it not to schedule any traffic in the subframe intended for retransmission.

- Cooperative diversity [55, 56] — the authors of these contributions postulate the use of relaying (with extensions such as rate adaptation) in order to increase the probability of decoding. The communication model involves broadcasting of the message to both the destination and the relay node. Then, in case of the NACK, the message can be retransmitted from either the source, or relay or both.

The above works are meant to complement the overview of the more recent publications which can be found in paper F.

## 3.1 Summary of Contributions

Within this PhD project, the following papers on the topic of retransmission schemes in grant-based access have been written:

**Paper E:** Radosław Kotaba, Carles Navarro Manchón, Nuno Manuel Kiilerich Pratas, Tommaso Balercia and Petar Popovski,
"Improving spectral efficiency in URLLC via NOMA-based retransmissions", 2019 IEEE International Conference on Communications (ICC), pp. 1–7, 2019. (*published*)
**Paper F:** Radosław Kotaba, Carles Navarro Manchón, Tommaso Balercia and Petar Popovski,
"How URLLC can Benefit from NOMA-based Retransmissions", IEEE Transactions on Wireless Communications, Vol. 20, No. 3, pp. 1684–1699, 2021. (*published*)

The second area of research, which concentrated on grant-based uplink access, took as a starting point the fact that in some URLLC applications a small number of retransmissions might be feasible. With that in mind, in contributions E, F we set out to devise an HARQ scheme with limited number of rounds that would focus on providing high reliability guarantees, while still being spectrally efficient. We achieved this by exploiting two mechanisms. Firstly, we allowed the uplink packets to be scheduled non-orthogonally following a power-domain NOMA paradigm. Secondly, we assumed that it is possible to send an extended feedback (i.e. more than one bit) in the downlink. In the scheme proposed in Paper E, the BS pairs retransmissions (if there are any) with packets of other users and schedules them on the same resource. The rationale behind this approach is that the receiver is able to combine retransmissions with earlier versions of the packets (either in power (CC) or code (IR) domain), so the retransmissions themselves typically do not need to be performed with high power. Consequently, they are not overly disruptive to the other signals when transmitted non-orthogonally. As one of the main contributions of this work we developed an analytical formula, which allows to calculate the error probability for each of the two non-orthogonal UEs assuming SIC is used and taking into account previous unsuccessful packet transmissions. Based on that formula, we solve numerically an optimization problem to find minimum powers that fulfill imposed reliability target. The extended feedback is then used to inform the UEs how much power they should use in their next transmissions[6].

In Paper F, we considered a generalized approach where any two packets can be paired, not only initial transmissions with retransmissions. Furthermore, we developed an analytical method of determining optimal error targets for the individual HARQ rounds (which are then used to find the minimum transmit powers as described before). We also extended our earlier work which focused solely on the case with no CSI, and analyzed the

[6] We also assumed that BS communicates the resource allocation as a part of the feedback, but since this is a grant-based access, the scheduling information would be provided in the control signal anyway.

scenario where instantaneous channel realizations are known. In the latter, the errors become noise-dominated and finite-blocklength effects have to be considered.

**Future work:** The Paper F considers a very comprehensive system model that includes scheduling process at the BS. The approach is aimed to be optimal and checks all the possible configurations to ultimately select the one minimizing total power. However, the drawback of this solutions is its complexity. In the future work, it would be interesting to investigate other heuristic and suboptimal approaches to determine the scheduling, e.g. based on machine learning methods, and compare with the baseline.

# Chapter 4

# Massive Access

## 1 Background

Massive Access, or massive machine type communications (mMTC), is another major category of use cases that became a staple of 5G. Unlike URLLC, however, it does not generally impose very strict latency and reliability requirements. Instead, the main challenge stems from the need to provide connectivity to a massive number of densely deployed devices. According to [57] and [58] the figures considered could be as high as 300,000 simultaneous connections in a single cell and a density of more than 1,000,000 units/km$^2$ respectively. On the other hand, mMTC is characterized by small packet sizes (payloads) and infrequent transmissions, such that the average number of simultaneously active users is only a small fraction of the total population. Additional constraint is that mMTC devices are typically simple, battery-operated, low-power terminals whose lifespan should exceed 10 years. This battery life requirement is based on the average activity of 200 bytes in UL and 20 bytes in DL per day, with individual packets consisting of few tens of bytes [59].

The consequence of these requirements is that the signalling overhead of the traditional protocols becomes comparable, or even dominant component of the overall communication. As such, a significant part of the total battery capacity would be wasted on auxiliary procedures rather than the transmission of actual data. Considering that these enhancements were devised to improve the spectral efficiency in the first place, this clearly defeats their purpose. Therefore, it has been established that supporting mMTC will require new, leaner solutions, with a special focus on simplified air interface, possibly at the cost of increased complexity at the BS [57].

Ideally, the device should be able to wake up from a power efficient mode, carry out the minimum necessary procedures (such as synchronization), transmit the data in a grant-free manner, wait for the feedback and perform retransmissions (optional, depending on a use case) and then return to a battery efficient mode. This suggests the usage of non-orthogonal, random access-based transmission protocols. However, due to the massive number of terminals, they may lead to a potentially large set of simultaneously active, uncoordinated users competing for the shared channel. In practice, it might be even more challenging since the devices' activity in mMTC scenarios is often triggered by an event and thus correlated, e.g. a sensor network can react to an abnormal reading and send multiple alarm messages [60]. In that sense the number of active users is not only random, but can exhibit a high variance.

## 1.1 Massive Access Enablers and State-of-the-Art

The requirements and characteristics of mMTC again point to the solutions that favor direct data transmission through random access, in particular, those based on slotted ALOHA and its extensions (cf. Section 2 of Chapter 3). Unlike in URLLC where it was motivated by strict latency constraints, here the desire to avoid multi-step resource allocation procedure is due to the associated overhead. However, because of the much larger population of devices and very low transmission rates which both entail long frames, the specific solutions are different.

An important category are those that rely on compressive sensing (CS) techniques to perform multi user activity detection (MUD) and channel estimation [61]. Compressed sensing, which by now is a mature and well researched concept, exploits the *sparsity* of the signal reflected in that only a small subset of users is active at a time (compared to the overall population). Consequently, it allows to reconstruct the signal with relatively few samples that would otherwise make the problem underdetermined. Since its inception, many algorithms to tackle CS estimation problem have been proposed [62] with variable complexity and precision. Further enhancements facilitating CS techniques involve massive MIMO, which allows to exploit sparsity of the angular domain as well [61]. Lastly, a particularly relevant for mMTC class of CS-MUD algorithms are those that incorporate non-coherent data transmission [63, 64] thus providing a full, comprehensive access solution. On the other hand, in the schemes with more traditional coherent data transmission, optimized multiple access coding and modulation techniques have been proposed, such as sparse code multiple access (SCMA) [65].

However, compressive sensing-based MUD techniques, especially those that do not incorporate data transmission, exhibit poor scalability as the number of active devices grows [66]. Furthermore, there has been an interest in examining the fundamental limits and performance bounds of massive random access from the information-theoretic point of view. This has inspired works such as [67], which coins the term many-access channel (MnAC) and [68, 69] that introduces unsourced random access (URA) channel. In a way, both attempt to address the inconsistency caused by the assumption that the number of users $N$ tends to infinity, which appears when analyzing mMTC within the classical ALOHA framework. Namely, since identification requires $\lceil \log_2 N \rceil$ bits, the length of the packet cannot have a fixed and relatively short length in such a setting. In [67] this is circumvented by making the number of devices a function of the codeword length. Meanwhile, in [68] it is assumed that the packets do not include the ID of the transmitter to keep the blocklength fixed, thus making the scheme *unsourced* and precluding user identification. This can be a problem from the point of view of reliability and security, which has been addressed in Papers G and H. On the other hand, the lack of the IDs makes it possible for all devices to share the same codebook, which allows to simplify their transmitters and reduce the communication overhead. Additionally, in some use cases the inability to identify the source can be an advantage, for example in [70]. These features has made the URA schemes of practical interest and inspired their implementations [71, 72].

While the data transmission schemes in massive access scenarios have been fairly well studied and they continue to attract the attention of many researchers, the other procedures, in particular the acknowledgement and feedback schemes, have been somewhat neglected in comparison. Meanwhile, feedback is a prerequisite for performing retransmissions and when the number of devices is massive, providing it in an efficient, yet reliable, manner becomes challenging. The common approach is to jointly encode the acknowledgements of all users using source coding, which has been explored for example in [73].

Last but not least, with regard to small payloads, there are two important practical enhancements. First, is the introduction of the already discussed flexible frame structure and mini slots [17], which provide higher granularity than LTE and are better suited for very short packets. Secondly, there has been a lot of focus recently on channel coding schemes optimized specifically towards short block-lengths [74]. Indeed, due to the emphasis on increasingly larger payloads in earlier communication systems, the research efforts were concentrated primarily on channel codes that approach the capacity for long packets, while neglecting the fact that they may suffer significant penalty when applied to shorter data.

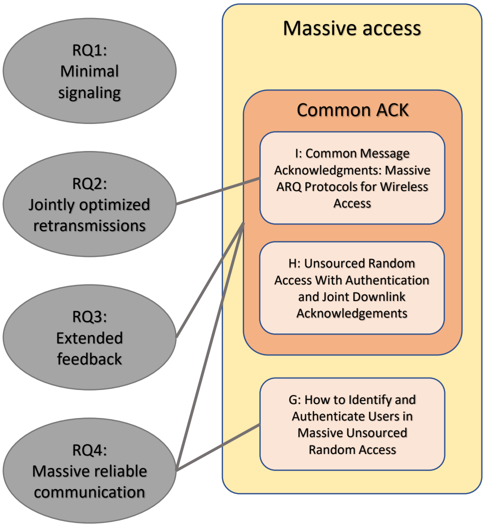

Massive access
Common ACK
I: Common Message Acknowledgments: Massive ARQ Protocols for Wireless Access
H: Unsourced Random Access With Authentication and Joint Downlink Acknowledgements
G: How to Identify and Authenticate Users in Massive Unsourced Random Access
RQ1: Minimal signaling
RQ2: Jointly optimized retransmissions
RQ3: Extended feedback
RQ4: Massive reliable communication

**Fig. 4.1:** Schematic representation of the contributions in the area of Massive Access and their relation to research questions.

## 1.2 Summary of Contributions

Within this PhD project, the following papers on the topic of reliable massive access have been written:
**Paper G:** Radosław Kotaba, Anders E. Kalør, Petar Popovski, Israel Leyva-Mayorga, Beatriz Soret, Maxime Guillaud and Luis G. Ordóñez,
"How to Identify and Authenticate Users in Massive Unsourced Random Access", IEEE Communications Letters Vol. 25, No. 12 pp. 3795–3799, 2021. (*published*)
**Paper H:** Radosław Kotaba, Anders E. Kalør, Petar Popovski, Israel Leyva-Mayorga, Beatriz Soret, Maxime Guillaud and Luis G. Ordóñez,
"Unsourced Random Access With Authentication and Joint Downlink Acknowledgements", 55$^{th}$ Asilomar Conference on Signals, Systems, and Computers, 2021. (*awaiting publication*)
**Paper I:** Anders E. Kalør, Radosław Kotaba and Petar Popovski,
"Common Message Acknowledgments: Massive ARQ Protocols for Wireless Access", IEEE Transactions on Communications. (*submitted*)

The URA offers a way to simplify the terminals, however in its basic form it also takes away the functionality of identifying the source of the transmission and authenticate its packets. To address that, in Paper G we proposed a method that reintroduces these functionalities, while preserving the main

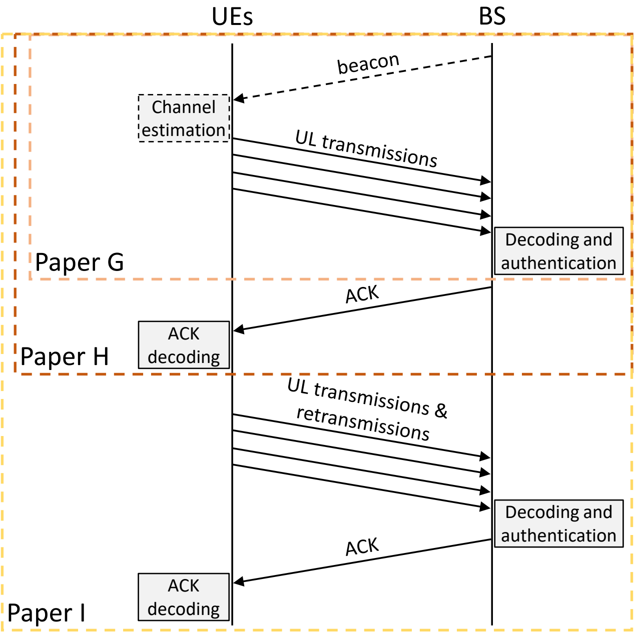


**Fig. 4.2:** The diagram of the considered protocol mapped to the individual papers.

advantage of the URA, i.e. keeping the packet short by avoiding the need to include explicit ID field. This is done at the cost of shifting the complexity towards the receiver (BS), who has to perform many additional checks to determine the ID of the sender.

In Papers G and H we generalize the definition of reliability such that for the packet to be considered successful, it must be decoded without errors and associated with the correct sender. Moreover, we distinguish between simple decoding errors, which in our scheme can be detected due to the additional authentication step, and more severe mis-identification and mis-authentication events. In the former, the packet is genuine but associated with incorrect user, while in the latter the packet contains errors and has been erroneously accepted by the BS. Interestingly, even though the lack of explicit address in the proposed scheme introduces a small probability of mis-authentication, overall our solution is able to provide higher reliability than the traditional approach which does include the ID. This is because the shortening of the packet effectively decreases the rate (in bits per channel use) thus making the transmission more robust.

In Paper H we build on the previous contribution by following the URA uplink phase with a jointly encoded downlink acknowledgment. We extend the analysis by investigating a full two-way success probability, i.e. involving the decoding and authentication at the BS, and reception of the downlink acknowledgment by UEs. Furthermore, for a fixed total number of channel uses we examine how to optimally divide them between UL and DL phases. We compare a naïve scheme, where the feedback is simply a concatenation of the IDs of the UEs being acknowledged, and the encoding approach based on Bloom filter, and show that the latter can deliver sizeable improvement of the total success probability.

The concept of jointly encoded acknowledgments is investigated in great detail in Paper I. Our work is motivated by the fact current approaches are not suitable and largely inefficient as the number of simultaneously active users becomes massive. However, by allowing a small number of false positives and encoding the acknowledgements jointly rather than individually, we are able to significantly reduce the size of the feedback message, which otherwise can take up considerable number of bits. In our work we explore a range of schemes varying in terms of complexity. The most sophisticated one among them, which involves defining and solving a set of linear equations in the Galois field, is able to match the information-theoretic lower bound. In the second part of Paper I, we devise an ARQ protocol for massive access based on the joint downlink acknowledgments. We show that despite the false positives introduced by the compression of the feedback, the overall reliability of such an ARQ protocol is significantly improved. This is because a (two-way) communication is considered to be successful only when the UE receives correct acknowledgement, which makes robust transmission of the feedback particularly crucial. That robustness is achieved by using lower coding rate enabled by the reduction of the number of bits.

**Future work:** The future work on authentication in the URA framework could exploit machine learning to speed up the process of identifying the source. Based on the traffic patterns, correlation etc. a preliminary list of the most likely sources could be made to guide and facilitate the search.

# Chapter 5

# Final Remarks

## 1 Conclusions

In the following, let us summarize the contributions of this thesis and how they connect to the research questions and the overall problem posed in Chapter 2.

In response to the first research question, our proposed solution for the uplink URLLC transmissions is to rely on a common pool of shared resources which is used in a grant-free manner, instead of many, dedicated, orthogonal allocations which do not scale very well with the number of users and can be inefficient when the traffic is intermittent. In the proposed solution the signalling is kept to minimum in order to facilitate low latency communications. However, the important thing is that even this small amount of signalling, which is used to pre-assign access patterns, is able to tremendously improve the reliability (by coordinating the interference) and, perhaps counter intuitively, simplify the system design compared to the scheme where transmissions are fully random.

The research questions 2 and 3 have been approached from two different angles – URLLC and mMTC. In the context of URLLC, a comprehensive solution has been developed in which the base station jointly optimizes the HARQ processes of many users. The application of NOMA paradigm and SIC processing in conjunction with novel transmit power and error target optimization techniques have lead to a system whose performance is far superior to that of the traditional OMA-based one. In particular, it can support almost twice as much URLLC traffic, exhibits higher availability, is more spectrally efficient and in some cases even more power efficient than the OMA counterpart. This is enabled by exploiting rich feedback to instruct the individual devices how much power they should use for their subsequent transmissions.

When it comes to mMTC, the optimization of the feedback serves a different purpose. Namely, the goal is to compress it by encoding the acknowledgements of all active users jointly (and also by introducing small probability of false positives to shorten it even further). Then, having fewer bits, the feedback can be transmitted more robustly, which has a significant, positive impact on the reliability of the overall system. The importance of the reliable acknowledgement is twofold: firstly, because transmission is considered successful only if it is decoded by the BS *and* the corresponding device receives an ACK; secondly, because it enables retransmissions. Finally, the proposed system can be improved further by dividing the fixed number of channel uses optimally between the uplink transmission and downlink feedback phases.

The fourth research question, which contemplates reliable massive access, has been partially addressed in the previous paragraph. I.e., it is our belief that one of its enablers lies in the design of schemes that rely on optimized, jointly encoded acknowledgements. Furthermore, we have recognized unsourced random access as a promising physical layer solution to enable massive uplink connectivity, noting that it also allows to simplify the transmitters and shorten the packets. This makes URA a suitable choice for mMTC. However, the fact that URA does not natively support the identification and authentication may compromise its reliability and make it prone to malicious attacks. To address this issue, a mechanism was proposed which reintroduces these functionalities, although at the cost of higher complexity at the BS.

Lastly, there is a dashed line connection between the research question 4 and the work on Steiner systems in Fig. 3.1. While the Paper D did not explicitly consider massive number of devices, it should be noted that the properties of the $(t, K, M)$ Steiner codes could make them a viable solution also in that regime if the pool of resources $M$ is sufficiently large. For example, a $(3, 5, 400)$ system contains over million access patterns. It should be noted however, that a general construction for arbitrary parameters $(t, K, M)$ is not known, which can be an obstacle. Nevertheless, there are certain infinite families for which simple construction methods exist, such as $(2, q, q^n)$ and $(3, q+1, q^n+1)$ with $q$ a prime number, and many others are tabularized [75]. Alternatively, the problem of finding suitable patterns can always be tackled by splitting the large pool of resources into smaller subpools and finding a Steiner system for each of them independently (though it might not be optimal).

## 2 Future Work and Outlook

Perhaps the most appealing (and admittedly, at times overwhelming) thing about research is that it is never really finished. Despite the solid amount of scientific output that make up this thesis and new contributions appear-ing virtually every day in various outlets, there seems to be no shortage of potential new topics. Several possible future directions have already been discussed in the respective sections on URLLC and Massive Access that can be viewed as direct and natural extensions of the work presented in this thesis. Here, I would like to focus on the more futuristic prospects in the area of reliable communication.

There is a consensus that as the 5G becomes more and more widely deployed and mature, the boundaries between use cases will start to dissolve leading to new modes of operation such as MBRLLC and mURLLC. The natural question is then what will the enablers of such ambitious modes of operation be.

In recent years, one concept has been particularly standing out in terms of popularity, and that is Reconfigurable Intelligent Surfaces (RIS). Vast amount of resources and various, often dedicated, international projects [76] have been launched to explore their capabilities. In short, RIS is a special surface that has the ability to actively modify the impinging radio waves by changing their phases, amplitudes and directions. With respect to the ultra-reliable communication they have the potential to firstly, expand the service availability by providing coverage in difficult-to-reach areas impacted by blockages as well as enabling more traditional range extension. Secondly, they can act in an adaptive and event-driven manner by "highlighting" specific UR(LL)C users/groups of users (or alternatively mute the non-URLLC ones) whenever the need to do so arises. In the context of massive access, RIS can provide additional means of separating the signals of concurrently transmitting devices.

Another hot topic revolves around semantics-oriented communication facilitated by machine learning. The idea there is that based on the context and situation, some information can be inferred without the need to explicitly communicate it, just like in real life. This could be exploited in two ways. If reliability is the main objective, then being able to infer the data would act as a second line of defense against failures by providing a chance to recover the missing information. On the other hand, for throughput maximization some information could be intentionally left out with the hopes that receiver can still recover their meaning. Furthermore, by leveraging the knowledge of the environment and “understanding” the communication objective at a given moment, predictive optimization of the network or particular links becomes possible.

Lastly, the most futuristic vision involves quantum technology. In addition to the quantum cryptography which will improve the security (an important component of overall reliability as we argue in Papers G–I), it is foreseen that quantum communication and computing will also play an important role e.g. in the form of quantum-based wireless sensor networks. By leveraging entanglement and parallelism properties, previously unattainable communication and computation speeds could be achieved.

# References

[1] "Ericsson Mobility report," Tech. Rep., Nov. 2021. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/november-2021

[2] C. Directorate-General for Communications Networks and Technology, "5G Observatory Quarterly Report 13 Up to October 2021," European Commission, Tech. Rep. [Online]. Available: https://5gobservatory.eu/wp-content/uploads/2021/11/5G-Obs-PhaseIII_Quarterly-report-13_final-version-11112021.pdf

[3] S. E. Elayoubi, M. Fallgren, P. Spapis, G. Zimmermann, D. Martín-Sacristán, C. Yang, S. Jeux, P. Agyapong, L. Campoy, Y. Qi, and S. Singh, "5G service requirements and operational use cases: Analysis and METIS II vision," in *2016 European Conference on Networks and Communications (EuCNC)*, 2016, pp. 158–162.

[4] "Deliverable D1.5 Updated scenarios, requirements and KPIs for 5G mobile and wireless system with recommendations for future investigations," Tech. Rep., Apr. 2015. [Online]. Available: https://metis2020.com/wp-content/uploads/deliverables/METIS_D1.5_v1.pdf

[5] A. Khalifeh, K. A. Aldahdouh, K. A. Darabkh, and W. Al-Sit, "A Survey of 5G Emerging Wireless Technologies Featuring LoRaWAN, Sigfox, NB-IoT and LTE-M," in *2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET)*, 2019, pp. 561–566.

[6] M. Lauridsen, I. Z. Kovacs, P. Mogensen, M. Sorensen, and S. Holst, "Coverage and Capacity Analysis of LTE-M and NB-IoT in a Rural Area," in *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall)*, 2016, pp. 1–5.

[7] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[8] P. Mogensen, W. Na, I. Z. Kovacs, F. Frederiksen, A. Pokhariyal, K. I. Pedersen, T. Kolding, K. Hugl, and M. Kuusela, "LTE Capacity Compared to the Shannon Bound," in *2007 IEEE 65th Vehicular Technology Conference - VTC2007-Spring*, 2007, pp. 1234–1238.

[9] B. Hochwald and S. ten Brink, "Achieving near-capacity on a multiple-antenna channel," *IEEE Transactions on Communications*, vol. 51, no. 3, pp. 389–399, 2003.

[10] "New Services & Applications with 5G Ultra-Reliable Low Latency Communications," 5G Americas, Tech. Rep., Nov. 2018. [Online]. Available: https://www.5gamericas.org/wp-content/uploads/2019/07/5G_Americas_URLLLC_White_Paper_Final__updateJW.pdf

[11] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1391–1396.

[12] 3GPP, "Service requirements for cyber-physical control applications in vertical domains; Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.104, 2021, v18.3.0.

[13] "Verticals URLLC Use Cases and Requirements," NGMN Alliance, Tech. Rep., July. 2019, v 2.5.4. [Online]. Available: https://www.ngmn.org/wp-content/uploads/200210-Verticals-URLLC-Requirements-v2.5.4.pdf

[14] M. Lauridsen, L. C. Gimenez, I. Rodriguez, T. B. Sorensen, and P. Mogensen, "From LTE to 5G for Connected Mobility," *IEEE Communications Magazine*, vol. 55, no. 3, pp. 156–162, 2017.

[15] S. Parkvall, E. Dahlman, A. Furuskar, and M. Frenne, "NR: The New 5G Radio Access Technology," *IEEE Communications Standards Magazine*, vol. 1, no. 4, pp. 24–30, 2017.

[16] 3GPP, "NR;Physical channels and modulation," 3rd Generation Partnership Project (3GPP), TS 38.221, 2021, v16.8.0.

[17] K. Pedersen, F. Frederiksen, G. Berardinelli, and P. Mogensen, "A Flexible Frame Structure for 5G Wide Area," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, 2015, pp. 1–5.

[18] M. Polese, M. Giordani, M. Mezzavilla, S. Rangan, and M. Zorzi, "Improved Handover Through Dual Connectivity in 5G mmWave Mobile Networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 9, pp. 2069–2084, 2017.

[19] J. Kim, G. Lee, S. Kim, T. Taleb, S. Choi, and S. Bahk, "Two-Step Random Access for 5G System: Latest Trends and Challenges," *IEEE Network*, vol. 35, no. 1, pp. 273–279, 2021.

[20] 3GPP, "NR; Medium Access Control (MAC) protocol specification," 3rd Generation Partnership Project (3GPP), TS 38.321, 2021, v16.7.0.

[21] A. Bayesteh, E. Yi, H. Nikopour, and H. Baligh, "Blind detection of SCMA for uplink grant-free multiple-access," in *2014 11th International Symposium on Wireless Communications Systems (ISWCS)*, 2014, pp. 853–857.

[22] D. Jiang, H. Wang, E. Malkamaki, and E. Tuomaala, "Principle and Performance of Semi-Persistent Scheduling for VoIP in LTE System," in *2007 International Conference on Wireless Communications, Networking and Mobile Computing*, 2007, pp. 2861–2864.

[23] A. Ksentini, P. A. Frangoudis, A. PC, and N. Nikaein, "Providing Low Latency Guarantees for Slicing-Ready 5G Systems via Two-Level MAC Scheduling," *IEEE Network*, vol. 32, no. 6, pp. 116–123, 2018.

[24] C.-F. Liu, M. Bennis, M. Debbah, and H. V. Poor, "Dynamic Task Offloading and Resource Allocation for Ultra-Reliable Low-Latency Edge Computing," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4132–4150, 2019.

[25] I. Parvez, A. Rahmati, I. Guvenc, A. I. Sarwat, and H. Dai, "A Survey on Low Latency Towards 5G: RAN, Core Network and Caching Solutions," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 3098–3130, 2018.

[26] N. A. Johansson, Y.-P. E. Wang, E. Eriksson, and M. Hessler, "Radio access for ultra-reliable and low-latency 5G communications," in *2015 IEEE International Conference on Communication Workshop (ICCW)*, 2015, pp. 1184–1189.

[27] A.-S. Bana, G. Xu, E. D. Carvalho, and P. Popovski, “Ultra Reliable Low Latency Communications in Massive Multi-Antenna Systems,” in *2018 52nd Asilomar Conference on Signals, Systems, and Computers*, 2018, pp. 188–192.

[28] M. Sawahashi, Y. Kishiyama, A. Morimoto, D. Nishikawa, and M. Tanno, “Coordinated multipoint transmission/reception techniques for LTE-advanced [Coordinated and Distributed MIMO],” *IEEE Wireless Communications*, vol. 17, no. 3, pp. 26–34, 2010.

[29] A. Anand and G. de Veciana, “Resource Allocation and HARQ Optimization for URLLC Traffic in 5G Wireless Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, 2018.

[30] R. Abreu, G. Berardinelli, T. Jacobsen, K. Pedersen, and P. Mogensen, “A Blind Retransmission Scheme for Ultra-Reliable and Low Latency Communications,” in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, 2018, pp. 1–5.

[31] S. Lin and P. Yu, “A Hybrid ARQ Scheme with Parity Retransmission for Error Control of Satellite Channels,” *IEEE Transactions on Communications*, vol. 30, no. 7, pp. 1701–1719, 1982.

[32] A. Benjebbour, Y. Saito, Y. Kishiyama, A. Li, A. Harada, and T. Nakamura, “Concept and practical considerations of non-orthogonal multiple access (NOMA) for future radio access,” in *2013 International Symposium on Intelligent Signal Processing and Communication Systems*, 2013, pp. 770–774.

[33] N. Abramson, “THE ALOHA SYSTEM: Another Alternative for Computer Communications,” in *Proceedings of the November 17-19, 1970, Fall Joint Computer Conference*, ser. AFIPS ’70 (Fall). New York, NY, USA: Association for Computing Machinery, 1970, p. 281–285. [Online]. Available: https://doi.org/10.1145/1478462.1478502

[34] L. G. Roberts, “ALOHA Packet System with and without Slots and Capture,” vol. 5, no. 2, 1975. [Online]. Available: https://doi.org/10.1145/1024916.1024920

[35] E. Casini, R. De Gaudenzi, and O. Del Rio Herrero, “Contention Resolution Diversity Slotted ALOHA (CRDSA): An Enhanced Random Access Schemefor Satellite Access Packet Networks,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1408–1419, 2007.

[36] G. Liva, “Graph-Based Analysis and Optimization of Contention Resolution Diversity Slotted ALOHA,” *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 477–487, 2011.

[37] M. Chiani, G. Liva, and E. Paolini, “The marriage between random access and codes on graphs: Coded Slotted Aloha,” in *2012 IEEE First AESS European Conference on Satellite Telecommunications (ESTEL)*, 2012, pp. 1–6.

[38] G. T. Peeters, R. Bocklandt, and B. Van Houdt, “Multiple Access Algorithms Without Feedback Using Combinatorial Designs,” *IEEE Transactions on Communications*, vol. 57, no. 9, pp. 2724–2733, 2009.

[39] C. Boyd, R. Vehkalahti, and O. Tirkkonen, “Interference Cancelling Codes for Ultra-Reliable Random Access,” *International Journal of Wireless Information Networks*, vol. 25, pp. 1–12, 12 2018.

[40] ——, “Grant-Free Access in URLLC with Combinatorial Codes and Interference Cancellation,” in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–5.

[41] E. Paolini, G. Liva, and A. Graell i Amat, “A structured irregular repetition slotted ALOHA scheme with low error floors,” in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[42] M. C. Lucas-Estañ, J. Gozálvez, and M. Sepulcre, “On the Capacity of 5G NR Grant-Free Scheduling with Shared Radio Resources to Support Ultra-Reliable and Low-Latency Communications,” *Sensors (Basel, Switzerland)*, vol. 19, 2019.

[43] R. Abreu, “Uplink Grant-free Access for Ultra-Reliable Low-Latency Communications in 5G: Radio Access and Resource Management Solutions,” Ph.D. dissertation, 2019, phD supervisor: Prof. Preben Mogensen, Aalborg University Assistant PhD supervisors: Assoc. Prof. Gilberto Berardinelli, Aalborg University Prof. Klaus Pedersen, Aalborg University.

[44] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, “Analyzing Grant-Free Access for URLLC Service,” *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 741–755, 2021.

[45] A. Anand, G. de Veciana, and S. Shakkottai, “Joint Scheduling of URLLC and eMBB Traffic in 5G Wireless Networks,” *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 477–490, 2020.

[46] R. Kassab, O. Simeone, and P. Popovski, “Coexistence of URLLC and eMBB Services in the C-RAN Uplink: An Information-Theoretic Study,” in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[47] E. Eraslan, B. Daneshrad, and C.-Y. Lou, “Performance Indicator for MIMO MMSE Receivers in the Presence of Channel Estimation Error,” *IEEE Wireless Communications Letters*, vol. 2, no. 2, pp. 211–214, 2013.

[48] C. Wang, E. K. Au, R. D. Murch, W. H. Mow, R. S. Cheng, and V. Lau, “On the Performance of the MIMO Zero-Forcing Receiver in the Presence of Channel Estimation Error,” *IEEE Transactions on Wireless Communications*, vol. 6, no. 3, pp. 805–810, 2007.

[49] S. Parkvall, E. Dahlman, P. Frenger, P. Beming, and M. Persson, “The evolution of WCDMA towards higher speed downlink packet data access,” in *IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings (Cat. No.01CH37202)*, vol. 3, 2001, pp. 2287–2291 vol.3.

[50] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, “On the Resource Utilization of Multi-Connectivity Transmission for URLLC Services in 5G New Radio,” in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, 2019, pp. 1–6.

[51] G. Berardinelli, S. R. Khosravirad, K. I. Pedersen, F. Frederiksen, and P. Mogensen, “Enabling Early HARQ Feedback in 5G Networks,” in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*, 2016, pp. 1–5.

[52] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, “Rate Allocation and Adaptation for Incremental Redundancy Truncated HARQ,” *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2580–2590, 2013.

[53] P. Popovski, "Delayed Channel State Information: Incremental redundancy with backtrack retransmission," in *2014 IEEE International Conference on Communications (ICC)*, 2014, pp. 2045–2051.

[54] B. Soret, G. Pocovi, K. I. Pedersen, and P. Mogensen, "Increasing Reliability by Means of Root Cause Aware HARQ and Interference Coordination," in *2015 IEEE 82nd Vehicular Technology Conference (VTC2015-Fall)*, 2015, pp. 1–5.

[55] S. R. Khosravirad, L. Szczecinski, and F. Labeau, "Rate Adaptation for Cooperative HARQ," *IEEE Transactions on Communications*, vol. 62, no. 5, pp. 1469–1479, 2014.

[56] Y.-L. Chung and Z. Tsai, "Cooperative Diversity with Fast HARQ for Delay-Sensitive Flows," in *2010 IEEE 71st Vehicular Technology Conference*, 2010, pp. 1–5.

[57] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.

[58] "Minimum requirements related to technical performance for IMT-2020 radio interface(s)," ITU-R, Tech. Rep., Nov. 2017, report ITU-R M.2410-0. [Online]. Available: https://www.itu.int/dms_pub/itu-r/opb/rep/R-REP-M.2410-2017-PDF-E.pdf

[59] 3GPP, "Study on Scenarios and Requirements for Next Generation Access Technologies," 3rd Generation Partnership Project (3GPP), TR 38.913, 2020, v16.0.0.

[60] K. Stern, A. E. Kalør, B. Soret, and P. Popovski, "Massive Random Access with Common Alarm Messages," in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1–5.

[61] M. Ke, Z. Gao, Y. Wu, X. Gao, and R. Schober, "Compressive Sensing-Based Adaptive Active User Detection and Channel Estimation: Massive Access Meets Massive MIMO," *IEEE Transactions on Signal Processing*, vol. 68, pp. 764–779, 2020.

[62] S. Qaisar, R. M. Bilal, W. Iqbal, M. Naureen, and S. Lee, "Compressive sensing: From theory to applications, a survey," *Journal of Communications and Networks*, vol. 15, no. 5, pp. 443–456, 2013.

[63] K. Senel and E. G. Larsson, "Grant-Free Massive MTC-Enabled Massive MIMO: A Compressive Sensing Approach," *IEEE Transactions on Communications*, vol. 66, no. 12, pp. 6164–6175, 2018.

[64] F. Monsees, M. Woltering, C. Bockelmann, and A. Dekorsy, "A potential solution for MTC: Multi-Carrier Compressed Sensing Multi-User Detection," in *2015 49th Asilomar Conference on Signals, Systems and Computers*, 2015, pp. 18–22.

[65] H. Nikopour and H. Baligh, "Sparse code multiple access," in *2013 IEEE 24th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2013, pp. 332–336.

[66] J. Liu, H.-Y. Cheng, C.-C. Liao, and A.-Y. A. Wu, "Scalable compressive sensing-based multi-user detection scheme for Internet-of-Things applications," in *2015 IEEE Workshop on Signal Processing Systems (SiPS)*, 2015, pp. 1–6.

[67] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian Many-Access Channels," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3516–3539, 2017.

[68] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527.

[69] A. Fengler, S. Haghighatshoar, P. Jung, and G. Caire, "Grant-Free Massive Random Access With a Massive MIMO Receiver," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 23–30.

[70] F. Li, B. Luo, and P. Liu, "Secure Information Aggregation for Smart Grids Using Homomorphic Encryption," in *2010 First IEEE International Conference on Smart Grid Communications*, 2010, pp. 327–332.

[71] A. Decurninge, I. Land, and M. Guillaud, "Tensor-Based Modulation for Unsourced Massive Random Access," *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 552–556, 2021.

[72] R. Calderbank and A. Thompson, "CHIRRUP: a practical algorithm for unsourced multiple access," 2019.

[73] J. Kang and W. Yu, "Minimum Feedback for Collision-Free Scheduling in Massive Random Access," *IEEE Transactions on Information Theory*, vol. 67, no. 12, pp. 8094–8108, 2021.

[74] M. Shirvanimoghaddam, M. S. Mohammadi, R. Abbas, A. Minja, C. Yue, B. Matuz, G. Han, Z. Lin, W. Liu, Y. Li, S. Johnson, and B. Vucetic, "Short Block-Length Codes for Ultra-Reliable Low Latency Communications," *IEEE Communications Magazine*, vol. 57, no. 2, pp. 130–137, 2019.

[75] "La Jolla Covering Repository, Steiner Systems," https://www.dmgordon.org/steiner/, accessed: 2022-03-04.

[76] "Reconfigurable Intelligent Sustainable Environments for 6G Wireless Networks (RISE-6G)," https://rise-6g.eu/, accessed: 2022-02-17.

# Part II

# Papers

# Paper A

## Uplink Transmissions in URLLC Systems with Shared Diversity Resources

Radosław Kotaba, Carles Navarro Manchón, Tommaso Balercia and Petar Popovski

The paper has been published in the
*IEEE Wireless Communications Letters* Vol. 7, No. 4, pp. 590–593, 2018.

# Abstract

*5G features flagship use cases with Ultra Reliable Low Latency Communication (URLLC), supported through high diversity. When multiple URLLC connections are only intermittently active, dedicating many diversity resources to a single connection leads to inefficient operation. We address this problem through shared diversity resources and compare it to per-link dedicated diversity. Two receiver types are considered, MMSE (minimum mean squared error) and MMSE-SIC (successive interference cancellation). Outage probability is evaluated by assuming channel estimation errors. The results show that it is possible to remain close to the reliability of reference system with a relatively low amount of pre-allocated resources.*

***Keywords***— URLLC, resource allocation and interference management, HARQ, transmit diversity, resource sharing.

# 1 Introduction

The advent of 5G opens up new possibilities and gives rise to a new category of use cases termed ultra reliable low latency communications (URLLC) [1]. Such services are characterized by very stringent requirements of e.g. 1 ms end-to-end latency and 99.999% reliability [2], which will be very challenging to accomplish using just the technologies and protocols of 4G and legacy systems [3].

High reliability requires use of some form of diversity. The way in which legacy systems achieve it is through hybrid automatic repeat request (HARQ), which involves exchange of feedback messages (ACK/NACK) that can trigger necessary retransmissions. However, such approach introduces latency that may not be affordable in many use cases. Another source of latency is connected to the scheduling request and grant procedure that needs to be performed before any transmission in the uplink can happen. Consequently, for extremely demanding applications some preallocation of the resources resembling that of semi-persistent scheduling [4] will be necessary in order to simultaneously cope with the reliability and latency requirements. However, such preallocation cannot be based on naïve assignment of dedicated resources to each user, as it could easily exhaust the available bandwidth and entails very poor system utilization when users are active only sporadically.

In this paper we provide an analysis of different uplink transmission schemes, taking as a baseline the traditional one used in LTE where each transmission and subsequent retransmissions are assigned dedicated resources. We compare it to a novel instance of hybrid schemes, which we coin *transmissions with shared diversity resources* (TSDR), and show that they offer significant savings of resources (which translate to lower latencies) while not compromising the performance. Inspired by the modeling of MIMO

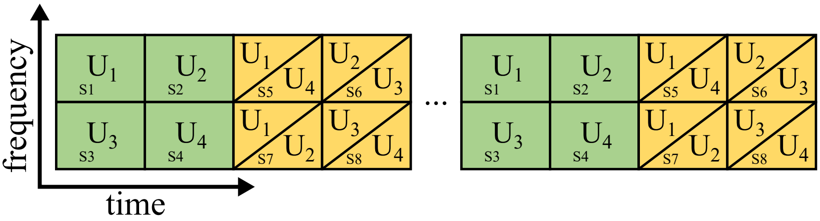


**Fig. A.1:** Example of resource allocation for $N = 4$ users over $M = 8$ slots. User $U_i$ performs $k_i = 2$ shared transmissions, $i = 1, \ldots, 4$. Green color denotes dedicated slots and yellow shared ones.

transmission [5], we propose an original, semianalytic evaluation framework which accommodates all the schemes of interest and allows us to numerically evaluate their performance in terms of outage probability. The framework allows for evaluation of the schemes assuming different conventional receivers, such as MMSE and MMSE with SIC, and takes into account impairments caused by realistic effect of non-ideal channel estimation [6] [7].

Throughout the paper the following notation is used: boldface uppercase and lowercase letters to denote matrices and vectors respectively, $\circ$ to denote Hadamard (entry-wise) product, $(\cdot)^\dagger$ to denote Moore-Penrose pseudoinverse, $(\cdot)^H$ to denote conjugate transpose, $(\cdot)_{i,j}$ to denote the $(i, j)^{th}$ entry of the matrix, $\mathbf{I}_N$ to denote identity matrix of size $N \times N$.

## 2 System Model

We analyze a system consisting of a single cell serving $N$ URLLC-type users transmitting in the uplink. At their disposal are periodic frames composed of $M$ preallocated slots each consisting of $K$ channel uses. The channel is modeled as Rayleigh fading and constant over all $K$ uses of the slot. Each user is assumed to be active in a frame with only a certain probability $p_i$. When active, user $i$ will transmit $k_i + 1$ replicas of the packet on a subset of available slots. Although we assume that each user has the same packet length equal to 1 slot, it can be easily generalized as long as the slot is kept as the smallest schedulable unit of transmission (no partial utilization). A toy example with a specific resource allocation is presented in Fig. A.1. It is further assumed that the duration of the frame is adjusted to the deadline i.e. transmission which is successful by the end of the frame is guaranteed to fulfill the latency constraint and dropped otherwise. The channel output can be written as:

$$\mathbf{Y} = \mathbf{H}\mathbf{X} + \mathbf{N} \tag{A.1}$$

where $\mathbf{Y} \in \mathbb{C}^{M \times K}$ is a received signal, $\mathbf{X} \in \mathbb{C}^{N \times K}$ with its $i^{th}$ row containing the $i^{th}$ user's complex modulated symbols and $E\left[|x_{i,j}|^2\right] = P_x$, $\mathbf{H} \in \mathbb{C}^{M\times N}$ with $H_{i,j}$ denoting the channel gain of the $j^{th}$ user in the $i^{th}$ slot, and $\mathbf{N} \in \mathbb{C}^{M\times K}$ is an additive white Gaussian noise with zero mean and variance $\sigma^2$. The channel matrix $\mathbf{H}$ can be written as:

$$\mathbf{H} = \mathbf{G} \circ (\mathbf{SP}) \tag{A.2}$$

where $\mathbf{G}$ models the underlying uncorrelated Rayleigh flat fading channel, i.e. its entries are independent and identically distributed (i.i.d) zero mean circularly symmetric complex Gaussian (ZMCSCG) variables with unit variance, $\mathbf{S} \in \{0,1\}^{M\times N}$ is a 'mask' that corresponds to the access pattern of the scheduling scheme, i.e $S_{i,j}$ is 1 when the $j^{th}$ user transmits in the $i^{th}$ slot and $\mathbf{P} = diag\left((k_1+1)^{-1/2}, \ldots, (k_N+1)^{-1/2}\right)$ is a normalization matrix ensuring that the total transmitted power per user is independent of the number of transmissions.

## 2.1 Transmission schemes

The authors of this contribution postulate the use of *transmission with shared diversity resources*, that involves splitting the $M$ resources into dedicated and shared portions. This way each user is guaranteed at least one uninterfered transmission and a configurable number of secondary transmissions in the shared part. An example of TSDR is presented in Fig. A.1 and the corresponding matrix is:

$$\mathbf{S} = \begin{pmatrix} \mathbf{I}_4 \\ \mathbf{S}_{sch} \end{pmatrix}, \quad \mathbf{S}_{sch} = \begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}. \tag{A.3}$$

To benchmark the performance of TSDR we consider three other schemes:

### Fully dedicated

As a baseline we consider a simple scheme where every transmission is assigned a distinct slot ensuring no mutual interference. This corresponds to the matrix $\mathbf{S}$ having $M = N + \sum_{i=1}^{N} k_i$ rows with a Hamming weight of 1 each. Obviously this is the most robust scheme but requires the highest number of resources for fixed $k_i$'s.

### Fully shared

On the other side of the spectrum is a fully shared scenario where each user is instructed to transmit its data on all of the available resources. This corresponds to the matrix **S** consisting of only 1′s. Such a scheme requires the least resources for fixed $k_i$′s.

### Random

In this scheme user equipments (UEs) select a new subset of slots for transmission at random in each frame which entails different realizations of matrix **S**. Such a scheme gives maximum flexibility to the users but the activity detection and data decoding is more challenging for the receiver which is forced to perform it blindly, as it doesn′t know **S** in advance.

In all of the schemes, we assume that some initial random access procedure with parameter configuration has been already performed for each user (including $k_i$, $M$, pilot assignment and, for all non-random, also **S**). Such step is necessary only once at the beginning (registering of the device) and stay valid until the resources are no longer needed by the UE and can be released.

## 3 Performance analysis methodology

For the purpose of analysis, we can look at the model and presented schemes from the point of view of MIMO system, where each User Equipment (UE) corresponds to a single transmit antenna, and each time-frequency slot is served by a different virtual receive antenna. Due to this structural similarity we are able to analyze their performance using results originally derived for MIMO.

In our evaluations we consider two types of receivers: MMSE offering a relatively good performance at a reasonable complexity, and a MMSE-SIC which is an iterative receiver achieving better results at the cost of an increased complexity.

To estimate the received signal of the form (A.1), receiver applies MMSE detection matrix **F** given by [5]:

$$\mathbf{F} = \left(\mathbf{H}^{\mathrm{H}}\mathbf{C}_n^{-1}\mathbf{H} + \frac{1}{P_x}\mathbf{I}_N\right)^{-1}\mathbf{H}^{\mathrm{H}}\mathbf{C}_n^{-1} \tag{A.4}$$

where $\mathbf{C}_n$ is the covariance matrix of the noise. The resulting estimate is the original signal contaminated by noise and interference from other users:

$$\widehat{\mathbf{X}} = \mathbf{FY} = \mathbf{FHX} + \mathbf{FN} \tag{A.5}$$

We include in our analysis the effects of imperfect channel estimation, which are expected to be relevant when resources are shared by multiple users. Following [8] we consider that $N$ out of $K$ symbols in each slot are used to transmit the training sequences which constitute rows of an $N \times N$

matrix $\mathbf{X}_{tr}$. The sequences of all $N$ users are orthogonal and have a total power $P_p$ i.e. $\mathbf{X}_{tr}\mathbf{X}_{tr}^{\mathrm{H}} = P_p\mathbf{I}_N$. The channel estimate $\widehat{\mathbf{H}}$ is obtained by applying a simple Maximum Likelihood (ML) estimator to the received training signal:

$$\widehat{\mathbf{H}} = \mathbf{Y}_{tr}\mathbf{X}_{tr}^{\dagger} = (\mathbf{H}\mathbf{X}_{tr} + \mathbf{N})\mathbf{X}_{tr}^{\dagger} = \mathbf{H} + \underbrace{\frac{1}{P_p}\mathbf{N}\mathbf{X}_{tr}^{\mathrm{H}}}_{\mathbf{\Delta H}} \tag{A.6}$$

where each entry of the error matrix $\mathbf{\Delta H}$ is i.i.d complex normal variable with variance $\sigma_H^2 = \frac{\sigma^2}{P_p}$. Consequently, the noisy channel estimate $\widehat{\mathbf{H}}$ introduces the distortion $\mathbf{\Delta F}$ to the detection matrix such that the estimate of $\mathbf{X}$ becomes:

$$\widehat{\mathbf{X}} \cong (\mathbf{F} + \mathbf{\Delta F})(\mathbf{H}\mathbf{X} + \mathbf{N}) = \mathbf{F}\mathbf{H}\mathbf{X} + \widehat{\mathbf{N}} \tag{A.7}$$

The post-processing SINR (PPSINR) of each stream that can be derived from (A.7) takes the form:

$$\mathrm{SINR}(i) = \frac{P_x K \left|(\mathbf{F}\mathbf{H})_{i,i}\right|^2}{P_x K \sum\limits_{j \neq i} \left|(\mathbf{F}\mathbf{H})_{i,j}\right|^2 + (E[\widehat{\mathbf{N}}\widehat{\mathbf{N}}^{\mathrm{H}}])_{i,i}} \tag{A.8}$$

The PPSINR for MMSE-SIC receiver is obtained using the same formula (A.8) but the procedure is iterative with optimal ordering [9], i.e. at the end of each iteration stream with the highest PPSINR $i_{max}$ is removed from $\mathbf{Y}$ by subtracting $\widehat{\mathbf{h}}_{i_{max}} x_{i_{max}}$. The decoding process is then repeated with fewer interfering streams (corresponding column of $\widehat{\mathbf{H}}$ removed) and slightly increased noise due to the residual term $\mathbf{\Delta h}_{i_{max}} x_{i_{max}}$.

For low values of $p_i$ the chance that at most one UE is active is relatively high leading to a simple SIMO system. Following [7] we can approximate this case by:

$$\mathrm{SINR}_{\mathrm{SIMO}}(i) \sim \frac{P_x}{\sigma^2 + \sigma_H^2 P_x}\chi^2_{2(k_i+1)} \tag{A.9}$$

where $\chi_l^2$ is a chi-squared distributed random variable with $l$ degrees of freedom.

Using the capacity formula for AWGN channel with i.i.d. ZMCSCG input signal process, the achievable rate is upper bounded by $\mathcal{R}_{max} = \log_2(1 + \mathrm{SINR}(i))$. Since for URLLC we are very often interested in outage measures of the system rather than pure throughput, the performance metric we will be using in the following section is the outage probability:

$$p_{out}(i) = Pr\left\{R > \mathcal{R}_{max}\left[\mathrm{SINR}(i)\right]\right\} \tag{A.10}$$

i.e. the probability that the rate $R$ (in bits/s/Hz) at which UE transmitted its data was higher than the instantaneous maximum achievable rate.

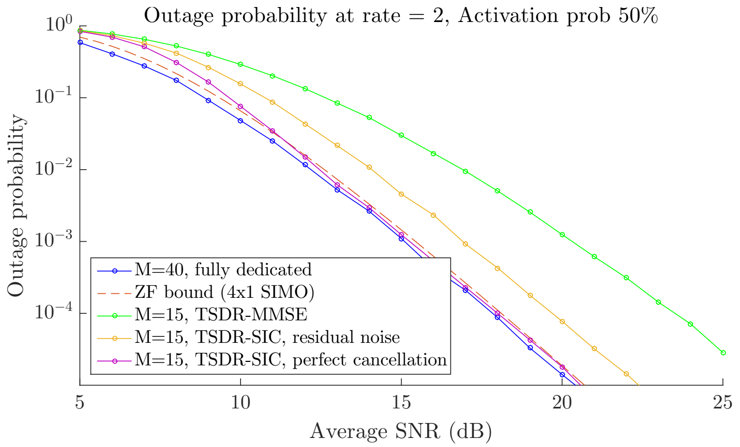


**Fig. A.2:** Performance of fully dedicated scheme and TSDR scheme with different receiver complexity for $N = 10$ and $k_i = 3$

Finally, we remark that explicit analysis of the latency is not the goal of this paper. Instead, we focus on analyzing how many slots M are necessary and how to best utilize them with respect to certain reliability targets. Taking into account other factors such as receiver processing delay, slot duration (determined by the subcarrier spacing and number of constituting OFDM symbols) allows to arrange the slots on a time-frequency grid so that a particular latency target is met.

# 4 Results

In this section, we present and discuss the results obtained through extensive simulations based on the analysis outlined in previous sections. The channel realizations **H** are generated as ZMCSCG according to (A.2) and with appropriate masks dependent on the scheme. The symbol power for each user is fixed to $P_x = 1$ while $\sigma^2$ is varied accordingly to SNR. For the purpose of calculating $\sigma_H^2$ the pilot power is set to $P_p = 4P_x$ so that the quality of channel estimation also depends on SNR. In the outage probability investigations, we select a relatively low transmission rate of 2 bits/s/Hz, which captures the robustness and low payload sizes of considered URLLC use cases.

In Fig. A.2 we show the gains of using advanced SIC receivers in combination with schemes based on shared diversity resources. As a baseline we consider the performance of fully dedicated scheme and compare it with TSDR operating over reduced number of slots $M$ and the same total number of transmissions per user $k_i$. We can see that with no SIC, which corresponds to the plain MMSE receiver, the performance is visibly degraded. However,

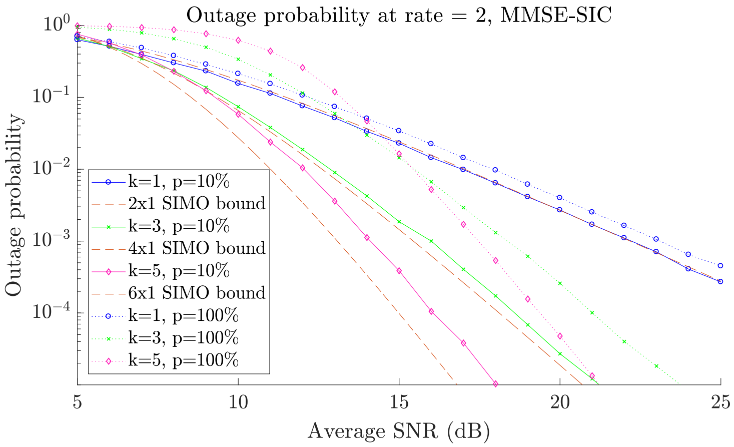


**Fig. A.3:** Performance of TSDR schemes with variable number of secondary transmissions and fixed $M = 15$ and $N = 10$.

using a more advanced receiver allows to approach the performance of dedicated scheme with almost three times less resources at a cost of moderate increase in complexity. To highlight the significance of imperfect channel estimation we provide the curves for both ideal SIC and the one introducing residual interference. In the rest of our evaluations we consider only the non-ideal one as it is more interesting to analyze and more realistic[1], while still significantly outperforming the MMSE receiver.

In Fig. A.3 we analyze the interplay between the channel estimation errors, number of shared transmissions $k_i$ and user activation probability. As shown by our analysis, channel estimation errors limit the interference cancellation capabilities of the receiver. In fact, one of the most important findings of this contribution is that, due to those imperfections, increasing $k_i$ offers diminishing returns in terms of diversity and causes larger dependency on activation probability. Consequently, TSDR with higher degree of resource sharing (higher $k_i$) will observe more severe performance drop with increased $p_i$, which might be of importance if the traffic is bursty rather than uniform. For the outage probabilities of interest this degradation can be quite significant (e.g., 3dB of SNR for $k_i = 5$ and 2dB for $k_i = 3$ at $10^{-5}$ outage probability).

Fig. A.4 compares TSDR and the idealized random scheme described in section 2.1 in terms of their dependency on users' activation probability. We can see that for higher values of $k_i$ randomization has an advantage since it allows to avoid too congested slots. However, we note that practical realiza-

[1] In practice, the gap could be reduced in several ways. Simplest method involves increasing the number of shared slots while keeping $k_i$ fixed to reduce the amount of interference. Another solution is to dedicate more resources to the pilots. Lastly, one could invest more computational power and use the successfully decoded stream as new pilots to refine the channel estimate.

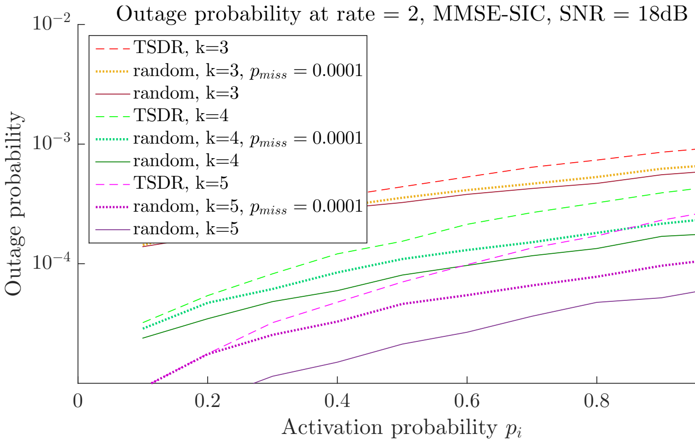


**Fig. A.4:** Impact of the users' activation probability on the performance of TSDR and random schemes with fixed SNR = 18dB and $N = 10$.

tions of such random schemes will require the base station to perform blind activity detection and decoding which inevitably will lead to false positives and false negatives. To give some insight, we consider also a simplified model where each packet replica has a probability of miss-detection $p_{miss}$ in which case the corresponding entry $\widehat{\mathrm{H}}_{i,j}$ is erroneously set to 0 and consequently $\Delta\mathrm{H}_{i,j} = -\mathrm{H}_{i,j}$. As shown in Fig. A.4 the impact on performance is significant even for low values of $p_{miss}$. Another issue connected with random access arises when the number of available pilots is limited which causes sporadic collisions and pilot contamination between users. TSDR and other coordinated schemes offer a way to avoid that.

Lastly, in Fig. A.5 we present our findings regarding the maximum number of supported users $N$ fulfilling the outage probability target of $10^{-5}$ at 20dB SNR as a function of available resources $M$. To meet the requirements with fully dedicated scheme each user must transmit in total $k_i + 1 = 5$ replicas of the packet, which entails very poor scaling of the system where $N = \lfloor M/5 \rfloor$. When using TSDR the behavior of maximum $N$ is much more linear as for every four slots invested it allows to add approximately three new users (over the simulated range the exact relationship is $N = 1 + \lceil 3(M-5)/4 \rceil$). For the fully shared scheme, the number of users $N$ is linear with $M$ thus achieving an upper bound (we do not consider here the underdetermined systems where $N > M$). However this scheme requires that $k_i + 1 = M$ which quickly becomes computationally prohibitive. On the same figure we also provide the achievable average capacity per user as dictated by their PPSINR. We can see that TSDR significantly outperforms the fully shared scheme in that metric. The results can be interpreted as follows: more replicas lead to lower mean and variance of the PPSINR (making the curves in

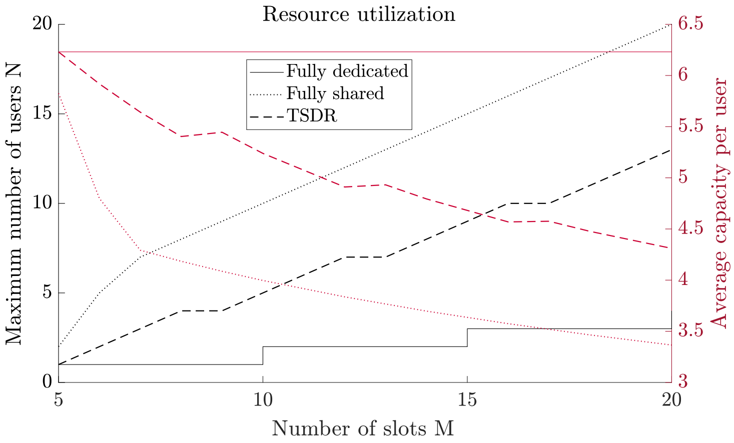


**Fig. A.5:** Maximum number of users $N$ that achieve the outage probability target of $10^{-5}$ at 20dB SNR and their average capacity.

Fig. A.3 steeper and shifted to the right). This could be dangerous if the SNR cannot be reliably estimated due to, for example, large fluctuations of the inter-cell interference.

## 5 Conclusion

In this publication we propose a novel uplink transmission scheme, TSDR, in which resources are shared by users in a coordinated manner. The scheme relies on the usage of advanced (SIC) receiver processing in order to achieve the URLLC requirements. We show that TSDR offers very large saving of resources compared to schemes in which users have dedicated resources for transmission. At the same time, it strikes a balance between excessive complexity imposed by random schemes and computational burden of fully shared scheme. Furthermore, our analysis reveals the importance of accounting for channel estimation errors in the design of the air interface, especially when advanced receivers are considered.

## References

[1] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity*, 2014, pp. 146–151.

[2] 3GPP, "Service requirements for the 5G system; Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.261, 2017, v2.0.0.

[3] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1391–1396.

[4] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. John Wiley & Sons Ltd, 2011.

[5] T. Brown, E. De Carvalho, and P. Kyritsi, *Practical Guide to the MIMO Radio Channel*. John Wiley & Sons Ltd, 2012.

[6] E. Eraslan, B. Daneshrad, and C.-Y. Lou, "Performance Indicator for MIMO MMSE Receivers in the Presence of Channel Estimation Error," *IEEE Wireless Communications Letters*, vol. 2, no. 2, pp. 211–214, 2013.

[7] C. Wang, E. K. Au, R. D. Murch, W. H. Mow, R. S. Cheng, and V. Lau, "On the Performance of the MIMO Zero-Forcing Receiver in the Presence of Channel Estimation Error," *IEEE Transactions on Wireless Communications*, vol. 6, no. 3, pp. 805–810, 2007.

[8] T. L. Marzetta, "BLAST Training: Estimating Channel Characteristics for High Capacity Space-Time Wireless," in *Proc. 37th Annual Allerton Conference on Communications, Control, and Computing*, 1999, pp. 958–966.

[9] P. Wolniansky, G. Foschini, G. Golden, and R. Valenzuela, "V-BLAST: an architecture for realizing very high data rates over the rich-scattering wireless channel," in *1998 URSI International Symposium on Signals, Systems, and Electronics. Conference Proceedings (Cat. No.98EX167)*, 1998, pp. 295–300.

# Paper B

## Enhancing Performance of Uplink URLLC Systems via Shared Diversity Transmissions and Multiple Antenna Processing

Radosław Kotaba, Carles Navarro Manchón and Petar Popovski

The paper has been published in the
*53rd Asilomar Conference on Signals, Systems, and Computers*, pp. 1409–1415,
2019.

# Abstract

*In this work we investigate the reliability aspects of uplink multi-user MIMO communication over a preallocated pool of time-frequency resources shared by a group of ultra-reliable low-latency devices. To achieve sufficient diversity, users perform multiple transmissions of their packets over a shared pool of time-frequency resources in a non-orthogonal manner. The preallocation allows users to employ fast, grant-free type of access, while sharing improves the overall spectral efficiency. The multiple transmit opportunities enhance the robustness of communication through incremental redundancy. On the base station side, we consider the performance of a minimum mean square error (MMSE) receiver, chosen for its relative simplicity. In addition to a baseline scheme in which devices randomly select the resources without coordination, we consider two other approaches based on preassigned access patterns: i) one in which all resources are utilized evenly and with equal power, and ii) another, where the spectrum is divided into high and low contention portions and users benefit from having few reliable transmissions and few diversity resources. In particular, we focus on evaluating the performance limits of the schemes as the number of antennas at the base station grows.*

# 1 Introduction

Ultra-reliable low latency communication (URLLC) is a new category of use cases in the latest, fifth generation cellular standard [1], and it encompasses the most demanding types of applications including (but not limited to): Industry 4.0 scenarios (factory automation, motion control) [2], tele-surgery (based on haptic feedback) and vehicular-to-anything (V2X) in the Automotive industry [3].

Achieving spectrally efficient URLLC is inherently difficult, which is the key takeaway from [4]. In fact, the solutions implemented in practical systems make simultaneous low latency and high reliability contradictory as they trade one for another. This is especially true for uplink (UL) traffic which in the classical cellular networks is fully managed by the centralized base station (BS). A comprehensive overview of the challenges faced by URLLC can be found in [5] where the authors discuss in detail various enablers and their tradeoffs.

Among the most promising and at the same time disruptive techniques is the grant-free access, which gives devices the ability to perform transmissions without prior scheduling [6]. Indeed, the requirement to perform scheduling grant handshake is one of the largest bottlenecks in the design of low latency systems. The price to pay for avoiding it is a significantly reduced control that the BS has over interference.

Since its inception the topic of grant-free access has garnered a lot of attention. Several designs and implementations were considered, earliest of which build on the legacy concept known as slotted ALOHA [7] and its extension coded random access (CRA) [8]. The actual grant-free scheme as defined for 5G NR, utilizing $k$-repetitions over shared resources and aperiodic traffic has been studied in [9] [10]. The former contribution focuses on the collision aspect (from the combinatorial point of view), while the latter provides a realistic assessment through system-level simulations in an outdoor urban micro scenario. In [11] a hybrid approach is studied where devices initiate the communication grant-free and switch to coordinated access for retransmissions.

The idea of introducing some coordination in the form of preassigned access patterns is discussed in several works [12] [13]. The former coins the term transmissions with shared diversity resources (TSDR) and focuses on their performance in the presence of imperfect CSI. The latter analyzes a special type of access patterns based on the code construction according to Steiner system.

In this work we analyze a multi-user URLLC system where the diversity required to achieve high reliability is provided by a combination of multiple receive antennas and multiple redundant transmissions of the packet over a shared pool of resources. We start by pointing out the shortcomings of the naive grant-free access scheme, which does not take into account potential pilot collisions, and postulate that preassigned access patterns should be used instead to avoid them. Aided by the recent work of [14], we develop an original analytical framework that allows to evaluate the outage performance of such multi-user, multi-transmission system when the BS utilizes MMSE processing. To the best of the authors' knowledge such results haven't been obtained before and until recently could only be treated under ordinary zero-forcing (ZF). We apply the developed tools to analyze two types of access patterns: uniform patterns, evenly utilizing the whole resource pool; and a generalized version of TSDR which combines slots with higher and lower amount of interference. The obtained results clearly show the superiority of preassigned patterns over the naïve grant-free in terms of outage probability. Moreover, the approach based on TSDR has a potential for reducing the complexity and effective latency compared to the uniform ones.

Throughout the paper the following notation is used: $(\cdot)^{\mathrm{H}}$ to denote conjugate transpose, $(\cdot)_{i,j}$ to denote the element in $i$-th row and $j$-th column of the matrix, bold uppercase letters to denote matrices respectively. $\mathbf{I}_N$ denotes an $N \times N$ identity matrix. $\|\cdot\|_0$ denotes the $\ell_0$ pseudonorm. $\mathrm{E}[\cdot]$ denotes the expected value.

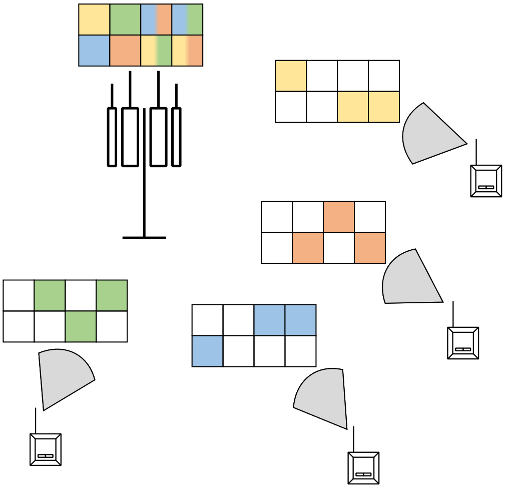


**Fig. B.1:** Example of the grant-free access scheme with $k = 3$ multiple transmissions over a pool of $L = 8$ TF-blocks.

## 2 System Model

We consider a single base station (BS) serving $N$ URLLC-type users. The base station is equipped with $M$ antennas while each of the devices (UEs) has a single antenna. The access channel is divided into periodic frames composed of $L$ slots also referred to as time-frequency (TF) blocks. Each such block is further composed of $K$ channel uses. In this work we consider the case of Rayleigh block fading channel, where the realizations of the channel coefficients are independent between different slots and UEs. We assume that all UEs transmit with the same rate $R$ and a worst case scenario is considered where all of them are active in each frame. To harvest diversity and consequently achieve reliability each user transmits its packet using $k$ out of $L$ slots in a frame. The packets can be identical, constituting a form of $k$-repetition coding, or contain different coded symbols (redundancy versions) of the original message. Throughout the paper we will refer to the former and latter scheme as Chase Combining (CC) and Incremental Redundancy (IR) respectively.

At the receiver, the baseband representation of the channel output during $l$-th slot of the $d$-th frame can be written as:

$$\mathbf{Y}_{d,l} = \mathbf{H}_{d,l}\mathbf{X}_{d,l} + \mathbf{N}_{d,l} = \left(\mathbf{R}_{RX}\mathbf{G}_{d,l}\mathbf{P}_{d,l}\right)\mathbf{X}_{d,l} + \mathbf{N}_{d,l} \tag{B.1}$$

where $\mathbf{Y}_{d,l} \in \mathbb{C}^{M\times K}$ are received symbols, $\mathbf{X}_{d,l} \in \mathbb{C}^{N\times K}$ are the transmitted complex modulated symbols normalized such that $\mathrm{E}\left[\left|x_{i,j}\right|^2\right] = 1$, $\mathbf{N}_{d,l} \in$

$\mathbb{C}^{M\times K}$ is the additive white Gaussian noise (AWGN) with zero mean and variance $\sigma^2$ and $\mathbf{H}_{d,l} \in \mathbb{C}^{M\times N}$ are the channel gains between $N$ users and $M$ antennas. The component $\mathbf{H}_{d,l}$ can be further represented as a product of $\mathbf{G}_{d,l} \in \mathbb{C}^{M\times N}$, which is zero-mean circularly-symmetric complex Gaussian (ZMCSCG) and models the underlying uncorrelated Rayleigh flat fading channel, $\mathbf{P}_{d,l} = diag\left((P_{d,l,1})^{1/2}, \ldots, (P_{d,l,N})^{1/2}\right)$ is a diagonal matrix of transmit powers and $\mathbf{R}_{RX}$ is a square Toeplitz matrix with parameter $\rho$ denoting receive antenna correlation. The packet of each UE is subject to the total transmit power constraint such that independently of the total number of transmissions within a frame

$$\sum_{l=1}^{L} P_{d,l,i} = P_{tot}, \qquad \forall i \in \{1, \ldots, N\}, \forall d. \tag{B.2}$$

Note that some of the $P_{d,l,i} = 0$ which reflects the fact that each UE uses only $k$ among $L$ available slots, specifically $\sum_{l=1}^{L} \left\|P_{d,l,i}\right\|_0 = k$.

In the remainder of the paper we will omit the frame index $d$ as the transmissions within a single frame are self contained and independent (i.e. UEs are not allowed to transmit the same packet over multiple frames as this would violate the latency constraint).

## 2.1 Receiver processing

As the use case on which we are focusing in this paper is URLLC, the relevant metric for our system is outage probability. When IR transmission mode is being used the outage can be defined as[1]

$$p_{out_i} = \Pr\left\{R > \sum_{l=1}^{L} \ln\left(1 + \mathrm{SINR}_{l,i}\right)\right\} \tag{B.3}$$

where $R$ is the transmission rate (in nats) used to encode the packet. The $\mathrm{SINR}_{l,i}$ is understood as the post-processing signal-to-interference-plus-noise ratio of user $i$ in slot $l$ and can be computed as

$$\mathrm{SINR}_{l,i} = \frac{\left|(\mathbf{F}_l\mathbf{H}_l)_{i,i}\right|^2}{\sum_{j=1,j\neq i}^{N} \left|(\mathbf{F}_l\mathbf{H}_l)_{i,j}\right|^2 + \sigma^2 \sum_{j=1}^{N} \left|(\mathbf{F}_l)_{i,j}\right|^2} \tag{B.4}$$

where $\mathbf{F}_l$ is the detection matrix employed by the BS. In this work we chose to focus on the minimum mean square error equalization method where

[1] In the case of CC experiencing independent interference, the summation appears inside the logarithm instead. Transmissions can also be processed jointly yielding single, combined post-processing SINR.

$\mathbf{F}_l = \left(\mathbf{H}_l^{\mathrm{H}}\mathbf{H}_l + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{H}^{\mathrm{H}}$, in which case (B.4) simplifies to

$$\mathrm{SINR}_{l,i} = \frac{1}{\left(\left(\mathbf{H}_l^{\mathrm{H}}\mathbf{H}_l + \sigma^2\mathbf{I}_N\right)^{-1}\right)_{i,i}} - 1. \tag{B.5}$$

It is further assumed that prior to the transmission BS and UEs possess only a statistical knowledge of the channel state information (CSI). Once the packets are received, and unless otherwise stated, BS is capable of perfectly estimating **H** from the available pilots.

# 3 Grant-Free Access

Grant-free transmissions are among the most frequently proposed enablers of the use cases requiring extremely low latencies. In its most basic version, grant-free involves dedicating a set of $L$ TF-blocks to create a common pool of resources which are then accessed by a group of UEs in random and uncoordinated manner. Whenever a device has a packet to send, it selects randomly $k$ out of $L$ slots and uses them to transmit its data. A toy example of such scheme is shown in Fig. B.1. In such a grant-free system with $N$ active devices, the contention level of the $l$th TF-block, denoted by $C_l$ and defined as the number of UEs using the $l$th TF-block, is a binomial random variable with success probability $p = \frac{k}{L}$ and $N$ trials s.t. $\mathrm{E}\left[C_l\right] = \frac{kN}{L}$.

In principle, having such a scheme is possible as long as the BS is equipped with enough antennas and has sufficiently accurate CSI, as it can resolve the (potentially numerous) collisions with proper multi-antenna processing. In a fully uncoordinated grant-free scheme, however, the UEs need to select at random not only TF-blocks but also one out of a finite number of pilot sequences. Collisions of the latter can be more severe as they lead to pilot contamination and hinder the use of multi-antenna detectors.

To illustrate these issues we will perform the following experiment. We simulate uncoordinated, grant-free random access by considering two situations: an idealized one in which the BS always has perfect CSI, regardless of the contention levels of each slot, and a realistic one in which users select one out of $D$ available pilot sequences in an uncoordinated manner. In the latter scenario, when pilot sequences collide in a slot, the information in the associated transmission is considered lost for the BS receiver. For additional details regarding the signal processing and issues related to the simulation of this scenario we refer the reader to the Appendix A.

The results of this experiment are shown in Fig. B.2. The simulations in this and other figures (unless otherwise stated) are done with $L = 12$ TF-blocks, $N = 24$ users, $M = 8$ receive antennas, $\mathbf{R}_{RX} = \mathbf{I}_M$ and for different number of replicas $k \in \{2, 3, 4\}$. The target rate is set to $R = \ln 2$ (which

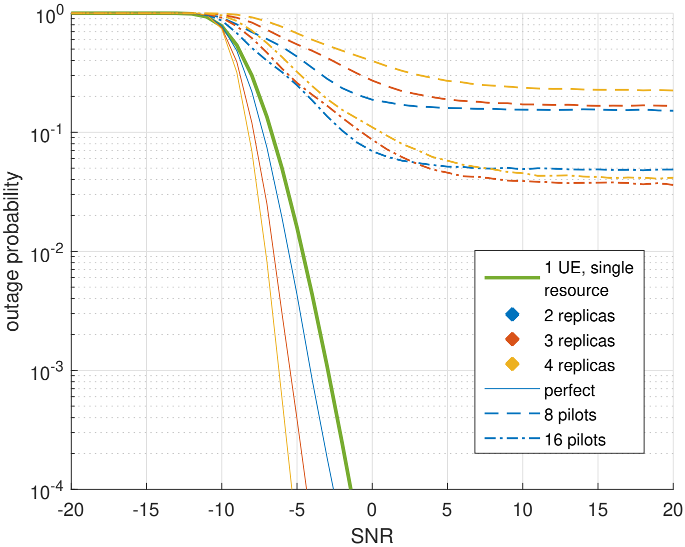


**Fig. B.2:** Outage probability performance of idealized and pilot-limited grant-free access. The average contention levels corresponding to 2,3, and 4 replicas are 4, 6, and 8 respectively. The one-shot transmission used for comparison is equivalent to $k = 1$ and $C_l = 1$

corresponds to 1 bit/channel use). The transmit power is $\frac{P_{tot}}{k}$ and is determined by the operating point (x-axis). From Fig. B.2 we can see that when the number of pilots is limited, the outage probability is severely degraded and exhibits plateauing with the level related to the probability that all replicas are lost due to collisions. Depending on the number $D$, transmitting more replicas $k$ may or may not help as it involves a trade-off between their number and increasing the chance of collision per transmission. We should also note that in practical systems increasing the number of available pilot sequences results either in a loss of spectral efficiency (larger percentage of resources dedicated to pilots) or in an increased transmission rate (and therefore reliability degradation) if a constant spectral efficiency is to be maintained.

As a reference, with thick green line we show also the performance of a single user transmitting over a dedicated, interference-free slot. Although idealized grant-free experiences much higher average contention (which translates to lower diversity order per replica) and can even lead to situations where $C_l > M$ it clearly outperforms the so called one-shot approach. It is even more surprising considering that the latter offers only half the rate of grant-free ($\frac{24R}{12}$ vs $\frac{R}{1}$).

# 4 Preassigned Access Patterns

As confirmed by the experiment in the previous section, the idea of pooling resources has clearly a great potential but is reliant on the availability of CSI. To address the main flaw of grant-free access, we consider in this section the case where users have preassigned access patterns (known and assigned by the BS) rather than selecting them randomly themselves. These patterns can be considered fixed or at least changing on a much larger timescale than the duration of the frame[2].

The specific design of patterns determines the maximum contention level $C_l$ of each TF-block. Most importantly, unlike random selection, $C_l$ is deterministic and controlled by the BS, and can therefore be adapted to the number of available pilot sequences. In particular, since the number of simultaneously transmitting devices in a TF-block can be made lower or equal to the number of pilots, the pilot sequences can be preassigned to UEs, in a way that ensures they won't collide with each other.

The determinism of $C_l$ significantly simplifies the problem and allows to derive some analytical tools and results which are the focus of the following subsection. Then, in subsections 4.2 and 4.3 we consider two special cases of access patterns and apply the aforementioned tools to assess their performance.

## 4.1 Outage probability analysis

Recently, the authors of [14] were able to obtain a closed-form expression for the pdf of the SINR provided by the MMSE equalizer in an uncorrelated Rayleigh fading setting with $C_l \leq M$ and equal power allocation between users. This surprisingly simple result yields

$$f_{\mathrm{SINR}_{l,i}}(x) = \frac{x^{M-1}e^{-\frac{x}{\gamma_l}}}{(1+x)^{C_l}} \sum_{a=0}^{C_l-1} \binom{C_l-1}{a} \frac{\gamma_l^{a-M}}{(M-a-1)!} \left(\frac{M}{M-a}+x\right) \tag{B.6}$$

where $\gamma_l = P_l/\sigma^2$ is the (same) average SNR of the UEs active in a slot $l$.[3]

Since in this work we define the outage criterion in terms of mutual information rather than SINR directly (cf. (B.3)) we need to make a simple transformation. Let $\mathrm{MI}_{l,i} = g(\mathrm{SINR}_{l,i}) = \ln(1+\mathrm{SINR}_{l,i})$ be the mutual information (in nats) of the $i$-th user message obtained from the $l$-th TF-block.

[2] In practice, they could be assigned when the device first registers with the BS and then updated periodically to adapt to the varying total population of the URLLC users.

[3] Strictly speaking, in our scenario the parameter $\gamma_l$ can have two values: either $P_l/\sigma^2$ or 0. In the latter case, the pdf should be replaced by a Dirac delta distribution (and consequently Heavyside step function for cdf).

Consequently $\mathrm{SINR}_{l,i} = g^{-1}(\mathrm{MI}_{l,i}) = e^{\mathrm{MI}_{l,i}} - 1$. Using then the pdf transformation $f_Y(y) = f_X\left(g^{-1}(y)\right)\frac{d\left(g^{-1}(y)\right)}{dy}$ we obtain a new pdf

$$f_{\mathrm{MI}_{l,i}}(x) = \frac{(e^x-1)^{M-1}e^{-\frac{e^x-1}{\gamma_l}}}{e^{x(C_l-1)}} \sum_{a=0}^{C_l-1} \binom{C_l-1}{a} \frac{\gamma_l^{a-M}}{(M-a-1)!} \left(\frac{a}{M-a} + e^x\right). \tag{B.7}$$

Work [14] provides also the cdf of the SINR albeit the expression is slightly more complex. After adapting it to our scenario the expression reads

$$\begin{aligned} F_{\mathrm{MI}_{l,i}}(x) = {} & \mathcal{I}(M, M+1-C_l, \frac{1}{\gamma_l}, e^x - 1) \sum_{a=0}^{C_l-1} \binom{C_l-1}{a} \frac{M\gamma_l^{a-M}}{(M-a)!} \\ & + \mathcal{I}(M+1, M+2-C_l, \frac{1}{\gamma_l}, e^x - 1) \sum_{a=0}^{C_l-1} \binom{C_l-1}{a} \frac{\gamma_l^{a-M}}{(M-a-1)!} \end{aligned} \tag{B.8}$$

where $\mathcal{I}(a,b,c,x) = \int_0^x e^{-ct}t^{a-1}(t+1)^{b-a-1}dt$. Since user $i$'s total mutual information $\mathrm{MI}_i^{total}$ is a sum of contributions from the $k$ packets transmitted by the user, we eventually rewrite the outage probability (B.3) as

$$\begin{aligned} p_{out_i} &= \Pr\left\{R > \mathrm{MI}_i^{total}\right\} = \Pr\left\{R > \sum_{l \in \mathcal{L}_i} \mathrm{MI}_{l,i}\right\} \\ &= \left(F_{\mathrm{MI}_{l_1^i,i}} * f_{\mathrm{MI}_{l_2^i,i}} * \cdots * f_{\mathrm{MI}_{l_k^i,i}}\right)(x)\Big|_{x=R} \end{aligned} \tag{B.9}$$

where $\mathcal{L}_i = \{l_1^i, l_2^i, \ldots, l_k^i\}$ is the set of indices of TF-blocks where user $i$ transmits.

In addition to the exact outage probability given by (B.9), which might be cumbersome to evaluate for larger number of replicas $k$ as it requires multiple numerical convolutions/integrations, we provide here also it's Chernoff bound. Let us start by deriving the moment generating function of the mutual information

$$
\begin{aligned}
M_{\mathrm{MI}_{l,i}}(t) &= \mathrm{E}\left[e^{t\ln\left(1+\mathrm{SINR}_{l,i}\right)}\right] = \int_0^\infty (1+x)^t f_{\mathrm{SINR}_{l,i}}(x)dx \\
&= \int_0^\infty \frac{x^{M-1}e^{-\frac{x}{\gamma_l}}}{(1+x)^{C_l-t}}dx \sum_{a=0}^{C_l-1}\binom{C_l-1}{a}\frac{M\gamma_l^{a-M}}{(M-a)!} \\
&+ \int_0^\infty \frac{x^{M}e^{-\frac{x}{\gamma_l}}}{(1+x)^{C_l-t}}dx \sum_{a=0}^{C_l-1}\binom{C_l-1}{a}\frac{\gamma_l^{a-M}}{(M-a-1)!} \\
&= U(M, M+1-C_l+t, \frac{1}{\gamma_l}) \sum_{a=0}^{C_l-1}\binom{C_l-1}{a}\frac{M!\gamma_l^{a-M}}{(M-a)!} \\
&+ U(M+1, M+2-C_l+t, \frac{1}{\gamma_l}) \sum_{a=0}^{C_l-1}\binom{C_l-1}{a}\frac{M!\gamma_l^{a-M}}{(M-a-1)!}
\end{aligned}
\tag{B.10}
$$

where $U(a, b, c)$ is the confluent hypergeometric function of the second kind. The outage probability can be then upper-bounded as

$$
\begin{aligned}
p_{out_i} &= \Pr\left\{R > \mathrm{MI}_i^{total}\right\} \\
&= \Pr\left\{e^{-t\sum_{l\in\mathcal{L}_i}\mathrm{MI}_{l,i}} > e^{-tR}\right\} \qquad , \; t \in \mathbb{R}^+ \\
&\le \min_{t>0} \frac{\mathrm{E}\left[e^{-t\sum_{l\in\mathcal{L}_i}\mathrm{MI}_{l,i}}\right]}{e^{-tR}} = \min_{t>0} \frac{\prod_{l\in\mathcal{L}_i}\mathrm{E}\left[e^{-t\mathrm{MI}_{l,i}}\right]}{e^{-tR}} \\
&= \min_{t>0} e^{tR} \prod_{l\in\mathcal{L}_i} M_{\mathrm{MI}_{l,i}}(-t).
\end{aligned}
\tag{B.11}
$$

## 4.2 Uniform Patterns

We can now apply the tools developed in the preceding section to some specific cases of grant-free access with preallocated patterns. We will start with the most straightforward approach in which the patterns use the $L$ available slots evenly (in other words, their covering is uniform). Consequently, $C_l = \frac{kN}{L}$ for all $l \in \{1, \dots, L\}$. Similarly, we will consider the same transmit power for each replica, which yields $P_{l,i} = \frac{P_{tot}}{k}$ for all $l \in \{1, \dots, L\}$ and $i \in \{1, \dots, N\}$. With the slight abuse of notation, this allows to simplify (B.9) as

$$
p_{out} = \left.\left(F_{\mathrm{MI}} * \underbrace{f_{\mathrm{MI}} * \cdots * f_{\mathrm{MI}}}_{k-1}\right)(x)\right|_{x=R}
\tag{B.12}
$$

where the CDF $F_{\mathrm{MI}}$ and all $k-1$ pdfs $f_{\mathrm{MI}}$ are defined as in (B.8), (B.7) and with identical parameters $C_l$, $\gamma_l$. In a similar manner, the Chernoff bound

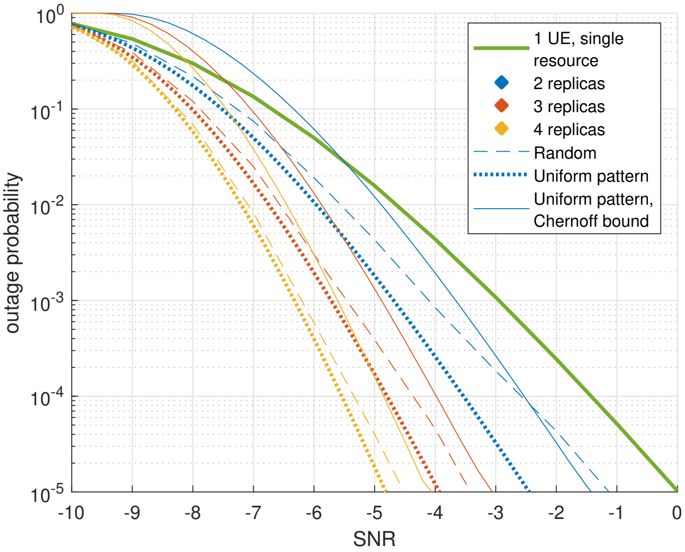


**Fig. B.3:** Outage probability performance of uncoordinated grant-free (idealized) and with pre-assigned patterns.

simplifies to

$$p_{out} \leq \min_{t>0} e^{tR} \left(M_{\text{MI}}\left(-t\right)\right)^{k} \tag{B.13}$$

In Fig.B.3 we compare the performance of the idealized grant-free with random selection (no pilot collisions) and the just described approach based on uniform patterns. The parameters used are the same as before with $L = 12$, $N = 24$, $M = 8$ and no antenna correlation. We can see that not only we were able to recover the performance of the idealized random scheme but even improve on it. This is because the ability to coordinate interference allows to avoid too heavily congested TF-blocks and protects against the most detrimental cases where $C_l > M$. With the solid lines of the appropriate color we provide also the Chernoff upper bound on the outage probability of the scheme with deterministic patterns. The bound offers reasonably good approximation by being around 1dB from the actual curve and the gap decreases with higher number of replicas.

## 4.3 Transmissions with shared diversity resources

In addition to the regular patterns evenly utilizing all resources, in this work we extend also the concept originally introduced in [12] coined transmissions with shared diversity resources. There, the main idea was to distribute available TF-blocks in such a way that each UE had one dedicated, interference-free slot and additional transmissions were performed on the remaining $L-N$ diversity resources. Clearly, such a scheme requires at least as many TF-blocks as the total number of users, which does not scale very well with the number of served UEs (though such approach was also justified by the fact that only a single receive antenna was considered).

The fact that we consider a BS with multiple antennas allows for relaxing the requirement of fully dedicated resources. To that end, we divide the TF-blocks into two parts: $L_L$ blocks having lower contention and $L_H$ blocks with higher contention ($L_l + L_H = L$). Consequently, each UE will be assigned an access pattern which consists of $k_L$ transmissions located in the first part and $k_H$ transmissions somewhere in the second part ($k_L + k_H = k$). The contention levels for the two types of slots can be calculated in a similar way as before and with these new parameters are

$$C_T = \frac{k_T N}{L_T}, \quad T \in \{L, H\}. \tag{B.14}$$

The example shown in Fig. B.1 can be viewed as one instance of this scheme where $L_L = 4$, $L_H = 4$, $k_L = 1$, $k_H = 2$, $C_L = 1$ and $C_H = 2$.

Due to the introduced asymmetry we will also consider an unequal power allocation: namely users will transmit the two types of packets with powers $P_L$ and $P_H$ respectively. The modified outage probability (B.9) corresponding with the described scheme is given by

$$\begin{aligned} p_{out} &= \Pr\left\{ R > \sum_{l \in \mathcal{L}_{L_i}} \mathrm{MI}_{l,i} + \sum_{l \in \mathcal{L}_{H_i}} \mathrm{MI}_{l,i} \right\} \\ &= \left( F_{\mathrm{MI}_L} * \underbrace{f_{\mathrm{MI}_L} * \cdots * f_{\mathrm{MI}_L}}_{k_L - 1} * \underbrace{f_{\mathrm{MI}_H} * \cdots * f_{\mathrm{MI}_H}}_{k_H} \right)(x) \Bigg|_{x=R} \end{aligned} \tag{B.15}$$

where in a similar manner as before we denote the indices of low and high contention slots of user $i$ with $\mathcal{L}_{L_i}$ and $\mathcal{L}_{H_i}$ respectively. In the last expression, $F_{\mathrm{MI}_L}$ and first $k_L - 1$ $f_{\mathrm{MI}_L}$'s are evaluated with parameters $C_l = C_L$, $\gamma_l = \frac{P_L}{\sigma^2}$ and the next $k_H$ pdfs are evaluated with parameters $C_l = C_H$, $\gamma_l = \frac{P_H}{\sigma^2}$, The optimal power allocation minimizing the outage probability (B.15) can be found by solving the problem

$$\min_{\{P_L, P_H\}} \quad p_{out} \tag{B.16a}$$

$$\text{s.t.} \quad k_L P_L + k_H P_H = P_{tot} \tag{B.16b}$$

Across the range of scenarios considered for this paper, we found the powers $P_L$, $P_H$ obtained through (B.16) to be only slightly different from the equal

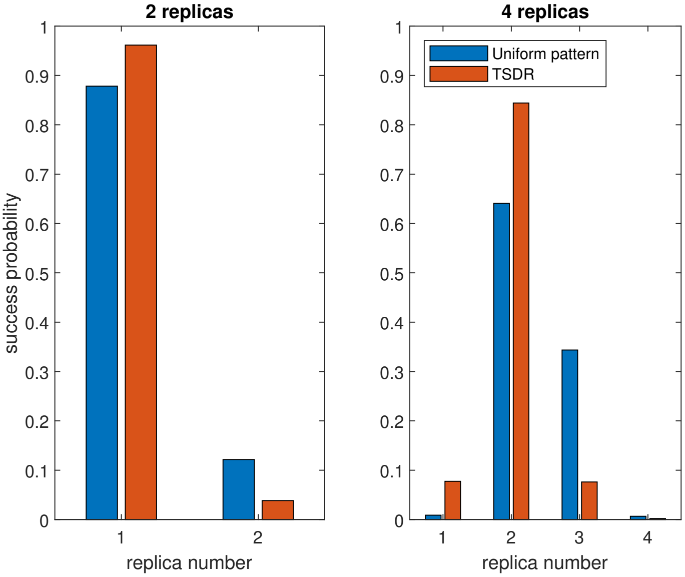


**Fig. B.4:** Probability of decoding with a given replica number

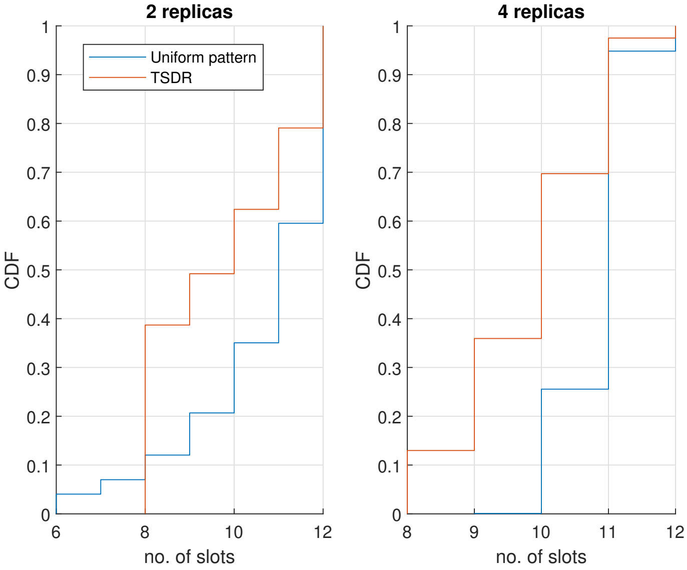


**Fig. B.5:** CDF of the number of slots required until all packets are decoded

power allocation case. Namely, for the $P_{tot}$ in the range of interest[4] (i.e. providing outage probability $10^{-3}$ or lower) $\frac{P_L}{P_H}$ is between 1.1 and 1.2 which corresponds to their absolute values being around 5% to 10% off from the uniform $P_l = \frac{P_{tot}}{k}$.

Next, we compare the performance of the TSDR access scheme with optimal powers to that of the uniform access patterns described in Subsection 4.2. In terms of outage probability, TSDR performs 0.1dB - 0.2dB worse than the scheme with uniform patterns which is a marginal difference.

However, the asymmetric patterns turn out to offer some other, less obvious benefits, which we will now demonstrate. For the purpose of the subsequent discussion we downselected two representative scenarios (2 and 4 replicas) which have the following parameters. In case of both 2 transmissions and 4 transmissions the split between resources is $L_L = 8$ and $L_H = 4$ while $k_L = k_H = 1$ in the former and $k_L = k_H = 2$ in the latter scenario. Consequently, the low and high contention levels are $C_L = 3$, $C_H = 6$ and $C_L = 6$, $C_H = 12$ respectively.

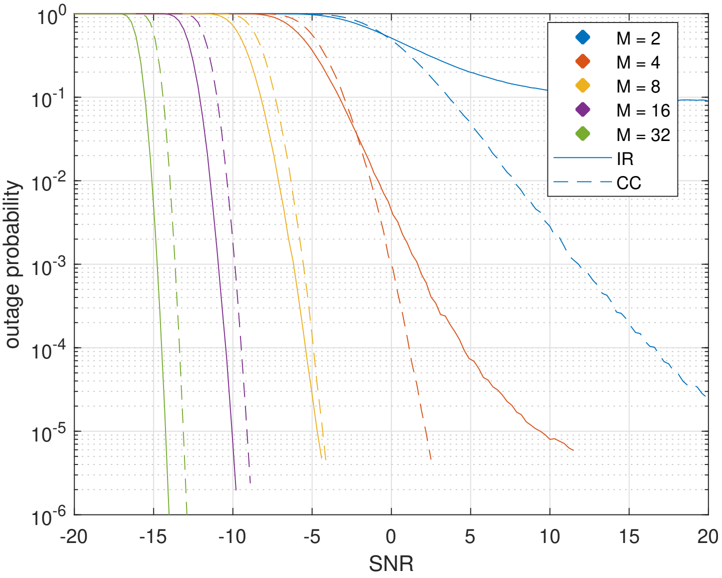
In Fig. B.4 we show the probability of decoding a packet with a given replica number. The immediate observation is that TSDR, which employs asymmetric patterns, yields a much higher chance to decode the packet early. For instance, when UEs transmit 4 times, BS will need more than two replicas only $\sim$ 7% of the time when TSDR is used compared to $\sim$ 35% with the uniform scheme. From the practical point of view, these results translate to lower effective latency of individual packets for TSDR. The two factors responsible for this effect are the higher $P_L$ and (simultaneously) lower $C_L$.

In Fig. B.5 we present the CDF of the number of received TF-blocks required to decode all packets. This type of performance can be viewed as an indicator of two other metrics. One is the total, system-wide latency and the other is the complexity as each additional TF-block entails more processing. Again, the approach based on asymmetric power and patterns offers tangible gains.

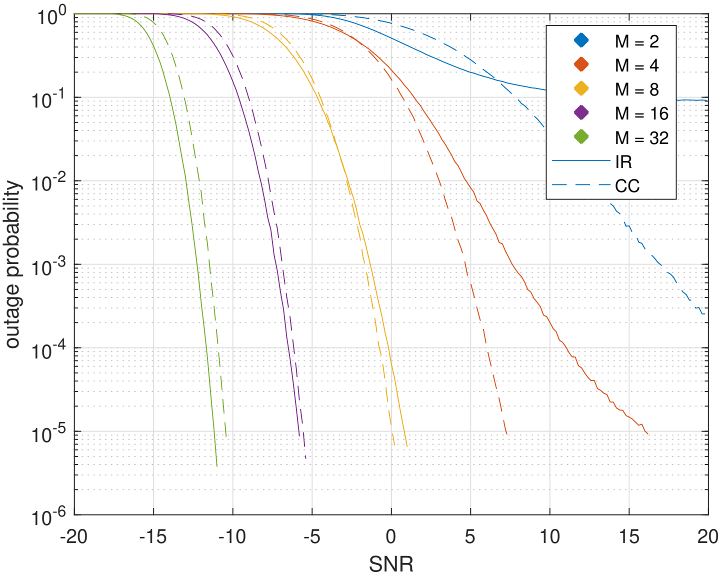
Lastly, in Fig. B.6 we investigate the impact of the antenna correlation and their total number on the TSDR scheme. The chosen scenario is the one with 4 replicas, and non-uniform access patterns with contention levels $C_L = 6$ and $C_H = 12$. With low number of antennas (less than the contention level) the performance is degraded and exhibits plateauing, which is to be expected as the BS receiver does not have the required degrees of freedom to separate all transmissions. An interesting case is the one with $M = 8$ as we arrive at a situation where for some transmissions $C_l = C_H > M$, and yet the penalty in terms of outage probability is not as prominent as one would expect.

In the situations with fewer antennas (or high correlation between them,

[4] For lower $P_{tot}$ the difference between powers is more significant, however this is not the region of operation relevant to URLLC



(a)



(b)

**Fig. B.6:** The effect of varying number of receive antennas $M$ in case of (a) no correlation and (b) $\rho = 0.85$

which reduces their effective number) the performance can be recovered to some extent by switching from Incremental Redundancy to the Chase Combining mode of operation. Since in CC all replicas of the packet are identical, then instead of considering matrices $\mathbf{H}_l$ from each slot individually, they can be stacked together to obtain a single $\mathbf{H} \in \mathbb{C}^{LM \times N}$ similarly as in [12]. This way, the transmissions are processed jointly based on a total of $LM$ measurements rather than by solving many underdetermined problems. By comparing Fig. B.6(a) with B.6(b) to assess the impact of correlation, we can conclude that the degradation at $10^{-5}$ outage ranges from 3dB for $M = 32$ antennas, up to 7dB for $M = 4$. Again, the case with $M = 8$ is the most interesting as it shows that with a reduction in the effective number of antennas, CC becomes preferable to IR.

# 5 Conclusions

In this work we have studied the reliability aspects of MIMO URLLC systems with grant-free access, operating in either uncoordinated manner or with preassigned channel resources. Using recently derived results on the SINR distribution of MMSE multi-antenna receivers, we have derived analytical results and bounds on the outage probability of the studied schemes under different conditions. Our results show that, although a totally uncoordinated scheme performs well when perfect CSI is assumed, preallocation of the channel resources provides an effective way to avoid pilot sequence collisions and to limit the maximum contention levels in each slot. In addition, we have also found that dividing the resource pool into two types of resources, with low and high contention levels, can help reducing the receive processing latency with virtually no loss in terms of reliability. Overall, we conclude that the combination of multi-antenna processing at the receiver with intelligent design of the preallocated resources can significantly boost the performance of URLLC systems, even in the presence of strong receive antenna correlation.

# A Appendix

Let $d_{l,i}$ denote a specific sequence chosen by user $i$ who is active in slot $l$. Furthermore, let us denote by $\mathcal{J}_l$ a subset of indexes of the active UEs who selected the same sequence as some other UE (e.g. if users 1,2,3,5,7,9 transmitted in the same TF-block $l$ and: 1, 3, 7 used the same sequence $a$; 2 and 5 used sequence $b$; 9 used sequence $c$ then $\mathcal{J}_l = \{1, 2, 3, 5, 7\}$). More formally $\mathcal{J}_l = \left\{ i : (\exists j) \left[ d_{l,i} = d_{l,j} \ \wedge \ P_{l,i}, P_{l,j} > 0 \right] \right\}$. We assume that the receiver is not able to estimate the channel coefficients of the users involved in pilot collisions. As a consequence, the corresponding transmissions are lost and become a part of the noise.

From the signal processing point of view, we deal with this case by defining a new matrix $\mathcal{H}_l$ which is the original $\mathbf{H}_l$ with columns $\mathcal{J}_l$ set to $\mathbf{0}$. Note, that the optimal MMSE detector in this case is also different [15] and has a form $\mathbf{F}_l = \left(\mathcal{H}_l{}^{\mathrm{H}}\mathbf{\Sigma}_l^{-1}\mathcal{H}_l + \mathbf{I}_N\right)^{-1}\mathcal{H}_l{}^{\mathrm{H}}\mathbf{\Sigma}_l^{-1}$ where $\mathbf{\Sigma}_l = \sigma^2\mathbf{I} + \sum_{a\in\mathcal{J}_l}(\mathbf{h}_l)_a(\mathbf{h}_l)_a^{\mathrm{H}}$ is the new $\mathbb{C}^{M\times M}$ noise covariance matrix with $(\mathbf{h}_l)_a$ being the columns of $\mathbf{H}_l$. This matrix is in fact unknown due to the assumption stated earlier, however in a simplified scenario with no antenna correlation $\mathbf{\Sigma}_l$ becomes diagonal with $i$-th diagonal element equal to $\sigma^2 + \sum_{a\in\mathcal{J}_l}|(\mathbf{H}_l)_{i,a}|^2$ (which requires from the BS only the knowledge of the total magnitude of the combined noise-plus-interference).

# Acknowledgement

The work of Radosław Kotaba and Petar Popovski was funded by the European Research Council (ERC), Consolidator Grant Nr. 648382 (WILLOW), under the European Union Horizon 2020 research and innovation program.

# References

[1] 3GPP, “NR;NR and NG-RAN Overall description; Stage-2,” 3rd Generation Partnership Project (3GPP), TS 38.300, 2019, v15.7.0.

[2] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, “Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture,” *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[3] O. Kaiwartya, A. H. Abdullah, Y. Cao, A. Altameem, M. Prasad, C.-T. Lin, and X. Liu, “Internet of Vehicles: Motivation, Layered Architecture, Network Model, Challenges, and Future Aspects,” *IEEE Access*, vol. 4, pp. 5356–5373, 2016.

[4] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, “Fundamental tradeoffs among reliability, latency and throughput in cellular networks,” in *2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1391–1396.

[5] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[6] 3GPP, "Study on latency reduction techniques for LTE," 3rd Generation Partnership Project (3GPP), TR 36.881, 2016, v14.0.0.

[7] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 182–185, 2018.

[8] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: applying codes on graphs to design random access protocols," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 144–150, 2015.

[9] M. C. Lucas-Estañ, J. Gozálvez, and M. Sepulcre, "On the Capacity of 5G NR Grant-Free Scheduling with Shared Radio Resources to Support Ultra-Reliable and Low-Latency Communications," *Sensors (Basel, Switzerland)*, vol. 19, 2019.

[10] T. Jacobsen, R. Abreu, G. Berardinelli, K. Pedersen, P. Mogensen, I. Z. Kovacs, and T. K. Madsen, "System Level Analysis of Uplink Grant-Free Transmission for URLLC," in *2017 IEEE Globecom Workshops (GC Wkshps)*, 2017, pp. 1–6.

[11] N. H. Mahmood, R. Abreu, R. Böhnke, M. Schubert, G. Berardinelli, and T. H. Jacobsen, "Uplink Grant-Free Access Solutions for URLLC services in 5G New Radio," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*, 2019, pp. 607–612.

[12] R. Kotaba, C. Navarro Manchón, T. Balercia, and P. Popovski, "Uplink Transmissions in URLLC Systems With Shared Diversity Resources," *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 590–593, 2018.

[13] C. Boyd, R. Kotaba, O. Tirkkonen, and P. Popovski, "Non-Orthogonal Contention-Based Access for URLLC Devices with Frequency Diversity," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019, pp. 1–5.

[14] H. Lim and D. Yoon, "On the Distribution of SINR for MMSE MIMO Systems," *IEEE Transactions on Communications*, vol. 67, no. 6, pp. 4035–4046, 2019.

[15] T. Brown, E. De Carvalho, and P. Kyritsi, *Practical Guide to the MIMO Radio Channel*. John Wiley & Sons Ltd, 2012.

# Paper C

## Non-Orthogonal Contention-Based Access for URLLC Devices with Frequency Diversity

Christopher Boyd, Radosław Kotaba, Olav Tirkkonen and Petar Popovski

The paper has been published in the
*IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019.

# Abstract

*We study coded multichannel random access schemes for ultra-reliable low-latency uplink transmissions. We concentrate on non-orthogonal access in the frequency domain, where users transmit over multiple orthogonal subchannels and inter-user collisions limit the available diversity. Two different models for contention-based random access over Rayleigh fading resources are investigated. First, a collision model is considered, in which the packet is replicated onto $K$ available resources, $K' \leq K$ of which are received without collision, and treated as diversity branches by a maximum-ratio combining (MRC) receiver. The resulting diversity degree $K'$ depends on the arrival process and coding strategy. In the second model, the slots subject to collisions are also used for MRC, such that the number of diversity branches $K$ is constant, but the resulting combined signal is affected by multiple access interference. In both models, the performance of random and deterministic repetition coding is compared. The results show that the deterministic coding approach can lead to a significantly superior performance when the arrival rate of the intermittent URLLC transmissions is low.*

***Keywords***— URLLC, grant-free, coded random access, MRC

# 1 Introduction

Machine-Type Communication (MTC) is one of the main technologies in 5th Generation (5G) mobile communication. Within this very general category, we can distinguish two main use cases [1] with widely differing requirements —massive MTC (mMTC), capable of supporting a large number of sporadically communicating devices, possibly battery-operated, and Ultra-Reliable Low Latency Communication (URLLC) enabling mission-critical MTC. Of the two, the latter especially has been igniting researchers' imaginations, as it would enable the implementation applications previously unattainable and considered futuristic, such as self-driving cars, remote surgery and telemetry, and more [2].

While downlink communications in a cellular setting is fairly flexible, a radical change in the uplink access protocol might be necessary in order to fulfill the requirements of the more demanding MTC use cases. One solution is communication based on random access. For mMTC, this is motivated by the sporadic, infrequent traffic patterns which require energy efficient protocols, and the fact that the control overhead involved in establishing the connection significantly exceeds the amount of actual data to be transmitted. In URLLC, traffic also contains elements of randomness and is characterised by intermittent activation, but the random access is a means of achieving low latency levels, which could not be possible with the scheduling request/grant procedure in place. However, a solution based on random access is inherently unreliable, as it is subject to interference and collisions, so the random access for URLLC needs to be augmented by other mechanisms that introduce redundancy to compensate for unavoidable collisions.

Multiple technologies to improve reliability of random access have been recently studied. Coded random access [3, 4] improves throughput and reliability by exploiting repetition coding and interference cancellation. Diversity slotted ALOHA has been considered by 3GPP as a potential solution for grant-free access [5], while in [6, 7], the possibility to increase access reliability by preassigning non-orthogonal access sequences to users has been investigated. In [8], a similar idea of preassigned patterns is treated, but with focus on the performance of successive interference cancellation (SIC) with imperfect channel state information (CSI). These technologies can be collectively called $K$-repetition schemes.

In this paper we explore the diversity aspects of random access schemes based on packet repetition. If access opportunities that are exploited in a $K$-repetition scheme are independently fading, e.g., if access packets are transmitted over distinct frequency domain resource blocks experiencing different fading conditions, the reliability of communication is improved by diversity gains, in addition to the possible collision mitigation benefits. However, due to the fact that devices access the medium in a random manner, possibly causing collisions, this leads to a communication channel where *the diversity degree is a random variable*, governed by the arrival process of other users. We treat such a communication model by analyzing the probability distribution of the diversity degree, distribution of the contention level (number of simultaneous interferers) and finally total outage probability, which is a metric especially relevant in URLLC context. We compare the performance of a simple receiver utilizing a destructive collision-model, which discards all overlapping packets, with a more advanced system capable of optimally combining the replicas based on their signal-to-interference-plus-noise-ratio (SINR). Furthermore, we provide such analysis for the two coding approaches: uncoordinated, random selection of subchannels and deterministic assignment of patterns to the users. The latter technique is based on a code construction given by a Steiner system, as in [7, 9].

## 2 System Model

We consider a communication system where $N$ URLLC users attempt to randomly access the uplink resources of a centralized receiver. Users active during a single timeslot transmit in an uncoordinated, grant-free fashion, and employ $K$-repetition coding of their access packets over $M$ orthogonal frequency subchannels. Users are slot synchronized to the receiver, and access packets are of equivalent size and occupy an entire subchannel. The users

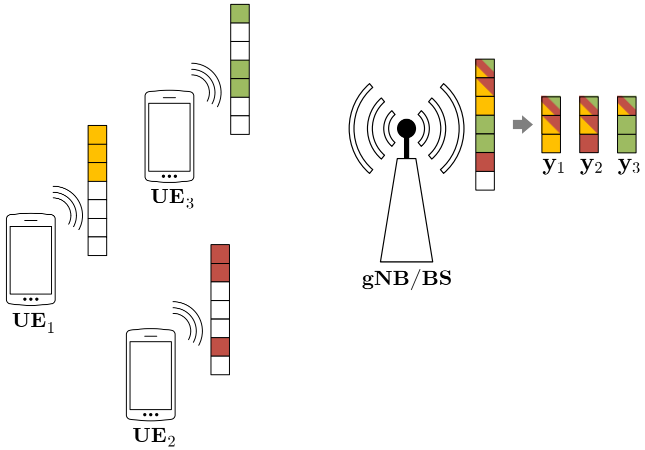


**Fig. C.1:** An example of the system with $N = 3$ active users, $M = 7$ frequency subchannels and $K = 3$ repetitions of each packet. Users 1 and 2 manage to have a single interference-free replica, while the $3^{rd}$ user has two. The subchannels where collisions occur might be used by the base station to further increase the reliability if a more advanced receiver processing is available.

are assumed to become active randomly, such that the number of users active during a time instance follows a Poisson process with intensity $\lambda$.

Repetition coding of access packets provides robustness to inter-user collisions and facilitates diversity gain, which are integral for reliability in contention-based access over fading channels. Here we consider two approaches to coding: (i) users transmit $K$ packet replicas randomly over the slotted frequency resources, as in ALOHA-type schemes, and (ii) users transmit their replicas according to a deterministic and uniquely preallocated pattern from a designed access code.

Two receiver models are investigated: (i) a destructive collision model, a PHY layer abstraction to the MAC layer which assumes that colliding packets are lost and only interference-free packets may be correctly received, and (ii) a multi-user interference (MUI) model, where all packets are used to decode the signal but their contribution depends on their effective SINRs. In this paper, maximum-ratio combining is applied to the different packet replicas to achieve this.

The received complex baseband signal corresponding to the symbol $x_j$ transmitted by a device $j$ reads

$$\mathbf{y}_j = \mathbf{h}_j x_j + \sum_{l=1}^{L_j} \mathbf{g}_{j,l} z_{j,l} + \mathbf{n}_j = \mathbf{h}_j x_j + \mathbf{i}_j, \tag{C.1}$$

where $L_j$ is a random number of interferers perceived by user $j$, $\mathbf{h}_j, \mathbf{g}_{j,l} \in \mathbb{C}^K$ are the channel gains of the signal of interest and its $l$-th interferer respectively (which are assumed to be known at the receiver), $z_{j,l} \in \mathbb{C}$ are the interfering symbols, $\mathbf{n}_j \in \mathbb{C}^K$ is the additive white Gaussian noise (AWGN) with zero mean and variance $N_0$, and $\mathbf{i}_j \in \mathbb{C}^K$ is the joint interference-plus-noise term. Note, that interferers might occupy only some of the slots of $j$, in which case the remaining entries of $\mathbf{g}_{j,l}$ are 0. We define the linear filter

$$\mathbf{f}_j = \mathbf{W}_j \mathbf{h}_j \tag{C.2}$$

where $\mathbf{W}_j = \mathrm{diag}[w_{j,1}, w_{j,2}, .., w_{j,K}]$ is a diagonal matrix of real-valued reliability weights that depend on the combining strategy. Applying the combiner yields

$$r_j = \mathbf{f}_j^H \mathbf{y}_j. \tag{C.3}$$

By assuming uncorrelated interference, the post combining SINR of $j$-th user's signal can be approximated by

$$\gamma_j = \frac{\left|\mathbf{h}_j^H \mathbf{W}_j \mathbf{h}_j\right|^2}{\mathrm{E}\left\{\left|\mathbf{h}_j^H \mathbf{W}_j \mathbf{i}_j\right|^2\right\}} = \frac{\left(\sum_{k=1}^K w_{j,k} \left|h_{j,k}\right|^2\right)^2}{\sum_{k=1}^K w_{j,k}^2 \left|h_{j,k}\right|^2 \left(\sum_{l=1}^{L_j} \left|g_{j,l,k}\right|^2 + N_0\right)},$$

where the reliability weights for MRC in the multi-user interference model that maximize $\gamma_j$ are given by

$$w_{j,k} = \frac{1}{\sum_{l=1}^{L_j} |g_{j,l,k}|^2 + N_0}. \tag{C.4}$$

In the destructive collision model only those $0 \leq K' \leq K$ replicas which were received collision-free can be combined together. Let us further denote by $\mathcal{I}_j$ a subset of indices which correspond to those packets. Then, the signal model can be simplified since $\mathbf{i}_j = \mathbf{n}_j$ and

$$w_{j,k} = \begin{cases} 1 & \text{for } k \in \mathcal{I}_j \\ 0 & \text{otherwise} \end{cases}, \tag{C.5}$$

resulting in the final signal-to-noise-ratio (SNR)

$$\gamma_j = \sum_{k \in \mathcal{I}_j} \frac{|h_{j,k}|^2}{N_0}. \tag{C.6}$$

In the remainder of this paper we will often discuss a signal from the perspective of a single device and omit the index $j$ whenever it does not create ambiguity.

# 3 Available Diversity after Collisions

Consider a timeslot in which a given user $U$ of population $N \geq 1$ is active and transmitting $K$ packet replicas randomly over the $M$ access resources, along with $N-1 \sim \text{Poisson}(\lambda)$ simultaneously active users. The probability that $K'$ of $K$ replicas are received without interference depends on the coding strategy.

## 3.1 Diversity Slotted ALOHA

In diversity slotted ALOHA (DSA), users transmit their $K$ packet replicas over the $M$ subchannels randomly, following a uniform distribution. From the perspective of user $U$, the probability that the $N-1$ other users collide is such a way that $K'$ of $K$ subchannels are unoccupied by packet replicas follows from the classical occupation problem, and is given by

$$p_r(K'|N) = \binom{K}{K_{\text{diff}}} \sum_{n=0}^{K_{\text{diff}}} (-1)^n a_n X_n^{N-1} \ , \tag{C.7}$$

where $K_{\text{diff}} = K - K'$, $a_n = \binom{K-K'}{n}$, $X_n = \binom{M-K'-n}{K} / \binom{M}{K}$, $p_r(K' \neq K|1) = 0$, and $p_r(K|1) = 1$.

The probability that user $U$ occupies $K'$ interference-free subchannels, conditioned on the arrival process, is

$$\begin{aligned} p_r(K') &= \sum_{N=1}^{\infty} p_r(K'|N) p(N-1) \\ &= \binom{K}{K_{\text{diff}}} \sum_{N=1}^{\infty} \sum_{n=0}^{K_{\text{diff}}} (-1)^n a_n \frac{(X_n \lambda)^{N-1}}{(N-1)!} e^{-\lambda} \end{aligned} \tag{C.8}$$

## 3.2 Designed Codes

Designed and uniquely preallocated user codes have been shown to outperform the random coding approach in a URLLC context [6, 10]. Such codes limit the number of supportable users in order to coordinate the interference over that population. The performance of combinatorial code designs such as Steiner system $S(t, K, M)$ as random access codes has been explored in [7, 9]. Here we consider Steiner $t = 2$ designs, as their highly symmetric structure makes for ready analysis. Note that $t > 2$ designs may produce significantly larger codes, and therefore be more practical in systems that need to support many devices simultaneously (which is typically not the case in URLLC).

Consider the scenario where each of the $N$ users is uniquely allocated a repetition pattern from a $S(2, K, M)$ code [11]. The maximum supportable

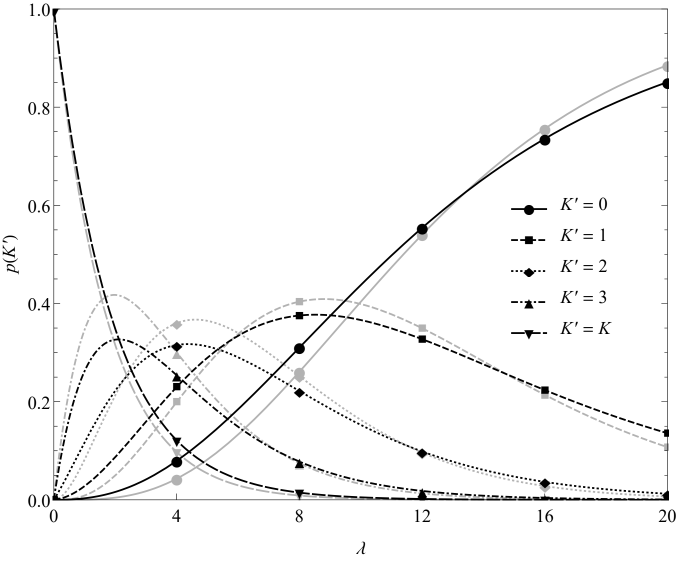


**Fig. C.2:** Probability distribution of the available diversity for DSA (black) and Steiner (grey) with $M = 25$ and $K = 4$ under Poisson arrivals.

user population is limited to $C = |S(2, K, M)| = M(M-1)/K(K-1)$. However, the structure of the code ensures that the number of users that may overlap with user $U$ in a single subchannel is at most $D = (M-K)/(K-1)$. When user $U$ is active, the $N-1$ simultaneously active users will be employing repetition patterns from the remaining $C-1$, of which at most $kD$ can overlap user $U$ in $k$ slots. The probability that user $U$ has $K'$ of $K$ diversity branches post collisions is therefore

$$p_{\text{det}}(K'|N) = \binom{K}{K_{\text{diff}}} \sum_{n=0}^{K_{\text{diff}}} (-1)^n a_n Y_n \, , \tag{C.9}$$

where $Y_n = \binom{(C-1)-D(n+K')}{N-1} / \binom{C-1}{N-1}$, and the same restrictions on $N$ as in (C.7) apply. The probability $p_{\text{det}}(K')$ can be found similarly to (C.8), but with $p_{\text{det}}(K'|N)$ in place of $p_r(K'|N)$.

Figure C.2 compares the probability of $K'$ diversity branches being available post collisions for the DSA scheme over $M = 25$ subchannels with repetition factor $K = 4$, and a deterministic coding scheme employing a $(2, 4, 25)$ Steiner system with $C = 50$, as a function of the arrival intensity $\lambda$. Evident in the plot is how, at lower intensities, the Steiner code trades-off the probability of producing the best outcome ($K = K'$) to increase the probability of good outcomes (e.g. $K = 3$), and reduce the probability of the worst outcome ($K' = 0$). Since this worst outcome is especially detrimental in the collision model, we can expect significant deterministic gain in this intensity region. As $\lambda$ approaches $M$, the structure of the deterministic code becomes a disadvantage. If the URLLC users are activated in an intermittent and uncorrelated manner, then the expected number of users simultaneously active during a given slot is $\lambda << M$.

# 4 Interferers per Subchannel

Consider again a user $U$ transmitting during a timeslot with $N-1$ other users. In the case of weighted MRC, we are interested in the number of packets from $L$ interferers present in the subchannels occupied by $U$. Let $0 \leq L' \leq L$ be the number of independently Rayleigh faded packets from the $N-1$ interfering users *in a given subchannel* of user $U$. The probability distribution of $L'$ depends on the coding strategy.

## 4.1 Diversity Slotted ALOHA

In DSA, the repetition coding procedure amounts to users independently selecting one of the $\binom{M}{K}$ possible binary patterns with replacement. As such, the maximum number of interferers observed by $U$ in one subchannel is $N-1$. The probability of $L'$ interfering packets in a given subchannel occupied by user $U$ is given by

$$p_r(L'|N) = \frac{\binom{M-1}{K}^{N-1-L'}}{\binom{M}{K}^{N-1}} \binom{M-1}{K-1}^{L'} \binom{N-1}{L'} . \tag{C.10}$$

Note that this is an approximation assuming the interferers select the subchannels independently. As in (C.8), the probability $p_r(L')$ is found by marginalizing over Poisson distributed $N$.

## 4.2 Designed Codes

With a finite set of $C$ deterministic access patterns, the probability that user $U$ sees $L'$ interferers in a given subchannel in which it is active, is the probability that $L'$ of the $N-1$ other users are using patterns from the $D$ that overlap in that subchannel. Since patterns are uniquely preallocated, selection from the $C-1$ available codes is done without replacement. The random variable $L'$ therefore follows the hypergeometric distribution, such that

$$p_{\text{det}}(L'|N) = \binom{C-1}{N-1}^{-1} \binom{D}{L'} \binom{C-1-D}{N-1-L'} , \tag{C.11}$$

and $p_{\text{det}}(L')$ is found as in (C.8).

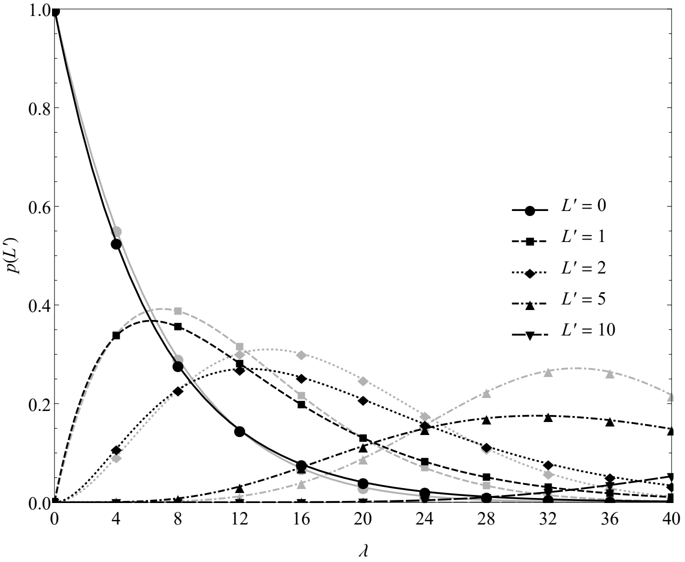


**Fig. C.3:** Probability distribution of the number of interferers in a subchannel for DSA (black) and Steiner (grey) with $M = 25$ and $K = 4$ under Poisson arrivals.

Figure C.3 compares the probability distributions of $L'$ for the DSA scheme over $M = 25$ subchannels with $K = 4$ and the $(2, 4, 25)$ Steiner code with $C = 50$, as functions of $\lambda$. Here, the Steiner codes slightly increases the probability of the best outcome ($L = 0$) at lower intensities, but decreases it as $\lambda$ approaches $M$. More pronounced is how the Steiner code increases the probability of lower numbers of interferers in certain windows of intensity (e.g. $\lambda \in [4, 20]$ for $L' = 1$, or $\lambda \in [8, 27]$ for $L' = 2$).

# 5 Diversity Combining

Lastly, let us analyze the outage probability performance of the different schemes. The outage probability is defined as the probability that the post-processing SINR $\gamma$ falls below a certain threshold $\theta$, i.e

$$p_{\text{out}} \quad = \quad p(\gamma < \theta) \ . \tag{C.12}$$

This metric will depend on both the coding technique as well as the applied receiver processing.

## 5.1 Collision Model

After discarding the packets which experienced collision, the remaining $K'$ replicas transmitted by user $U$ can be combined by the receiver. Assuming perfect CSI is available, and the SNR of a single packet is exponentially distributed (following the Rayleigh fading assumption), the post processing SNR has the distribution

$$p(\gamma) = \sum_{K'=0}^{K} p(\gamma|K')p_c(K') = \sum_{K'=0}^{K} \frac{1}{(K'-1)!}\frac{\gamma^{K'-1}}{\Gamma^{K'}}e^{-\gamma/\Gamma}p_c(K') \,, \tag{C.13}$$

where $\Gamma$ is the expected SNR per packet and $p_c(K')$ is either $p_r(K')$ or $p_{det}(K')$ depending on the scenario. The (C.13) follows from the fact that the sum of $K'$ exponentially distributed random variables with scale $\Gamma$ is $Gamma(K', \Gamma)$ distributed.

## 5.2 Multi-User Interference Model

In the case of MUI, obtaining closed form expressions of even the conditional SINR distribution is not feasible, as it quickly becomes intractable (i.e. for more than one interferer):

$$p(\gamma|L') = \int_{N_0\gamma}^{\infty} f_{\text{exp}}\left(x|\Gamma\right) f_{\text{gamma}}\left(\frac{x}{\gamma} - N_0|L', \Gamma\right) dx \,. \tag{C.14}$$

Furthermore, the full distribution would require marginalizing the convolution of individual SINRs of the replicas over all $N$ and all possible realizations of $L'_1, ..., L'_K$, that is:

$$\begin{aligned} p(\gamma) = \sum_{N=1}^{\infty} \sum_{L'_1}^{N-1} \cdots \sum_{L'_K}^{N-1} &\left(p(\cdot|L'_1) * \cdots * p(\cdot|L'_K)\right)(\gamma) \\ &\times p\left(L'_1, \ldots, L'_K|N\right) p(N) \,. \end{aligned} \tag{C.15}$$

To obtain $p(\gamma)$, and eventually $p_{\text{out}}$, for the multi-user interference model we resort to simulation. We generate multiple instances of $(N, \mathbf{H})$, i.e. number of transmitting devices, their channel gains and patterns accordingly (DSA or Steiner), and evaluate the effective SNR according to (C.4) and (C.4). Figures C.4 and C.5 show the outage probability performance as function of the SINR threshold $\theta$, for the MRC receiver in the collision and MUI models, with random and deterministic repetition coding. Additionally, included in the plots is the outage probability for a white noise approximated match filter (WN-MF), for which the reliability weights in (C.2) are set to $w_k = 1$ for all packets.

For low access intensity, represented here by Figure C.4 with $\lambda = 0.5$, the gains offered by deterministic codes are significant compared to the uncoordinated traffic, e.g. at $\theta = 5$dB the difference in offered reliability is more than an order of magnitude (and further two orders of magnitude compared to the white noise approximation). This gain diminishes as the intensity of

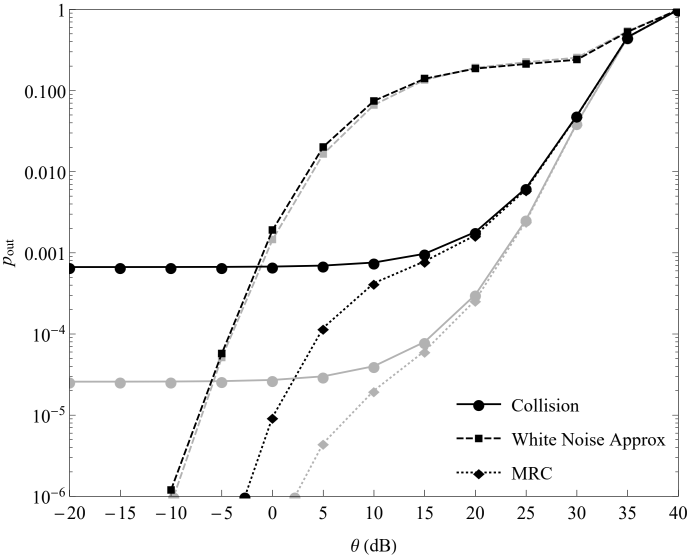


**Fig. C.4:** Probability of outage for DSA (black) and Steiner (grey) with $M = 25$, $K = 4$ and $\Gamma = 30$ dB, under Poisson arrivals with $\lambda = 0.5$.

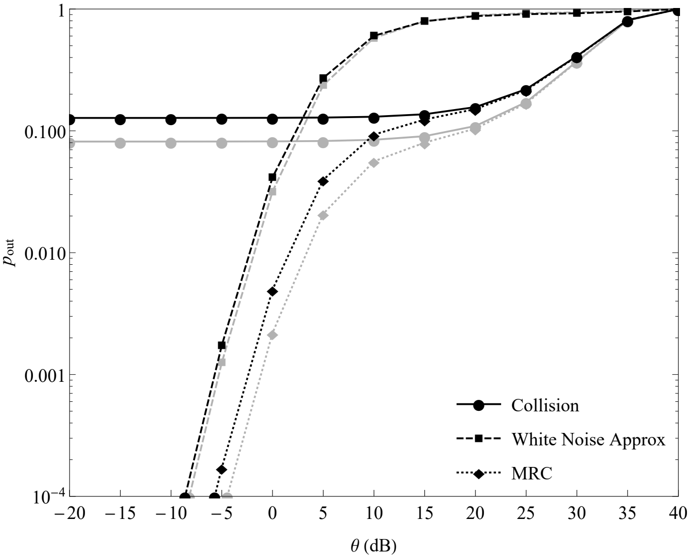


**Fig. C.5:** Probability of outage for DSA (black) and Steiner (grey) with $M = 25$, $K = 4$ and $\Gamma = 30$ dB, under Poisson arrivals with $\lambda = 5$.

traffic increases (cf. Figure C.5), since the structure of the Steiner code becomes irrelevant as the channel becomes flooded with packet replicas. The collision model exhibits a clear error floor related to $p(K' = 0)$.

With regard to required processing and complexity, the receiver in the collision model is the simplest—it only needs to detect whether or not there were collisions in the subchannels occupied by a given user, and measure the SNR of each of their $K'$ interference-free packets. The MRC receiver, however, requires accurate measurements of the channel gain (and phase), as well as a precise estimate of the interference and noise corrupting each packet. The white noise approximation requires precise channel information for each packet, but does not need an interference plus noise estimate. The white noise approximation shows to what degree neglecting interference plus noise whitening is detrimental to the performance of the MRC receiver. In collision channels of the type discussed here, noise plus interference whitening is of prime importance.

# 6 Conclusion

We have presented a study of multichannel random access mechanisms for supporting an uplink URLLC transmissions from a set of uncoordinated devices. The study treats two different models, one with destructive collisions and another where the interfered slots can also be used to contribute to the overall SINR through a combining process. Furthermore, we compared two different types of repetition coding: fully random, and utilizing deterministic access patterns, respectively. The latter, designed according to a Steiner system, leads to a significantly better outage performance (more than one order of magnitude) when the arrival rate of URLLC packets is low, and between 2-3 times better for moderate arrival rates. This is enabled by the properties of Steiner codes, which allow to coordinate the users and limit the probability of instances with particularly unfavourable interference conditions.

An interesting direction for future work can be identified in the model with non-destructive collisions. Namely, the current combining algorithm does not take into account that the interference is created from signals that are also packet replicas, just from different users. This fact can be used to devise a multi-user decoding, based on e.g. MMSE receiver. Going further, even more advanced receivers could be employed that are capable of cancelling the interference from different users, similar to the mechanisms applied in coded random access.

## Acknowledgement

This work was funded in part by the Academy of Finland (grant 299916) and EIT ICT (HII:ACTIVE) and by the European Research Council (ERC), Consolidator Grant Nr. 648382 (WILLOW), under the European Union Horizon 2020 research and innovation program.

## References

[1] P. Popovski, K. F. Trillingsgaard, O. Simeone, and G. Durisi, "5G Wireless Network Slicing for eMBB, URLLC, and mMTC: A Communication-Theoretic View," *IEEE Access*, vol. 6, pp. 55 765–55 779, 2018.

[2] 3GPP, "Service requirements for the 5G system; Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.261, 2018, v16.6.0.

[3] E. Paolini, C. Stefanovic, G. Liva, and P. Popovski, "Coded random access: applying codes on graphs to design random access protocols," *IEEE Communications Magazine*, vol. 53, no. 6, pp. 144–150, 2015.

[4] J. Choi, "Throughput Analysis for Coded Multichannel ALOHA Random Access," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1803–1806, 2017.

[5] B. Singh, O. Tirkkonen, Z. Li, and M. A. Uusitalo, "Contention-Based Access for Ultra-Reliable Low Latency Uplink Transmissions," *IEEE Wireless Communications Letters*, vol. 7, no. 2, pp. 182–185, 2018.

[6] C. Boyd, R. Vehkalahti, and O. Tirkkonen, "Interference Cancelling Codes for Ultra-Reliable Random Access," *International Journal of Wireless Information Networks*, vol. 25, pp. 422–433, 2018.

[7] ——, "Grant-Free Access in URLLC with Combinatorial Codes and Interference Cancellation," in *2018 IEEE Globecom Workshops (GC Wkshps)*, 2018, pp. 1–5.

[8] R. Kotaba, C. Navarro Manchón, T. Balercia, and P. Popovski, "Uplink Transmissions in URLLC Systems With Shared Diversity Resources," *IEEE Wireless Communications Letters*, vol. 7, no. 4, pp. 590–593, 2018.

[9] G. T. Peeters, R. Bocklandt, and B. Van Houdt, "Multiple Access Algorithms Without Feedback Using Combinatorial Designs," *IEEE Transactions on Communications*, vol. 57, no. 9, pp. 2724–2733, 2009.

[10] E. Paolini, G. Liva, and A. Graell i Amat, "A structured irregular repetition slotted ALOHA scheme with low error floors," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[11] La Jolla Covering Repository. Steiner Systems. [Online]. Available: https://ljcr.dmgordon.org/cover.php.

[12] O. Galinina, A. Turlikov, S. Andreev, and Y. Koucheryavy, "Multichannel random access with replications," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2538–2542.

# Paper D

## Deterministic Patterns for Multiple Access in Latency-Constrained Ultra-Reliable Communications

Radosław Kotaba, Roope Vehkalahti, Čedomir Stefanović, Olav Tirkkonen and Petar Popovski

The paper has been submitted to
IEEE Transactions on Communications, *arXiv:2202.09890*

# Abstract

*The grant-free access is envisioned as one of the enablers of the ultra-reliable low-latency communications. Yet, when there are many devices that tend to be active only intermittently, the fully orthogonal resource allocation is largely inefficient. The solution is to employ a common, shared pool of resources and account for the fact that some collisions and interference will inevitably occur. In this contribution we study the reliability aspects of such multi-user uplink communication scenario over a shared pool of channel resources, where intermittently active devices utilize multiple transmissions (K-repetition coding) to achieve diversity. We focus on two access methods – one where devices choose the K slots at random and one where the access patterns are deterministic and follow a specific code design, namely the Steiner System. We analyze the problem under two signal models that involve different complexity for the receiver. Firstly, a model which treats collisions as destructive, i.e. only those K′ among K transmissions that do not contain interference can be used and combined. Second, where receiver is capable of utilizing all K replicas and applies maximum ratio combining (MRC) treating interference as noise. Furthermore, in both cases we investigate the receiver with and without successive interference cancellation (SIC) capabilities. As one of the main contributions of this work, we develop useful approximations and bounds for the outage probabilities in the aforementioned scenarios that match very closely the simulation results. We also show that deterministic patterns have the potential to significantly outperform fully random selection, both in terms of raw performance and by simplifying the system design.*

***Keywords***— grant-free, radio resource management, uplink, ultra reliable low latency communication (URLLC)

# 1 Introduction

The latest generation of wireless systems, 5G networks, are becoming more and more widely adopted [1], while the researchers and the industry already plan their next steps by laying ground for technologies that will come next [2]. Importantly, the shift to 5G and beyond is not just about the need for higher data rates, but is driven by a new types of use cases and applications for which the support is required from the network. Among those, particularly relevant and, simultaneously, challenging are the use cases that fall under the category of ultra-reliable low-latency communications (URLLC). These are characterized by especially stringent end-to-end (E2E) latency constraints (between $0.5 - 2$ ms) and reliability (i.e. the probability of successful delivery of a packet) of 99.999% [3]. Among the most prominent URLLC applications are those that involve tactile interaction, intelligent transport and factory automation [3, 4].

The primary challenge in designing ultra-reliable systems with stringent latency constraints is to not overly compromise their spectral efficiency in the process [5]. This is particularly difficult to achieve in the uplink, which in traditional networks is centrally managed by the base station (BS) and relies on either explicit grants (more efficient but with high latency due to excessive signalling) or pre-allocation of resources (low latency but inefficient when the traffic is intermittent). As such, in order to fulfill the demanding latency and reliability targets, new uplink access protocols and modes of operation have to be devised for 5G and beyond.

One way to tackle the problem is to rely on random access based communication. Within that family, perhaps the most well-known approach is the grant-free access [6], in which user equipments (UEs) are allowed to transmit without prior, explicit scheduling. Instead, a certain portion of bandwidth is delineated and provided to a group of users who can use it whenever they have data to send. The benefit of that is significant reduction in the signalling overhead and connected with it latency. As a matter of fact, scheduling contributes the most to the E2E delay making it the main bottleneck when designing URLLC systems [7]. However, it should be noted, that grant-free as a solution is particularly suitable for and motivated by traffic that is relatively infrequent and irregular [8]. Due to the sharing of the resources and lack of coordination it is inherently less reliable than its grant-based counterparts and without explicit control from the BS, the uplink signals of URLLC users are prone to collisions and interference. Clearly, to compensate for that, additional mechanisms which will improve the reliability are needed.

The simplest and at the same time most effective solution is to introduce redundancy through multiple transmissions [9]. This improves the reliability in two ways. Firstly, by increasing the chance that at least some of the replicas reach the BS uninterfered, and secondly, by providing diversity that allows to combat the negative effects of the fading channel. The second technique builds upon the concept of multiple transmissions. Instead of letting UEs select the resources from the pool completely at random, the idea is to structure the transmissions of the individual users into *access patterns*. These access patterns can be constructed in many ways and with different goals in mind, but in general they aim to provide certain reliability guarantees [10]. The drawback of such solution is that their assignment requires some coordination with the BS and signalling, which makes it less flexible than purely random selection. However, this operation can be integrated into the registration procedure that each device has to perform anyway when it first attaches to the BS or wakes up and re-synchronizes after being inactive for a prolonged period. Furthermore, even in a fully random scheme the device needs to be configured at least with the number of repetitions and portion of bandwidth where the grant-free pool is located.

Lastly, on the receiver side, there is a possibility to implement successive interference cancellation (SIC). With SIC, it is possible to iteratively decode signals by gradually removing the interference. Namely, in each round the packets that were successfully decoded in the preceding rounds can be subtracted (after proper equalization) from the received signal, thus improving the signal-to-interference-plus-noise (SINR) of the remaining ones. This is a powerful technique and especially relevant when dealing with traffic that is non-orthogonal by design [11].

## 1.1 Related work

Fundamentally, the grant-free techniques descend from one of them most well known concepts in the field of communication - slotted ALOHA [12]. Since its inception, many extensions have been proposed. One of them is the Content Resolution Diversity Slotted ALOHA [11, 13], which utilizes multiple transmissions (with the goal of achieving diversity) and the interference cancellation. In [14], authors analyze a variant of this scheme - Irregular Repetition Slotted ALOHA (IRSA) in which the number of repetitions is not fixed, but follows a certain discrete distribution. Furthermore, in [15] the analysis is extended to the Rayleigh fading channel and optimization of the repetition degree is presented. Another extension coined Coded Slotted ALOHA [16] involves transmitting different coded version of the packet (redundancy versions) rather than exact replicas, which allows to achieve better granularity in terms of transmission rate. In [17] the author analyzes the throughput of CSA in a multichannel Rayleigh fading scenario.

In the above works, the primary focus is on the maximization of spectral efficiency of the systems rather than achieving high reliability. The grant-free access methods as envisioned for 5G and designed specifically for URLLC have been thoroughly researched in [18]. In its thesis, author investigates different repetition and retransmission schemes in realistic scenarios based on detailed system level simulator. In [19], authors focus on the combinatorial aspects of the repetition-based 5G grant-free scheme, namely its probability of collisions, and evaluate achievable reliability and latency levels as a function of the number of UEs, amount of pre-allocated resources and number of replicas. Comparison to other schemes has been shown in [20], where authors jointly evaluate repetition coding, its proactive version (where the UEs have the possibility of early termination), and the more traditional Hybrid Automatic Repeat Request (HARQ) based on feedback and retransmissions. Most recently, in [21] the author extends the idea of repetition-based schemes towards network coding, making it better suited for scenarios where devices have more than one packet to transmit at a time. A different approach is considered in [22], where the resources are first assigned to a group of users based on sensing, and then between themselves UEs avoid the collisions by signalling transmission announcements.

In addition to fully random grant-free transmission schemes, some authors have investigated the access based on pre-allocated patterns and their optimal design. In [23], which is one of the earliest works, the pattern construction is based on combinatorial design. However, authors do not consider SIC and treat all the collisions as destructive. More recently, the designs oriented towards interference cancellation have been considered in [24] [25] and [26]. In [24] the patterns belong to the class of $(\leq M, 1, n)$-locally thin codes, in [25] are based on the $(t, K, M)$ Steiner Systems, while in [26] authors use LDPC codes. The idea to use deterministic access patterns also appeared in [27], where they are used in conjunction with multiple antenna processing at the BS. Furthermore, the channel resources are divided into high and low contention parts over which power optimization is additionally considered. Differently from the aforementioned, in [28] the patterns, are applied on the symbol-level rather than over slots.

## 1.2 Contributions

In this work we study the grant-free multiple access in which users apply access patterns, i.e. sequences consisting of multiple redundant transmissions, to achieve ultra-reliable communication. The framework involves a shared pool of resources - a short, periodic frame composed of limited number of slots, that makes our contribution relevant in scenarios with tight latency constrains, especially URLLC. We focus on the comparison between fully random selection of slots and a case of pre-assigned, deterministic patterns that are inspired by a specific code construction, known as the Steiner system. The latter is chosen due to its desirable properties, namely its construction ensures that two patterns can share at most $t - 1$ slots, where $t$ is a design parameter. In other words, it provides guarantees in terms of the amount of collisions/interference.

As a main contribution we provide a thorough analysis of the grant-free access based system in terms of its outage probability performance and spectral efficiency. In our analysis we consider two different signal models. One, resembling a traditional ALOHA system, where collisions are *destructive*, i.e. only the slots that contain a transmission of a single device can be used. Unlike in traditional ALOHA however, we allow the receiver to combine multiple collision-free replicas from a given user. In the second model, receiver is capable of utilizing all transmissions and performs their maximum ratio combining (MRC) accounting for different SINRs. Furthermore, for each of the two signal models we consider two subcases: with and without SIC at the receiver. In all of the aforementioned configurations access patterns given by a Steiner system exhibit clear gains over their fully random counterparts. Furthermore, their regular structure simplifies the overall system design. As such, we believe that Steiner systems make for a compelling solution in the design of grant-free access.

A particularly important contribution are the approximations and bounds developed for the collision model with SIC and Full MRC without SIC - two considerably non-trivial cases! The developed expressions match very closely the extensive simulation results obtained with Monte Carlo methods. The approximations are especially valuable in the context of URLLC, where relying on simulations alone is often not feasible due to the sheer number of samples required. To the best of the authors knowledge, these results are novel and have not been reported before.

This work extends our prior contribution [25] in several meaningful ways. Firstly, we provide an in-depth analysis and develop approximations and bounds that go well beyond the results reported earlier. We also present formal proofs of the combinatorial results in [25] that treat the distribution of the number of collision-free slots and number of interferers, and which were omitted due to space constraints. Secondly, we extend the scenario by considering receiver with SIC capabilities. We also broaden the scope by considering other Steiner systems with different parameters (frame length and number of repetitions). Lastly, we discuss the limitations of access methods based on Random selection highlighting issues with their practical implementation.

The rest of the paper is organized as follows. We start by introducing the system and signal model in Section 2. In Section 3 we introduce the two types of access patterns and discuss their properties. Then, in Section 4 we consider different receiver processing techniques and provide their thorough analysis in the context of the access patterns from Section 3. This is complemented by both analytical results and corresponding simulations. In Section 5, we discuss the deficiencies of the Random selection approach and the challenges wrt. its practical implementation. In Section 6 we compare different Steiner Systems using the analytical results derived earlier. Lastly, in Section 7 we offer final conclusions that close the paper.

## 2 System model

We consider a communication system with a single base station (BS) serving a population of $N$ intermittently active users (UEs) transmitting in the uplink. The access channel is composed of periodic frames, which are further broken down into $M$ access opportunities otherwise known as slots[1]. We assume that UEs are independently activated in a frame with probability $b$, such that the total number of active devices $U$ follows a binomial distribution $f_{\text{bin}}(u; b, N) = \binom{N}{u} b^u (1-b)^{N-u}$. Whenever active, a user selects $K$ out of $M$ slots in a frame and uses them to transmit its packet, thus employing a form

[1] Although in Fig. D.1 the slots seem to be arranged in time, this is not a requirement. The slots can also represent different frequencies/groups of frequencies (subcarriers) or be 2-dimensional constructs (similar to Resource Blocks in LTE/5G).

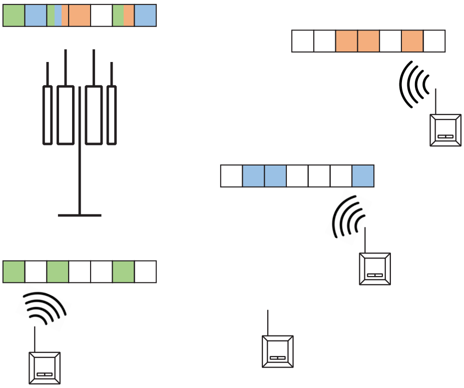


**Fig. D.1:** Example of the uplink access scenario with $K = 3$ multiple transmissions over a frame of $M = 7$ slots. There are $N = 4$ UEs out of which $U = 3$ happen to be active. Their transmissions cause collisions in slots 3 and 6.

of $K$-repetition coding. It is further assumed that all UEs transmit with the same rate $R$ measured in bits per channel use (c.u.). The described model is visualized in the example in Fig. D.1. In general $M$ is determined by the allowed latency, while $K$ is a design parameter that depends on the number of users $N$ and the activation probability $b$.

In this work we consider a Rayleigh block fading channel, where the realizations of channel coefficients are independent across slots and UEs. At the receiver, the baseband representation of the channel output in a frame[2] is modeled as

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w} \tag{D.1}$$

$$= \mathbf{G} \circ (\mathbf{V}\mathbf{A}\mathbf{P})\mathbf{x} + \mathbf{w} \in \mathbb{C}^{M\times 1} \tag{D.2}$$

where $\mathbf{x} \in \mathbb{C}^{N\times 1}$ is the vector of complex modulated symbols transmitted by users such that $\mathrm{E}\left[|x_n|^2\right] = 1$, $\mathbf{w} \in \mathbb{C}^{M\times 1}$ is the additive white Gaussian noise (AWGN) with zero mean and variance $\sigma^2$ and $\mathbf{H} \in \mathbb{C}^{M\times N}$ are the channel gains between the $N$ users and the base station in each of the $M$ slots. In (D.2), $\circ$ denotes entry-wise (Hadamard) product. The channel gains $\mathbf{H}$ can be represented as a product of $\mathbf{G} \in \mathbb{C}^{M\times N}$ - which is zero-mean, unit-variance, circularly-symmetric complex Gaussian (ZMCSCG) and models the underlying uncorrelated Rayleigh flat fading channel, $\mathbf{V} \in \{0,1\}^{M\times N}$ - matrix representing the access patterns of all users such that $V_{m,n} = 1$ if user

[2] We assume that the transmissions within a single frame are self contained and independent, i.e., UEs are not supposed to transmit their packets over multiple frames as that would violate the implicit latency constraint. Consequently, there is no need to introduce additional index to denote the frame number in the signal model.

$n$ is assigned to slot $m$, $\mathbf{A} \in \{0,1\}^{N \times N}$ - the diagonal matrix designating which users are active in a frame and $\mathbf{P} = \text{diag}\left((P_1)^{1/2}, \ldots, (P_N)^{1/2}\right)$ which is a diagonal matrix of square roots of average powers, i.e. signal amplitudes applied by UEs[3].

In the remainder of this work we assume that all active UEs use the same power $P_x$. Consequently, we can define the average received SNR

$$\theta = \frac{P_x}{\sigma^2}. \tag{D.3}$$

For a given user $n$, the choice of the slots which are used for transmission, i.e. the set of indices $\{j : V_{j,n} = 1\}$ constitute what we call an access pattern. The matrix of slot selections $\mathbf{V}$ may be fixed, such that UEs follow access patterns that were pre-assigned to them, or it may be random. Furthermore, since each UE uses only $K$ among $M$ available slots we have that $\sum_{m=1}^{M} V_{m,n} = K$.

# 3 Access Patterns: Random vs. Deterministic

In this section we introduce, and later on compare, two access methods that could be employed by the devices which try to communicate over a shared pool of resources (i.e., slots). In the following we describe one which relies on random selection, and one in which users have pre-assigned, deterministic access patterns.

## 3.1 Random selection

We start with an approach in which users transmit their $K$ packet replicas over the $M$ available slots by selecting slots uniformly at random. We are interested in determining the probability of having a certain number of interference-free slots, i.e. replicas which do not experience collisions with other UEs' replicas.

Consider a frame in which $U \geq 1$ out of a population of $N$ users is active and, without loss of generality, focus on a single, arbitrary user $u$.

**Lemma 1.** *From the point of view of a user $u$, the probability that the remaining $U - 1$ users do not cause collisions to exactly $K'$ out of $K$ of its replicas (collision-free*

[3] Note that a distance-dependent path loss term is absent in the signal model. Throughout this work we assume that UEs know the long-term statistics of the channel and based on that compensate for path loss accordingly. If we were to denote by $d_n$ the path loss of user $n$, then the actual transmit power would be $d_n P_n$ such that $P_n$ represents the average received power. Due to this compensation and for the sake of simplicity, we decide to omit the path loss term altogether.

*(CF)) when using random patterns (R) is given by*

$$p_{\mathrm{CF,R}}(K'|U) = \binom{K}{K'} \sum_{n=0}^{K-K'} (-1)^n a_n V_n, \tag{D.4}$$

*where* $a_n = \binom{K-K'}{n}$ *and* $V_n = \left(\binom{M-K'-n}{K} / \binom{M}{K}\right)^{U-1}$.

The proof of Lemma 1 can be found in Appendix B.

Another relevant metric when considering such a contention-based access and which may have an impact on the decoding is the distribution of the number of interferers $L$ in a given slot in which the reference user is active. Since users are free to select any of the $\binom{M}{K}$ possible sequences and do so independently from each other (with replacement) we have that $L \in [0, U-1]$ and

$$p_{\mathrm{I,R}}(L|U) = \frac{\binom{M-1}{K}^{U-1-L} \binom{M-1}{K-1}^{L} \binom{U-1}{L}}{\binom{M}{K}^{U-1}}. \tag{D.5}$$

## 3.2 Deterministic patterns

Another approach to the contention-based access over a shared pool of resources is the one in which UEs have fixed, pre-assigned access patterns. Such a solution is less flexible, as it requires coordination with the BS, who is responsible for assigning them, however it has the potential to greatly improve the overall reliability of the system. Typically, a pattern would be assigned when the device registers with the BS for the first time or wakes up and re-synchronizes after being in power-efficient mode. They can be also periodically updated. This is the case, for example, when the user population size changes and the resource pool needs to be adjusted; however such updates will occur relatively infrequently compared to the duration of the frame.

In this work we choose to focus on the patterns that are given by a specific block design known as Steiner system. A Steiner system $S(t, K, M)$ can be considered as a $M$-dimensional constant-weight code, where each codeword has $K$ ones, and for any two codewords $s_i, s_j \in S(t, K, M)$, $s_i \neq s_j$, we have $d(s_i, s_j) \geq 2K - 2(t-1)$, where $d(\cdot, \cdot)$ is the Hamming distance. In other words, two codewords (transmission patterns) can collide on at most $t-1$ positions.

For fixed $K$ and $M$, the lower the $t$, the smaller the codebook size and thus the number of supportable users. Specifically, we have that $C = |S(t, K, M)| = \binom{M}{t} / \binom{K}{t}$. Since in this work our focus is on URLLC applications, we limit our considerations to the case $t = 2$ as it provides high reliability and the support for a massive number of devices in not required[4].

Another property of the Steiner system is that its structure guarantees that the number of overlapping users in any given slot is at most $D = \binom{M-1}{t-1}/\binom{K-1}{t-1}$, i.e. there are exactly $D$ access patterns which include a certain slot, so even in the worst case scenario when all of them are active there are at most $D$ mutually interfering users. As we will elaborate later on, this is an important feature, as it allows to dedicate just the right amount of resources for the pilot sequences and ensure that no pilot collisions occur.

Analogously to the Random selection, we have the following results for the Steiner system.

**Lemma 2.** *With devices employing access patterns from a $S(t, K, M)$ Steiner system, the probability that an arbitrary user has $K'$ out of $K$ collision-free slots is*

$$p_{\mathrm{CF,S}}(K'|U) = \binom{K}{K-K'} \sum_{n=0}^{K-K'} (-1)^n a_n W_n \tag{D.6}$$

*where $W_n = \binom{(C-1)-(D-1)(n+K')}{U-1}/\binom{C-1}{U-1}$.*

The proof of the above Lemma is provided in Appendix C.

In terms of the number of interferers $L$, the situation is much more straightforward. In a slot in which an arbitrary user is active, there are only $D-1$ other patterns that could cause a collision and the selection is done without replacement due to the unique preassignment. Hence, it is the probability that $L$ out of $U-1$ devices select one of them while the rest of the devices select any of the remaining $C-D$ patterns:

$$p_{\mathrm{I,S}}(L|U) = \frac{\binom{D-1}{L}\binom{C-D}{U-1-L}}{\binom{C-1}{U-1}} \,. \tag{D.7}$$

In Fig. D.2 we compare a $S(2, 4, 25)$ Steiner system (solid line) and a corresponding Random selection (dashed) with the same frame length and number of repetitions.

In terms of the number of collision-free slots $K'$ shown in Fig. D.2(a), the main conclusion is that Steiner system reduces the probability of having the best ($K' = K$) and worst ($K' = 0$) outcome, while increasing the probability of having 'good' outcomes such as $K' = K-1$, $K' = K-2$. The ability to avoid the worst case scenarios is particularly important as the probability of $K' = 0$ is tied to the performance floor. Clearly, when all replicas are subject to collisions, increasing SNR is not effective and without a single packet that can be decoded, SIC cannot be applied. In that regard, the structure of the Steiner code becomes a disadvantage as the traffic intensity increases. However, as

[4] Technically, the highest reliability is provided when $t = 1$, however such case is trivial as it corresponds to the fully orthogonal allocation of resources.

will become evident later, in most cases ultra-reliability cannot be achieved if the traffic intensity is too high, regardless if the access scheme is based on Steiner system or Random selection. As such, we note that the region of interest is primarily low and medium traffic intensity, where the average number of activated users $bN < \frac{M}{2}$.

In Fig. D.2(b) and Fig. D.2(c) we compare the probability distribution of the number of interferers $L$, which has an impact on the SINR in the slots where collisions occur, as well as the utility of the SIC procedure. In other words, having more interferers makes it less likely that all of them can be removed. Once again, the Steiner system ensures that within the traffic intensities of interest, 'good' outcomes such as $L = 0, 1, 2$ are more likely, while really congested slots are rare. In fact, by inspecting Fig. D.2(c) one can see that unless $bN = 20$ or higher, the cdf of $L$ for the Steiner system is strictly above that of the Random selection. Furthermore, as already mentioned, with Steiner system the number of interferers is strictly limited to $D - 1$, which in this case is 7.

# 4 Receiver Processing

In this section we analyze different modes and processing techniques employed at the receiver. The metric on which we are focusing is outage probability, as it is particularly relevant for URLLC use cases. We define it as

$$p_{\mathrm{out}_i} = \Pr\left\{R > \log_2\left(1 + \mathrm{SINR}_i\right)\right\} \tag{D.8}$$

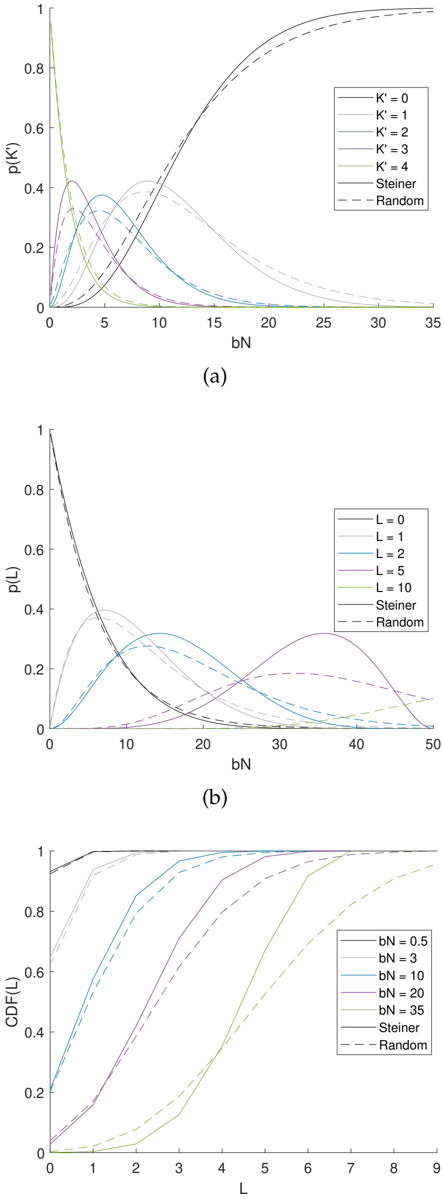
where $R$ is the transmission rate in bits per channel use at which the packet is encoded and $\mathrm{SINR}_i$ denotes the final, post-processing signal-to-interference-plus-noise ratio of user $i$'s packet. The exact definition and the means of computing it depend on the chosen scenario, which are the subject of the following subsections.

## 4.1 Collision model

We start with a simple model that entails a less computation-intensive processing method at the receiver. In the collision model, collisions are assumed to be *destructive*, so only the slots containing a transmission of a single device are considered. When the slot $m$ is interference free with only user $i$ transmitting, the received complex baseband signal in (D.2) simplifies to

$$y_m = \sqrt{P_x} g_{m,i} x_i + w_m \tag{D.9}$$

and the SNR of that signal is $\rho_{m,i} = \frac{P_x |g_{m,i}|^2}{\sigma^2} = \theta |g_{m,i}|^2$. Since channel coefficients are Rayleigh distributed r.v.'s, the SNR of each packet follows exponential distribution $f_{\exp}(\rho; \theta) = \frac{1}{\theta} e^{-\frac{\rho}{\theta}}$. As each device transmits $K$ times,



(a)

(b)

(c)

**Fig. D.2:** Comparison of the Steiner system (solid line) and Random selection (dotted) with $K = 4$ and $M = 25$ in terms of their distributions of interference free slots $K'$ and number of interferers $L$. In the first two subfigures, the x-axis represents the mean traffic intensity $bN$. The third subfigure is a CDF of $L$.

there might be up to $K' \leq K$ collision-free replicas, the probability of which is given by (D.4) and (D.6), and it is possible to combine the signals to improve the overall SNR. Denote by $\mathcal{J}_i$ the set of indices corresponding to the uninterfered transmissions of device $i$. Using maximum ratio combining (MRC), we obtain

$$\sum_{j\in\mathcal{J}_i} g_{j,i}^* y_j = \sqrt{P_x} x_i \sum_{j\in\mathcal{J}_i} |g_{j,i}|^2 + \sum_{j\in\mathcal{J}_i} g_{j,i}^* w_j \tag{D.10}$$

that yields the SNR $\rho_i = \theta \sum_{j\in\mathcal{J}_i} |g_{j,i}|^2$. As a sum of exponentially distributed r.v.'s, the total SNR after combining, conditioned on $K'$, has a gamma distribution $f_{\text{gam}}(\rho; K', \theta) = \frac{1}{\Gamma(K')\theta^{K'}} \rho^{K'-1} e^{-\frac{\rho}{\theta}}$.

As follows from (D.8), the decoding is unsuccessful whenever $\rho_i < 2^R - 1$, hence

$$p_{\text{out}}(R, \theta, U) = p_{\text{CF}}(0|U) + \sum_{K'=1}^{K} F_{\text{gam}}(2^R - 1; K', \theta) p_{\text{CF}}(K'|U) \tag{D.11}$$

where $p_{\text{CF}}(\cdot|U)$ is given by either (D.4) or (D.6), depending on whether random or Steiner patterns are used, respectively. The equation can be further marginalized over $U$ to account for the specific activation process.

In Fig. D.3 we present the results based on (D.11) that show the performance of deterministic patterns based on Steiner system, and random selection. The parameters chosen are: $M = 25$ slots, $K = 4$ repetitions and $N = C = 50$ users. The activation probabilities are $b = [0.02, 0.04, 0.1, 0.2]$ that translate to the mean number of active devices in a frame equal to $[1, 2, 5, 10]$ respectively. The transmission rate is $R = 2 \frac{bit}{c.u.}$. It becomes clear, that the properties of the Steiner system which were described in the previous section lead to tangible gains in terms of outage probability (up to an order of magnitude lower than Random selection). Nevertheless, even with the traffic intensity as low as 1 UE per frame, ultra-reliability is unattainable unless more sophisticated processing is applied.

## 4.2 Collision model with successive interference cancellation

A natural extension to the model put forward in the previous subsection is to introduce successive interference cancellation (SIC). SIC involves removing from the received signal the packets which were successfully decoded, including all of their $K$ replicas. This has the potential to greatly improve the performance, as it allows to remove the interference from the slots where collisions occurred and make them usable in the subsequent iterations of the decoding process.

A rigorous analysis of the models involving SIC is known to be inherently difficult [16] and the exact analytical results typically do not exist except for

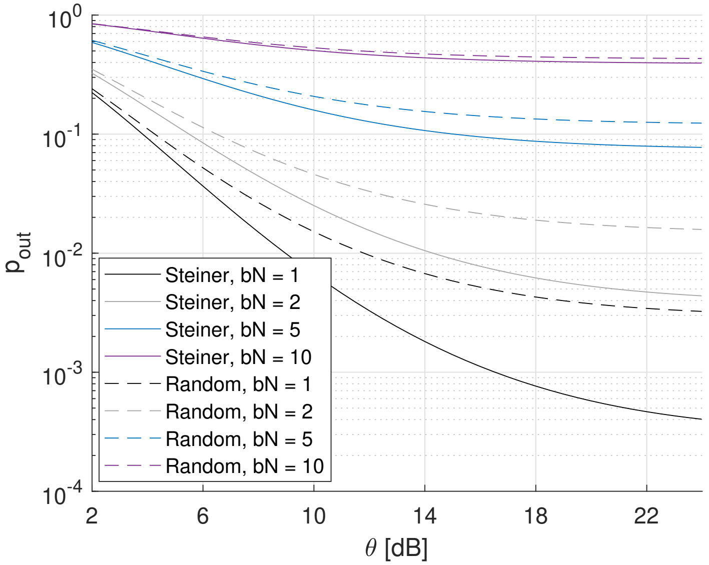


**Fig. D.3:** Outage probability performance of the system employing random and deterministic patterns as a function of the average received SNR for different mean number of active devices $bN$.

asymptotic cases and some simple special cases. Due to the multitude of possible configurations of the selected transmit patterns and different dependencies they create, the problem becomes intractable already when $U > 4$. This motivates us to look for approximate results. In the presented approach we consider the first few rounds of SIC. In each, we condition on the number of decoded users in the preceding rounds. Due to the unique structure of the Steiner system, which ensures that active users cover the frame uniformly, it is possible to simplify some of the steps by using averages. Indeed, rather than having to sum/integrate over possible outcomes of many variables, we can work with the averages in terms of the number of collision-free slots, the combined SNR, etc. This is in contrast to Random selection, where the covering is uniform with high probability only when there are many active users, while for low $U$ their replicas can be quite concentrated. Consequently, our approach is suitable only for Steiner systems and will not provide a good approximation if the patterns follow a Random selection.

Similarly as before, the analysis will be performed from a point of view of an arbitrary user. First, consider the case that $l_1 = 1, \ldots, U - 1 - S$ users are decoded in the first iteration of SIC. Here, $S$ denotes the number of users who for some reason are excluded from the procedure and cannot be cancelled (this will be explained in detail later on). If we treat all the $U - 1 - S$ users independently, each with a probability of outage

$p_{\text{out}}(R,\theta,U))$ given by (D.11), the distribution of $l_1$ can be approximated[5] as binomial $f_{\text{bin}}(l_1;U-1-S,1-p_{\text{out}}(R,\theta,U))$. For the remaining users that were not successful in the first round of SIC, it is important to account for the fact that they nevertheless accumulated some of the power already. Since they failed, their SNRs, given $K'$, are drawn from a truncated gamma distribution $\frac{f_{\text{gam}}(\rho;K',\theta)}{F_{\text{gam}}(2^R-1;K',\theta)}$. By taking its mean and marginalizing over $K'$, we can determine the mean residual SNR, i.e. the amount of signal power that is missing before the packet can be decoded:

$$\rho_{\text{res}} = \sum_{K'=0}^{K}\left(2^R-1-\theta\frac{\gamma\left(\frac{2^R-1}{\theta},K'+1\right)}{\gamma\left(\frac{2^R-1}{\theta},K'\right)}\right)p_{\text{CF}}(K'|U) \tag{D.12}$$

where $\gamma(s,x)=\int_0^x t^{s-1}e^{-t}dt$ is a lower incomplete gamma function. Furthermore, let us also define an expected number of collision-free slots per user, which is simply $\widehat{K}(U)=\sum_{K'=0}^{K}K'p_{\text{CF}}(K'|U)$. Since all the collision-free slots from the first round have been already taken into account, we are only interested in the new ones that were released after cancelling the $l_1$ successful users. On average, there will be $\widehat{K}(U)-\widehat{K}(U-l_1)$ new slots, so the probability of decoding a packet in the second round of SIC is

$$p_{\text{out},2}(l_1) = F_{\text{gam}}\left(\rho_{\text{res}};\widehat{K}(U)-\widehat{K}(U-l_1),\theta\right) \tag{D.13}$$

Similarly, we then consider the number of additional messages that can be decoded in the second iteration $l_2=0,1,...U-1-S-l_1$ which is given by $f_{\text{bin}}\left(l_2;U-1-S-l_1,1-p_{\text{out},2}(l_1)\right)$. We halt this procedure at the third iteration. At this point the device in focus observes the system with $U-l_1-l_2$ devices (including itself), however, since there was no attempt to decode its packet yet, it is subject to $p_{\text{out}}(R,\theta,U-l_1-l_2)$. By marginalizing over $l_1$ and $l_2$, the outage probability conditioned on $S$ is then

$$\begin{aligned} &p_{\text{out,SIC}}(R,U|S)\\ &\quad= p_{\text{out}}(R,\theta,U)^{U-S}+\sum_{l1=1}^{U-1-S}f_{\text{bin}}\big(l_1;U-1-S,1-p_{\text{out}}(R,\theta,U)\big)\\ &\qquad\times\sum_{l2=0}^{U-1-S-l_1}f_{\text{bin}}\big(l_2;U-1-S-l_1,1-p_{\text{out},2}(l_1)\big)\,p_{\text{out}}(R,\theta,U-l_1-l_2) \end{aligned} \tag{D.14}$$

where the first term corresponds to the case in which all users fail and SIC cannot proceed.

[5] Note that in general the decoding events are not independent. Eq. (D.11) is a weighted mean of all realizations of $K'$, however it is not possible to have a situation with 2 active users where $K'_1 \neq K'_2$.

The expression (D.14) with $S = 0$ approximates very well the simulation results when the number of active devices $U$ is low. This depends on the specific Steiner system, but typically it means no more than 7. As $U$ grows, the approximation and simulations start to diverge in the high SNR regime, with the latter exhibiting plateauing. The reason is due to the existence of *stopping sets* [29]. It is easy to imagine a situation where the access patterns overlap in such a way that there are no collision-free slots and consequently SIC cannot be applied. Formally, a stopping set $s^{(n)}$ of order $n$ is a subset of $n$ patterns such that in every slot there is either 0, or $\geq 2$ users; that is, the decoding cannot proceed as there is no slot with a single transmission only. Furthermore, let us denote by $T^{(n)}$ the set of all stopping sets of order $n$ for a given Steiner system and by $|T^{(n)}|$ its cardinality[6]. In order to take into account stopping sets and augment the expression (D.14), we need to consider three cases. If there is a stopping set of certain order $n$, then with probability $n/U$ the user in focus is its member and cannot be decoded. Conversely, with probability $1 - n/U$ the user is not involved in that stopping set and decoding is possible, however SIC is impaired since there are $S = n$ noncancelable users. Otherwise, if there are no stopping sets then $S = 0$ and there are no limitations on SIC.

Ultimately, we have the following approximate expression for the outage probability with SIC:

$$\begin{aligned} p_{\text{out,SIC}}(R, U) = \sum_{n \in \mathfrak{N}} q_1(n|U) \left( \frac{n}{U} + \frac{U-n}{U} p_{\text{out,SIC}}(R, U|n) \right) \\ + p_{\text{out,SIC}}(R, U|0) \left( 1 - \sum_{n \in \mathfrak{N}} q_1(n|U) \right) \end{aligned} \tag{D.15}$$

where summation is over $\mathfrak{N} = \{n \colon T^{(n)} \neq \emptyset\}$ and $q_1(n|U) \approx f_{\text{bin}}(1; \binom{U}{n}, \frac{|T^{(n)}|}{\binom{C}{n}})$ is the probability that there is a stopping set of order $n$ among $U$ active users. We note that this is an approximation, because the $\binom{U}{n}$ tuples are not independent. Additionally, we would like to bring to the reader's attention the fact that we are interested in "exactly one" stopping set of a given order and not "at least one". The reason is that stopping sets are closed under union, so their combination produces another stopping set of a higher order [29]. As such, by summing over $n$ we would count some of the stopping sets multiple times.

Perhaps the most important, however, is that in (D.15) it is not necessary to sum over whole $\mathfrak{N}$ to obtain a good approximation. In practice, performance is impacted primarily by the stopping sets of the lowest existing

[6] While it would be more precise to write $T^{(n)}_{S(t,K,M)}$, for brevity we decide to drop the subscript $S(t,K,M)$ (as we similarly do in the case of quantities $C$ and $D$). In this paper we always consider a single Steiner system at a time so this should not lead to any confusion.

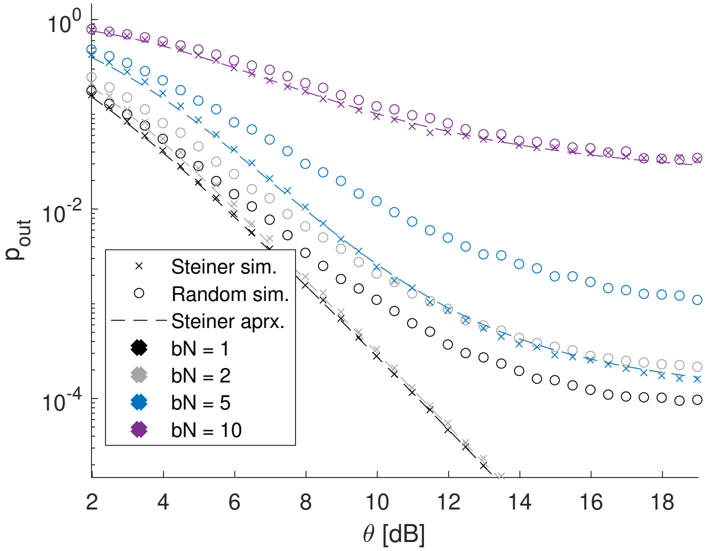


**Fig. D.4:** Comparison of the simulation results involving the exact procedure and the proposed approximation

order, which we denote by $n'$. They are decisive for two reasons. For a given Steiner system, the outage probability cannot be made arbitrarily low regardless of the SNR whenever $U \geq n'$. Secondly, simulations show that for $i < j$, $q_1(i|U) > q_1(j|U)$ and the difference can reach several orders of magnitude, making $q_1(n'|U)$ dominant overall. This is fortunate, since finding $T^{(n)}$ requires an exhaustive search, which for high $n$ becomes prohibitive. Consequently, when generating results, for each Steiner system we use only the stopping sets of the lowest existing order $n'$, and $n'+1$ whenever $T^{(n'+1)}$ is not empty (if $T^{(n'+1)} = \emptyset$, $T^{(n')}$ is sufficient).

In Fig. D.4 we plot the outage probability as given by the proposed approximation (dashed lines) and compare it to the results of the corresponding simulations in which the full procedure is implemented (markers). The derived approximations prove to be very close to the exact results across the whole SNR range and different traffic intensities. The improvement compared to a system without SIC (cf. Fig. D.3) is significant and allows to achieve ultra reliability at much lower SNRs. Particularly important is the fact that Random selection exhibits a performance floor even when the mean traffic intensity is as low as 1 user/frame. This is a consequence of the stopping sets, which in a Random selection can occur already with two users if they select exactly the same pattern. This has been discussed also in [10]. Conversely, in a Steiner system the number of collisions between two patterns is strictly limited, hence there are no stopping sets as long as the number of active users is sufficiently small. It is easy to show that with a maximum of $t-1$ collisions and $K$ repetitions at least

$$n' \geq \left\lceil \frac{K}{t-1} \right\rceil + 1 \tag{D.16}$$

users need to be active for the stopping set to occur. In practice, for some Steiner systems that number is even higher, e.g. in the used $S(2,4,25)$ no subset of size $< 7$ exist that would form a stopping set. Those issues as well as the results for other systems are further discussed in Section 6.

## 4.3 Model with Full Maximum Ratio Combining (MRC)

In the following we will consider a more involved model, in which the receiver is capable of using the totality of all the replicas (including those experiencing interference) and combines them using maximum ratio combining (MRC).

Without loss of generality, let us consider an arbitrary active user $i$ and one of its transmissions $j \in \{m : V_{m,i} = 1\}$. The SINR of this signal

$$\rho_{j,i} = \frac{P_x |g_{j,i}|^2}{\sum_{k\in\{1,\ldots,N\}\setminus\{i\}} V_{j,k} A_{k,k} P_x |g_{j,k}|^2 + \sigma^2} \tag{D.17}$$

is a random variable. Assuming there are $L$ interferers in slot $j$, i.e. $\sum_{k\in\{1,\ldots,N\}\setminus\{i\}} V_{j,k} A_{k,k} = L$, we can denote this SINR as $\frac{X}{Y+1}$, where $X$ follows the exponential distribution $f_{\text{exp}}(x;\theta)$ and $Y$ the gamma distribution $f_{\text{gam}}(y;L,\theta)$. Hence,

$$\begin{aligned} P\left(\frac{X}{Y+1} > z\right) &= \int_0^\infty \left(\int_{(y+1)z}^\infty \frac{1}{\theta} e^{-\frac{x}{\theta}} dx\right) \frac{1}{\Gamma(L)\theta^L} y^{L-1} e^{-\frac{y}{\theta}} dy \\ &= \frac{1}{\Gamma(L)\theta^L} \int_0^\infty y^{L-1} e^{-\frac{y}{\theta}} \cdot -e^{-\frac{x}{\theta}} \Big|_{(y+1)z}^\infty dy \\ &= \frac{1}{\Gamma(L)\theta^L} \int_0^\infty y^{L-1} e^{-\frac{y}{\theta}} e^{-\frac{(y+1)z}{\theta}} dy \\ &= \frac{e^{-\frac{z}{\theta}}}{\Gamma(L)\theta^L} \int_0^\infty y^{L-1} e^{-\frac{y(z+1)}{\theta}} dy \\ &= \frac{e^{-\frac{z}{\theta}}}{\Gamma(L)\theta^L} \cdot \frac{(L-1)!\theta^L}{(z+1)^L} = \frac{e^{-\frac{z}{\theta}}}{(z+1)^L} \end{aligned} \tag{D.18}$$

where the last integral can be computed by integrating it by parts $L-1$ times. Finally, by taking the derivative of $1 - P\left(\frac{X}{Y+1} > z\right)$ we obtain the pdf of the SINR given $L$ interferers:

$$f_{SINR}(z;\theta|L) = \frac{e^{-\frac{z}{\theta}}(\theta L + z + 1)}{\theta(z+1)^{L+1}} \tag{D.19}$$

and it can be seen that for a special case of $L = 0$ the expression reduces to a simple exponential distribution.

It would be tempting to consider the final SINR as a sum of $K$ independent RV's where SINR from a single slot is a mixture of $f_{SINR}(z;\theta|L)p_I(L|U)$ for $L = 0, \dots, U-1$. However, such a naïve approach does not provide a good approximation (especially when applied to the patterns from Steiner system) as it significantly underestimates the contribution of the interference-free slots, which contribute the most to the combined power[7]. Instead, let us first condition on the number of interference-free slots $K'$ (given by (D.4), (D.6)). Notice, however, that fixing $K'$ has an implications for the distribution of interferers in the remaining slots. Namely, by fixing $K'$ we implicitly reduce the number of available patterns (in case of Steiner system) or slots (in case of random selection). Hence, we introduce the modified version of the expressions (D.5), (D.7):

$$p_{\mathrm{I,R}}(L|K',U) = \frac{\binom{M-1-K'}{K}^{U-1-L}\binom{M-1-K'}{K-1}^{L}\binom{U-1}{L}}{\binom{M-K'}{K}^{U-1}} \tag{D.20}$$

$$p_{\mathrm{I,S}}(L|K',U) = \frac{\binom{D-1}{L}\binom{C-1-(D-1)(K'+1)}{U-1-L}}{\binom{C-1-(D-1)K'}{U-1}}\,. \tag{D.21}$$

Secondly, if the slot contains interference, then by definition $L > 0$ so the case $L = 0$ has to be excluded and the distribution re-normalized. Taking all this into account, the distribution of the SINR in the interfered slot becomes

$$f_{I-SINR}(z;\theta|U,K') = \sum_{L=1}^{U-1} f_{SINR}(z;\theta|L)\frac{p_{\mathrm{I}}(L|K',U)}{1-p_{\mathrm{I}}(0|K',U)}\,. \tag{D.22}$$

In order to obtain the distribution of the total SINR, we combine the above with the contribution from $K'$ interference-free slots and marginalize:

$$f_{tot,SINR}(z;\theta|U) = \sum_{K'=0}^{K} p_{\mathrm{CF}}(K'|U) f_{\mathrm{gam}}(z;K',\theta) * f_{I-SINR}(z;\theta|U,K')^{*(K-K')} \tag{D.23}$$

where $f^{*n} \overset{\mathrm{def}}{=} \underbrace{f * \cdots * f}_{n}$ and $f^{*0} = \delta$, with $\delta$ denoting the Dirac delta distribution. Similarly, we set $f_{\mathrm{gam}}(z;0,\theta) = \delta$ to keep the above expression (D.23)

[7] To see why, consider two users employing patterns from the Steiner system. For a given slot, there is a certain probability of collision, which means that by treating them independently we include cases with $1, \dots, t-1, \dots, K$ collisions. Meanwhile, Steiner system guarantees that two users will share at most $t-1$ slots.

compact. Finally, we have that

$$\begin{aligned} p_{out,cap}(R,\theta,U) = \sum_{K'=0}^{K} & p_{\mathrm{CF}}(K'|U) \\ & \times F_{\mathrm{gam}}(z;K',\theta) * f_{I-SINR}(z;\theta|U,K')^{*(K-K')}\Big|^{z=2^R-1} \end{aligned} \tag{D.24}$$

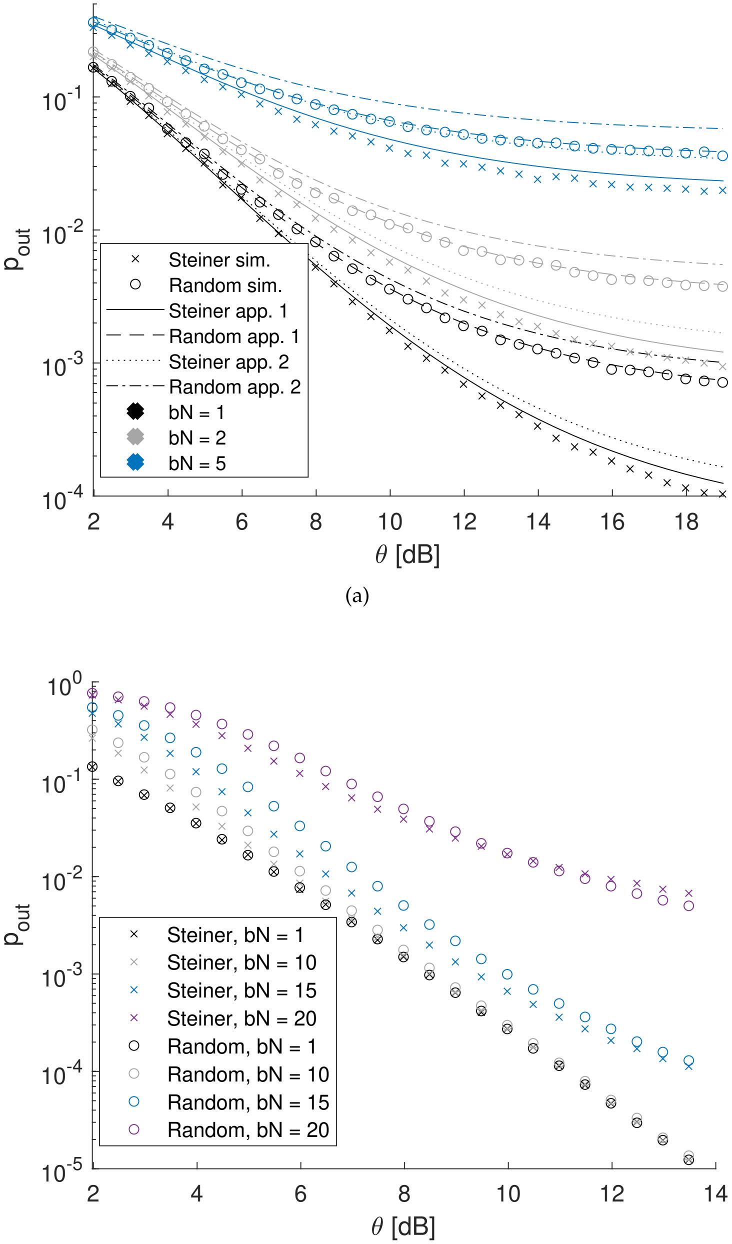
While not excessively complex, evaluation of (D.24) requires $(K-1)(K+2)/2$ numerical integrations. This is not an issue for the values of $K$ considered in this work, however to address this we provide in Appendix A an even simpler approximation that avoids the convolutions altogether.

In Fig. D.5(a) we present joint comparison of the simulation results and developed approximations. Once again, we consider the performance of the Steiner system and Random selection across the range of received SNRs and traffic intensities. The results from simulations, which implement the exact procedure, are given with markers. Solid and dashed lines (Steiner and Random respectively) correspond to the approximation based on eq. (D.24) (Approx. 1). Similarly, dotted and dash-dotted lines (Approx. 2) correspond to the simpler approximation by gamma distribution discussed in the Appendix A. Approx. 1 follows very closely the actual simulation results, however we note that in case of Steiner system the deviation is slightly larger than for Random selection. Albeit being simpler, the accuracy of Approx. 2 is not far off, especially for lower traffic intensities $bN$, which are of primary interest when ultra-reliability is considered.

For completeness, in Fig. D.5(b) we show the simulations results for the Full MRC model that leverages SIC. The improvement over non-SIC (cf. Fig. D.5(a)) processing is significant as the performance does not exhibit plateauing and is capable of achieving ultra-reliability even for traffic intensities as high as $bN = 10$. The superiority of Steiner system over Random selection, especially for high SNRs, is diminished, however this aspect is further discussed in Section 5.

# 5 System Design favors Steiner Sequences over Random Selection

Until now, the implicit assumption was that perfect channel estimates have been available. In reality channel estimates are never perfect, however, the quality of estimation can be improved by making the pilot sequences longer and/or investing in them more transmit power, as long as the pilot sequences of transmitting users are kept orthogonal. Collisions in the pilot domain are particularly problematic as they lead to pilot contamination and, consequently, very poor channel estimates. Typically, that means proper equal-



(a)

(b)

**Fig. D.5:** Outage probability in the Full MRC model (a) without SIC and (b) with SIC.

ization is not possible and the packet replicas involved in the collision are unusable i.e. cannot be combined with others through MRC or removed with SIC[8]. This is a significant challenge for random access schemes that rely on fully random selection of access patterns. Since any user can be active in any slot, the only way to avoid pilot collisions, would be to assign a unique orthogonal sequence to each of the $N$ devices. In many cases, however, this is not feasible or practical (e.g. with large population of intermittently active devices similar to the scenario addressed in this work). Instead, a common approach is to provide a pool of $Q < N$ pilot sequences from which users pick one at random every time they become active and accept, that some collisions will inevitably occur.

In that case, it is possible to provide a reasonable lower bound for the model utilizing SIC as $\theta \to \infty$. Let us again focus on an arbitrary user $u$ and one of its slots. We can distinguish two types of events. In the first case, whenever one or more interferers select the same pilot sequence as the user of interest $u$, the packet replica is lost. This is given by

$$p_{(\mathrm{I})} = \sum_{L=1}^{U-1} p_{\mathrm{I}}(L|U) \left(1 - \left(1 - \frac{1}{Q}\right)^{L}\right) . \tag{D.25}$$

The second type of event, is when user $u$'s pilots are intact, however there are some pilot collisions among the interferers. The probability that the slot is of this second kind is

$$p_{(\mathrm{II})} = \sum_{L=2}^{U-1} p_{\mathrm{I}}(L|U) \left(1 - \frac{\prod_{i=0}^{L-1} Q - i}{Q^{L}} - \left(1 - \left(1 - \frac{1}{Q}\right)^{L}\right)\right) . \tag{D.26}$$

In that case, even if the packets of interfering users can be decoded based on their other replicas, due to the lack of channel knowledge SIC cannot be applied, so the SINR is limited to at most $\frac{X}{Y_1 + \cdots + Y_{L'} + 1}$, where $L'$ is the number of mutually colliding (in the pilot domain) interferers. For the lower bound, we can fix $L' = 2$ and as $\theta \to \infty$ the '1' in the denominator can be dropped. Along with the fact that $X, Y_1, Y_2$ are exponentially distributed with the same scale parameter $\theta$, the SINR of user $u$'s replica follows a beta prime distribution $f_{BP}(x; \alpha, \beta) = \frac{x^{\alpha-1}(1+x)^{-\alpha-\beta}}{B(\alpha,\beta)}$ with $\alpha = 1$ and $\beta = 2$. Considering there can be $n = 0, \ldots, K$ such replicas (with the remaining ones being of the first type), the outage probability can be bounded by

$$p_{bound}(R, U) = p_{(\mathrm{I})}^{K} + \sum_{n=1}^{K} \binom{K}{n} p_{(\mathrm{I})}^{K-n} p_{(\mathrm{II})}^{n} F_{BP}(2^{R} - 1; n, 2) \tag{D.27}$$

[8] Note that even if the packet can be decoded based on replicas from other slots, it cannot be removed from the slots where pilot collisions occurred as the channel estimates are not available.

where $F_{BP}(\cdot;\cdot,\cdot)$ is the CDF of beta prime distribution and we leverage the fact that sum of its $n$ i.i.d variables is also beta-prime distributed, in this case, with $\alpha_n = n$, $\beta_n = 2$.

Unlike Random selection, the Steiner system guarantees that at most $D$ users can be active in any given slot. Furthermore, since each user has to be assigned a specific access pattern for their packet replicas, it can be simultaneously instructed which pilot sequence to use in which slot, thus eliminating any possibility of collisions with just $Q = D$ orthogonal pilot sequences. Recalling that in a Steiner system the number supportable users is $N = C = \binom{M}{t}/\binom{K}{t}$, while $D = \binom{M-1}{t-1}/\binom{K-1}{t-1}$, we have that $D = N\frac{K}{M}$, i.e. the number of pilot sequences required to ensure no collisions is reduced by a factor $K/M$ compared to the Random selection.

Another caveat is that, clearly, the receiver must know where each replica of each user is located in order to perform combining through MRC. Because with the Steiner system there is an association between pilot sequences and user IDs, observing a certain pilot sequence in a given slot automatically indicates which user is active and where to look for its remaining replicas. This is not the case when Random selection scheme is used so additional procedures might be required. One possibility is to look for the correlation between signals in different slots and combine those with the highest correlation score. However, such a solution is not perfect as it might miss some of the replicas or introduce false positives. Furthermore, it entails exhaustive search and, hence, high complexity. Alternatively, a unique ID that can be decoded independently of the rest of the payload could be added to each packet or, each slot could be preceded by an activity-indication phase. Clearly, the downside of this solution is the introduction of overhead.

Lastly, we remark that MRC processing is impeded when using Random selection. Note, that the SINR of the combined packet is equal to the sum of the SINRs of its constituents only when the interference in each is uncorrelated. This will not be the case if a given user collides with another in more than one slot. Whether the resulting SINR will be higher or lower than the sum will depend on whether the interference adds destructively or constructively. Interestingly, even though both are equally likely, the performance of the SIC is ultimately impaired[9]. To circumvent that, a more computationally-

[9] Consider the case with two users colliding in more than one slot and let us establish a baseline where the combined SINR is equal to the sum of SINRs in individual slots. As already mentioned, in reality, the SINR after MRC will be lower (we can call it negative MRC) or higher (positive MRC) than that baseline. Importantly, if the combining is negative(positive), it is negative(positive) for both users. Consider now 4 possible decoding outcomes than can happen in the baseline scenario. If both users are successful, then positive MRC would have no effect. Similarly, if only one of them succeeds but the other fails, positive MRC also wouldn't change anything since the successful user can be cancelled through SIC anyway. Only when both users fail the positive MRC can make a difference - that is if it makes at least one of them decodable and triggers SIC. Now let us consider negative MRC. When both users fail it has no effect. If

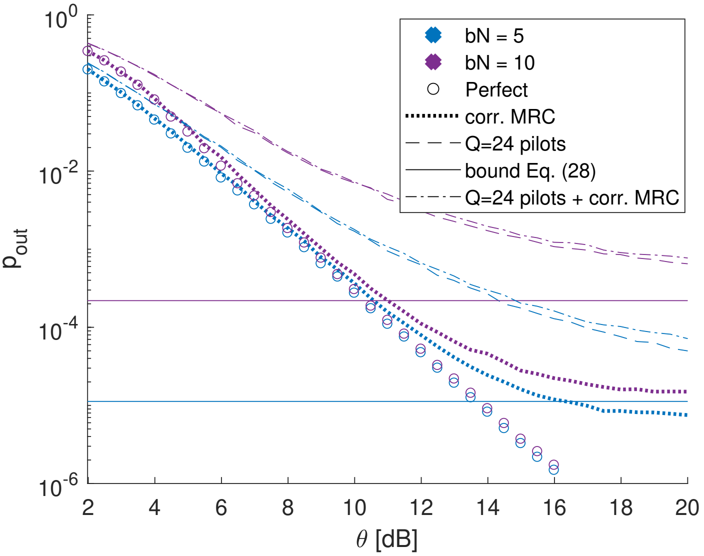


**Fig. D.6:** Impact of pilot collisions and realistic MRC processing on Random selection scheme. Although not shown here for the sake of readability, the performance of Steiner system matches, and for low $\theta$ even exceeds, the perfect Random scheme.

heavy equalization method such as zero-forcing (ZF) or minimum mean squared error (MMSE) would have to be used. Conversely, in a Steiner system with $t = 2$ each user is guaranteed to collide with another at most once so the interference is always uncorrelated.

In Fig. D.6 we demonstrate the impact of the above described issues. The markers represent the performance of the idealized version of the Random selection scheme and serve as a reference. The dotted lines show the actual performance of MRC in the presence of correlated interference. The dashed lines depict the scenario with finite pool of pilot sequences $Q = 24$ and the solid horizontal line is the corresponding lower bound as given by (D.27). Lastly, dash-dotted curves take into account both detrimental factors. The Fig. D.6 reveals that, in a more realistic setting, the performance of the Random selection would be significantly impaired and cannot match that of Steiner system (cf. Fig. D.5(b)).

there is one successful user, it can happen that negative MRC turns it into an undecodable one, thus making SIC impossible. Similarly, if both users are successful, negative MRC could make them both undecodable (in this case it is not enough to turn just one of them, as the SIC could still be applied). In the end, even though negative and positive MRC are equally probable, if the system uses SIC, the negative MRC has the potential to be dentrimental in three out of four cases, while the latter can only help in one of them.

**Table D.1:** Properties of Steiner Systems

| | $C$ | $D$ | $T^{(n)}$ |
|---|---|---|---|
| $S(2,5,25)$ | 30 | 6 | $\lvert T^{(9)}\rvert = 1150$ |
| $S(2,5,41)$ | 82 | 10 | $\lvert T^{(6)}\rvert = 41$, $\lvert T^{(7)}\rvert = 0$ |
| $S(2,4,25)$ | 50 | 8 | $\lvert T^{(7)}\rvert = 266$, $\lvert T^{(8)}\rvert = 1827$ |
| $S(2,4,28)$ | 63 | 9 | $\lvert T^{(5)}\rvert = 126$, $\lvert T^{(6)}\rvert = 630$ |
| $S(2,4,37)$ | 111 | 12 | $\lvert T^{(6)}\rvert = 37$, $\lvert T^{(7)}\rvert = 0$ |
| $S(2,3,25)$ | 100 | 12 | $\lvert T^{(4)}\rvert = 4$, $\lvert T^{(5)}\rvert = 92$ |
| $S(2,3,33)$ | 176 | 16 | $\lvert T^{(4)}\rvert = 429$, $\lvert T^{(5)}\rvert = 77$ |
| $S(2,3,39)$ | 247 | 19 | $\lvert T^{(4)}\rvert = 60$, $\lvert T^{(5)}\rvert = 132$ |

# 6 Performance Evaluation: Choice of Frame Parameters $M$ and $K$

Lastly, we consider Steiner systems with different configurations of the frame length $M$ and number of repetitions $K$. In Table D.1 we provide the relevant parameters for the systems used in this work, namely the number of patterns $C$, maximum number of interferers per slot $D$, order of the smallest existing stopping sets as well as their number. The patterns themselves can be found in [30]. The objective is to determine the highest supported rate $R$ for a given traffic intensity $bN$ and fixed mean SNR $\theta$, that fulfills certain target reliability $\epsilon_{tar}$:

$$\begin{aligned} \max \quad & R \\ \text{s.t.} \quad & \sum_{u=0}^{N} f_{\text{bin}}(u,b,N)p_{\text{out}}(R,\theta,u) \leq \epsilon_{tar} \end{aligned} \tag{D.28}$$

Complementary to rate which is per UE, we define also the spectral efficiency of the system given by $bN \cdot R/M$. In the figures, the results are plotted as a function of the absolute mean traffic intensity $bN$. We note that since $N$ is different for each Steiner System, so is the activation probability $b$. Furthermore, in order to jointly compare Steiner Systems with different number of repetitions $K$, we decide to normalize their mean received SNRs. The rationale is that, with $\theta$ being the same in each case, the systems with higher $K$ would use proportionally more energy and thus have an advantage. Consequently, we perform our evaluations by fixing $\theta K = 25$ dB. Lastly, as we are considering ultra-reliability we set $\epsilon_{tar} = 10^{-5}$.

We focus on two cases a) the Full MRC model without SIC and b) the collision model with SIC. The results are obtained based on the derived approximations (D.24) and (D.15) respectively which are applied to (D.28). To improve the readability of the figures, we do not show the results for Random selection schemes in this section, noting that they are always strictly worse than the corresponding Steiner system (cf. earlier discussions and Fig. D.3-D.5).

We start with the rate $R$ of the Full MRC model shown in Fig. D.7(a). As expected, for a given mean number of active users, the larger the frame and the number of repetitions, the higher the rate. Increasing the frame size decreases the chance of collisions, while increasing $K$ makes the transmission more robust and allows to harvest more diversity. The relationship, however, is not as straightforward when it comes to spectral efficiency shown in Fig.D.7(b). For a given number of repetitions $K$, increasing the frame length actually reduces the spectral efficiency, as the increase in rate is not enough to offset the extra resources (cf. $S(2,5,25)$ vs $S(2,5,41)$ or $S(2,4,25)$ vs $S(2,4,28)$ vs $S(2,4,37)$). Furthermore, even though higher $K$ itself is generally beneficial, the system with high $M$ and $K$ might be less spectrally efficient than the one with lower parameters when $bN$ is low (cf. $S(2,5,41)$ vs $S(2,4,25)$ or $S(2,4,28)$).

To provide further insight, in Fig. D.7(b) we also mark the traffic intensity $bN$ beyond which orthogonal resource allocation becomes more spectrally efficient than Steiner system. To find this value, first, we note that when resources are orthogonal, the maximum rate does not depend on the traffic intensity and is given by $R_{orth} = \log_2(F_{gam}^{-1}(\epsilon_{tar}; K, \theta) + 1)$, where $F^{-1}$ is the inverse CDF (quantile function). Since in a Steiner system the maximum number of users is $N = C$, the equivalent orthogonal allocation requires a frame of length $M_{orth} = CK$. For the sake of readability, we plot the spectral efficiency curve of the orthogonal system only for one representative case for each $K = 3, 4, 5$ (dashed curve) and for the rest simply mark the point at which $bN \cdot R/M$ intersects with $bN \cdot R_{orth}/M_{orth}$.

In Fig. D.8 the results corresponding to the collision model with SIC are shown. There are several notable differences. First, the maximum rate does not exhibit such a smooth degradation as in the Full MRC case. Instead, it stays very high and close to its absolute maximum (i.e. $\log_2(F_{gam}^{-1}(\epsilon_{tar}; K, \theta) + 1)$ as in the orthogonal allocation) and then goes abruptly to 0 once it reaches certain cut-off traffic intensity. The maximum supportable $bN$ of a Steiner system is a non-trivial function of the order and number of its stopping sets, frame length $M$, and number of patterns $C$. While the general trend is preserved, i.e. higher $K$ and $M$ lead to higher rates, there are some exceptions. Comparing $S(2,3,25)$, $S(2,3,33)$ and $S(2,3,39)$, one can see that the case with $M = 33$ actually performs the worst. This is tied to the particularly high number of stopping sets (cf. Table D.1). On the other hand, $S(2,4,37)$ can sustain higher traffic intensity than $S(2,5,41)$ despite shorter frame length. In this case, the reason lies in the lower $K$ and, consequently, higher number of patterns (111 vs 82). Even though the order and number of stopping sets

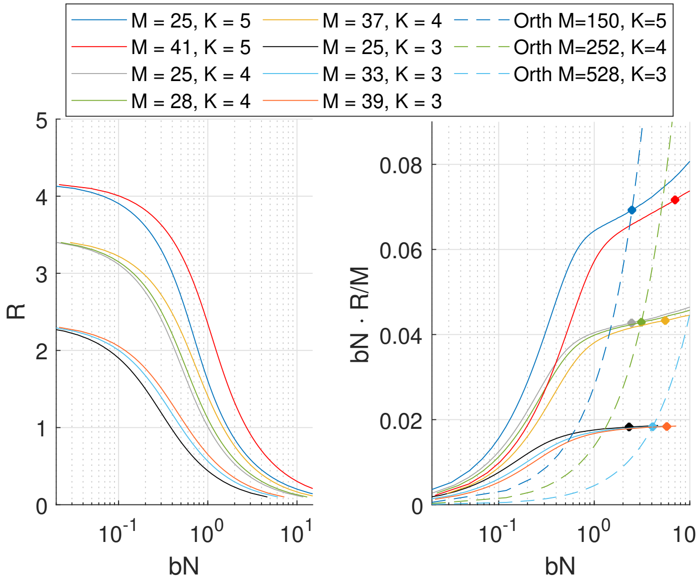


**Fig. D.7:** Full MRC model without SIC. (a) Maximum rate for which outage probability target $p_{out} \leq 10^{-5}$ is fulfilled as a function of the traffic intensity, and (b) corresponding spectral efficiency.

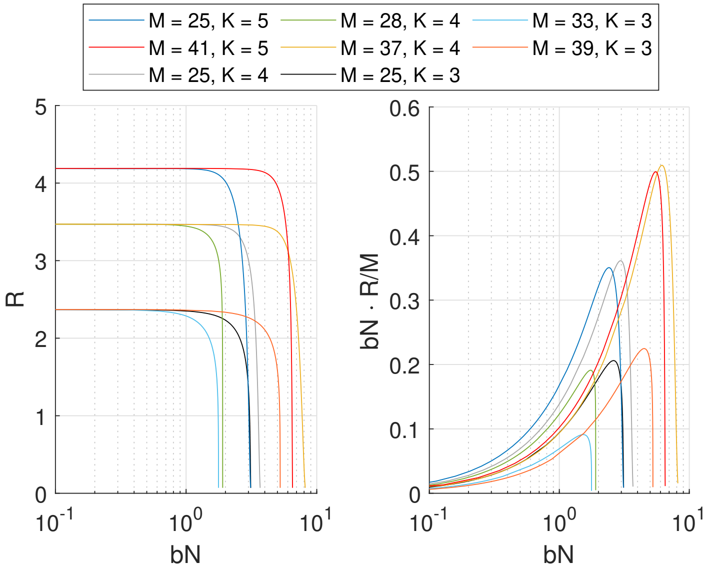


**Fig. D.8:** Collision model with SIC. (a) Maximum rate for which outage probability target $p_{out} \leq 10^{-5}$ is fulfilled as a function of the traffic intensity, and (b) corresponding spectral efficiency.

is similar, the probability of their occurrence is effectively lower in $S(2,4,37)$.

# 7 Conclusions

In this work we have proposed and investigated the usage of deterministic access patterns to provide ultra-reliable communication for a group of intermittently active users sharing a pool of resources. The patterns, which are a realization of the Steiner system, aim to control the number of collisions and interference among users. This feature leads to significant gains in terms of outage probability compared to an approach were the choice of channel resources is fully random. In our evaluations we have considered two different signal models - based on destructive collisions and Full MRC, and two receiver processing techniques - with and without SIC. As the second main contribution of this work, we have developed simple approximations for the outage probability in a collision model with SIC and Full MRC model without SIC that closely match the simulation results. Such approximations are particularly important in the context of ultra-reliable systems where the number of required samples/simulations needed to properly assess the performance is often infeasible.

# A Appendix

A relatively good approximation for the expression (D.23) can be obtained by first approximating (D.22) with a gamma distribution, i.e. finding $f_{\text{gam}}(z; k_{K'}, \alpha_{K'}) \approx f_{I-SINR}(z; \theta | U, K')$. This can be done e.g. by solving the following optimization problem to find the suitable coefficients:

$$\begin{aligned} \underset{k_{K'}, \alpha_{K'}}{\arg\min} \quad & \int_A \left[ f_{I-SINR}(z;\theta|U,K') - f_{\text{gam}}(z;k_{K'},\alpha_{K'}) \right]^2 dz \\ \text{s.t.} \quad & k_{K'} > 0, \quad \alpha_{K'} > 0 \end{aligned} \tag{D.29}$$

where $A = [0, 2^R - 1]$, since we are interested in the outage probability and hence concerned with a good fit only in that region. In Fig. D.9 we show as an example the results of such fitting for the Random selection in the SINR range $[0, 2^2 - 1]$. Goodness of fit tends to be the lowest for a medium number of interferers which is also reflected in the final approximation in Fig. D.5(a), as it becomes less tight the higher the traffic intensity $bN$.

Since the SINRs of individual interfered slots are i.i.d and for a given $K'$ each is now represented by a gamma distribution with parameters $k_{K'}$, $\alpha_{K'}$, we have simply that

$$f_{\text{gam}}(z; k_{K'}, \alpha_{K'})^{*(K-K')} = f_{\text{gam}}(z; (K-K')k_{K'}, \alpha_{K'}) \tag{D.30}$$

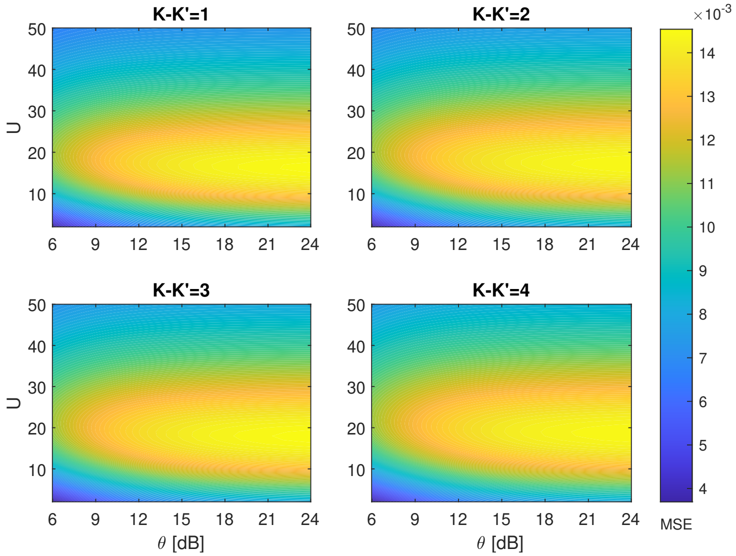


**Fig. D.9:** Results of least-squares approximation of (D.22) with gamma distribution.

Lastly, to combine (D.30) with the contribution from the interference-free slots $f_{\text{gam}}(z; K', \theta)$ we can use the result in [31] which provides the analytical expression for the sum of two gamma distributed RVs with arbitrary parameters:

$$\begin{aligned}
&\tilde{f}_{tot,SINR}(z;\theta|U) \\
&\quad = \sum_{K'=0}^{K} p(K'|U) f_{\text{gam}}(z; K', \theta) * f_{\text{gam}}(z; (K-K')k_{K'}, \alpha_{K'}) \\
&\quad = \sum_{K'=0}^{K} p(K'|U) \left(\frac{\alpha_{K'}}{\theta}\right)^{K'} \frac{\left(\frac{1}{\alpha_{K'}}\right)^{\kappa} z^{\kappa-1} e^{-\frac{z}{\alpha_{K'}}}}{\Gamma(\kappa)} {}_1F_1\left(K'; \kappa; \left(\frac{z}{\alpha_{K'}} - \frac{z}{\theta}\right)\right)
\end{aligned} \tag{D.31}$$

where $\kappa = K' + (K-K')k_{K'}$ and ${}_1F_1(\cdot;\cdot;\cdot)$ is a Kummer's confluent hypergeometric function. The CDF in this case becomes

$$\begin{aligned}
\tilde{F}_{tot,SINR}(z;\theta|U) &= \sum_{K'=0}^{K} p(K'|U) F_{\text{gam}}(z; K', \theta) * f_{\text{gam}}(z; (K-K')k_{K'}, \alpha_{K'}) \\
&= \sum_{K'=0}^{K} p(K'|U) \left(\frac{\alpha_{K'}}{\theta}\right)^{K'} {}_{[\frac{z}{\alpha_{K'}}]2}F_1\left(\kappa, K'; \kappa; \left(1 - \frac{\alpha_{K'}}{\theta}\right)\right)
\end{aligned} \tag{D.32}$$

where ${}_{[\cdot]2}F_1(\cdot,\cdot;\cdot;\cdot)$ is the incomplete Gauss hypergeometric function.

# B Appendix

## Proof of Lemma 1

*Proof.* Let us define a probability space where the *samples* are different ways the $U-1$ other users can select the locations of their packets inside the frame. We will call these samples configurations. An *event* is then a set of configurations and has a probability.

Let us now fix $K'$ of the $K$ slots of a given user $u$. Denote with $A_i$ an event that consists of all the configurations where the fixed set $K'$, and at least one of the remaining $K-K'$ slots, denoted by $i$, are free of interference. This gives us $K-K'$ different sets $A_i$.

Given now an $n$-element subset $J$ of $\{1,\ldots,K-K'\}$, then the probability for an event $\bigcap_{j\in J} A_j$ to appear is $P(\bigcap_{j\in J} A_j) = \left(\binom{M-K'-n}{K}/\binom{M}{K}\right)^{U-1} = V_n$, independent of the selected $J$ and $K'$. This can be directly seen as $\bigcap_{j\in J} A_j$ consists of all the configurations that leave at least fixed $n+K'$ slots of user $u$ interference free.

Let us denote with $S$ the set consisting of all the configurations where at least the fixed $K'$ slots are interference free. By the complementary form of the inclusion-exclusion principle we then have that

$$P\left(S \setminus \bigcup_{i=1}^{K-K'} A_i\right) = \sum_{n=0}^{K-K'} (-1)^n \binom{K-K'}{n} V_n, \tag{D.33}$$

where $V_0 = P(S) = \left(\binom{M-K'}{K}/\binom{M}{K}\right)^{U-1}$. Here the set $\bigcup_{i=1}^{K-K'} A_i$ is an event that contains all the configurations that leave the fixed $K'$ slots and at least one other of the slots occupied by the user $u$ free of interference. Then the set $S \setminus \bigcup_{i=1}^{K-K'} A_i$ is an event that consists of all the configurations, where the fixed $K'$ slots are free, but none of the other $K-K'$ occupied by the user $u$. In other words, it is the set of those configurations, where exactly $K'$ of $K$ slots are free. One can place $K'$ packets to $K$ available slots in $\binom{K}{K'}$ different ways. Hence the final result is gotten by multiplying (D.33) with the term $\binom{K}{K'}$. □

# C Appendix

## Proof of Lemma 2

*Proof.* The probability that a specific set of $K'$ slots selected by the user $u$ is not occupied by the packets of remaining $U-1$ users is

$$\left(\frac{\binom{C-1-K'(D-1)}{U-1}}{\binom{C-1}{U-1}}\right),$$

because there are $C$ patterns in total (including pattern of user $u$) and $K'(D-1)$ of them share a slot with the $K'$ slot set of pattern of user $u$. The proof follows that of Lemma 1 verbatim, after realizing that now

$$P(\bigcap_{j\in J} A_j) = \left(\frac{\binom{C-1-(D-1)(n+K')}{U-1}}{\binom{C-1}{U-1}}\right),$$

for any $n$ elements set of $J$. □

# References

[1] "Ericsson Mobility report," Tech. Rep., Nov. 2021. [Online]. Available: https://www.ericsson.com/en/reports-and-papers/mobility-report/reports/november-2021

[2] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, 2020.

[3] 3GPP, "Service requirements for the 5G system; Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.261, Dec. 2021, v18.5.0.

[4] "Verticals URLLC Use Cases and Requirements," NGMN Alliance, Tech. Rep., July. 2019. [Online]. Available: https://www.ngmn.org/publications/verticals-urllc-use-cases-and-requirements.html

[5] S. Schiessl, "Performance Trade-offs for Ultra-Reliable Low-Latency Communication Systems," Ph.D. dissertation, KTH Royal Institute of Technology, 2019. [Online]. Available: http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-251650

[6] G. Berardinelli, N. Huda Mahmood, R. Abreu, T. Jacobsen, K. Pedersen, I. Z. Kovács, and P. Mogensen, "Reliability Analysis of Uplink Grant-Free Transmission Over Shared Resources," *IEEE Access*, vol. 6, pp. 23 602–23 611, 2018.

[7] N. Patriciello, S. Lagen, L. Giupponi, and B. Bojovic, "The Impact of NR Scheduling Timings on End-to-End Delay for Uplink Traffic," in *2019 IEEE Global Communications Conference (GLOBECOM)*, 2019, pp. 1–6.

[8] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-Free Non-Orthogonal Multiple Access for IoT: A Survey," *IEEE Communications Surveys Tutorials*, vol. 22, no. 3, pp. 1805–1838, 2020.

[9] G. Choudhury and S. Rappaport, "Diversity ALOHA - A Random Access Scheme for Satellite Communications," *IEEE Transactions on Communications*, vol. 31, no. 3, pp. 450–457, 1983.

[10] C. Boyd, R. Vehkalahti, and O. Tirkkonen, "Interference Cancelling Codes for Ultra-Reliable Random Access," *International Journal of Wireless Information Networks*, vol. 25, pp. 1–12, 12 2018.

[11] E. Casini, R. De Gaudenzi, and O. Del Rio Herrero, "Contention Resolution Diversity Slotted ALOHA (CRDSA): An Enhanced Random Access Schemefor Satellite Access Packet Networks," *IEEE Transactions on Wireless Communications*, vol. 6, no. 4, pp. 1408–1419, 2007.

[12] L. G. Roberts, "ALOHA Packet System with and without Slots and Capture," vol. 5, no. 2, 1975. [Online]. Available: https://doi.org/10.1145/1024916.1024920

[13] G. Liva, "Graph-Based Analysis and Optimization of Contention Resolution Diversity Slotted ALOHA," *IEEE Transactions on Communications*, vol. 59, no. 2, pp. 477–487, 2011.

[14] M. Ghanbarinejad and C. Schlegel, "Irregular repetition slotted ALOHA with multiuser detection," in *2013 10th Annual Conference on Wireless On-demand Network Systems and Services (WONS)*, 2013, pp. 201–205.

[15] F. Clazzer, E. Paolini, I. Mambelli, and C. Stefanović, "Irregular repetition slotted ALOHA over the Rayleigh block fading channel with capture," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[16] E. Paolini, G. Liva, and M. Chiani, "Coded Slotted ALOHA: A Graph-Based Method for Uncoordinated Multiple Access," *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6815–6832, 2015.

[17] J. Choi, "Throughput Analysis for Coded Multichannel ALOHA Random Access," *IEEE Communications Letters*, vol. 21, no. 8, pp. 1803–1806, 2017.

[18] R. Abreu, "Uplink Grant-free Access for Ultra-Reliable Low-Latency Communications in 5G: Radio Access and Resource Management Solutions," Ph.D. dissertation, 2019, phD supervisor: Prof. Preben Mogensen, Aalborg University Assistant PhD supervisors: Assoc. Prof. Gilberto Berardinelli, Aalborg University Prof. Klaus Pedersen, Aalborg University.

[19] M. C. Lucas-Estañ, J. Gozálvez, and M. Sepulcre, "On the Capacity of 5G NR Grant-Free Scheduling with Shared Radio Resources to Support Ultra-Reliable and Low-Latency Communications," *Sensors (Basel, Switzerland)*, vol. 19, 2019.

[20] Y. Liu, Y. Deng, M. Elkashlan, A. Nallanathan, and G. K. Karagiannidis, "Analyzing Grant-Free Access for URLLC Service," *IEEE Journal on Selected Areas in Communications*, vol. 39, no. 3, pp. 741–755, 2021.

[21] J. Choi and J. Ding, "Network Coding for K-Repetition in Grant-Free Random Access," *IEEE Wireless Communications Letters*, vol. 10, no. 11, pp. 2557–2561, 2021.

[22] M. Lucas-Estan and J. Gozalvez, "Sensing-based Grant-Free Scheduling for Ultra Reliable Low Latency and Deterministic Beyond 5G Networks," *IEEE Transactions on Vehicular Technology*, pp. 1–1, 2022.

[23] G. T. Peeters, R. Bocklandt, and B. Van Houdt, "Multiple Access Algorithms Without Feedback Using Combinatorial Designs," *IEEE Transactions on Communications*, vol. 57, no. 9, pp. 2724–2733, 2009.

[24] C. Boyd, R. Vehkalahti, and O. Tirkkonen, "Combinatorial code designs for ultra-reliable IoT random access," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, 2017.

[25] C. Boyd, R. Kotaba, O. Tirkkonen, and P. Popovski, "Non-Orthogonal Contention-Based Access for URLLC Devices with Frequency Diversity," in *2019 IEEE 20th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2019, pp. 1–5.

[26] E. Paolini, G. Liva, and A. Graell i Amat, "A structured irregular repetition slotted ALOHA scheme with low error floors," in *2017 IEEE International Conference on Communications (ICC)*, 2017, pp. 1–6.

[27] R. Kotaba, C. N. Manchón, and P. Popovski, "Enhancing Performance of Uplink URLLC Systems via Shared Diversity Transmissions and Multiple Antenna Processing," in *2019 53rd Asilomar Conference on Signals, Systems, and Computers*, 2019, pp. 1409–1415.

[28] W. Tang, S. Kang, B. Ren, and X. Yue, "Uplink grant-free pattern division multiple access (GF-PDMA) for 5G radio access," *China Communications*, vol. 15, no. 4, pp. 153–163, 2018.

[29] C. Di, D. Proietti, I. Telatar, T. Richardson, and R. Urbanke, "Finite-length analysis of low-density parity-check codes on the binary erasure channel," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1570–1579, 2002.

[30] La Jolla Covering Repository. Steiner Systems. [Online]. Available: https://ljcr.dmgordon.org/cover.php.

[31] F. D. Salvo, "A characterization of the distribution of a weighted sum of gamma variables through multiple hypergeometric functions," *Integral Transforms and Special Functions*, vol. 19, no. 8, pp. 563–575, 2008.

# Paper E

## Improving spectral efficiency in URLLC via NOMA-based retransmissions

Radosław Kotaba, Carles Navarro Manchón, Nuno Manuel Kiilerich Pratas, Tommaso Balercia and Petar Popovski

The paper has been published in the
*2019 IEEE International Conference on Communications (ICC)*, pp. 1–7, 2019.

*The layout has been revised.*

# Abstract

*The requirement to accommodate ultra-reliable low latency communication (URLLC) is one of the most attractive, yet challenging, new features of upcoming 5G systems. A common way to achieve reliability is retransmission; however, the applicability of this mechanism is hindered by the strict latency requirements. Furthermore, the bandwidth is often limited and shared by multiple connections, which may put the packet into a retransmission queue, leading to even larger latency. We address this problem in an uplink setting by introducing the concept of non-orthogonal multiple access hybrid automatic repeat request (NOMA-HARQ). In essence, NOMA-HARQ allows newly incoming packets to share non-orthogonally the same resource with retransmitted packets. The reliability guarantees are preserved by designing a power optimization procedure that takes into account past transmission attempts as well as the time remaining until the deadline.*

# 1 Introduction

Next generation wireless networks (5G) aim at enabling new and previously unattainable use cases. Among them, probably the most challenging ones are those belonging to the class of ultra-reliable low-latency communications (URLLC). Examples of such applications are tactile interaction and industrial automation which require end-to-end latency of 1ms and reliability (defined as the probability of successful delivery of the packet before the deadline) in the order of 99.999% [1]. While the concept of URLLC started to emerge already a few years ago [2] there is still no agreed, default technology that will support it. Clearly, simple reuse and fine-tuning of legacy techniques is insufficient as was repeatedly proven [3], [4], and the reason for that is they were designed for systems with significantly different performance requirements. This sparked a lot of new ideas and discussions regarding the potential enablers for URLLC [5], [6].

The crucial aspect whenever stringent reliability guarantees are required is to provide a sufficient level of diversity. One of the most natural and, at the same time, efficient ways of achieving diversity in wireless communications is through retransmission mechanisms such as hybrid automatic repeat request (HARQ). Its usefulness has been realized early on and was introduced already in third generation system called high speed packet access (HSPA) [7]. While this technique inevitably introduces latency we note that the alternative, i.e. achieving high reliability with so called one-shot transmission [8], is very inefficient in terms of power and, whenever feasible, some form of HARQ is highly desirable. We foresee that with shortened transmission time intervals (TTIs), higher subcarrier spacing and improved processing times, a small number of retransmissions will still be possible in all except the most latency-stringent use cases.

On the one hand, HARQ greatly improves reliability as it allows to combine the faulty packets instead of discarding them and get closer to the success with each retransmission. On the other hand, whenever the message is close to being decoded, another full retransmission is likely a waste of resources. The straightforward idea to send a shorter packet containing just enough redundancy bits, while intuitive, is not easy to accomplish in practice and would require additional signaling to enable dynamically changing, arbitrary packet sizes. Instead, the legacy systems [9] prefer to work with fixed-size resources.

This resource inefficiency of HARQ can be alleviated with the help of non-orthogonal multiple access (NOMA) which superimposes, in a controlled manner, multiple transmissions over the same physical resources. Due to that, NOMA is also one of the most promising techniques in terms of ability to accommodate more users and latency reduction. The topic has been gaining increasing attention recently and a comprehensive overview can be found in [10] where different variants of power and code domain NOMA have been discussed. While its downlink version is much more popular within the research community due to the higher potential gains, in recent years increasing number of contributions claiming successful application of NOMA in the uplink have been reported [11], [12].

In this contribution we address the limitation of the legacy HARQ mechanism by proposing a novel technique based on the NOMA principle which we apply to the uplink scenario. The solution deliberately pairs two classes of messages: retransmissions and new packets, and schedules them over the same channel/physical resources, which improves spectral efficiency. At the core of NOMA-HARQ is a power optimization technique consisting of two subparts, termed online and offline optimization respectively. The goal of the former is to determine the optimal transmit power which should be applied by users during their next transmission, such that their respective fixed target error probabilities are met. The offline optimization is precomputed and involves finding an optimal sequence of error targets which should be pursued whenever packet is at its first, second, etc. attempt. In order to present the advantages of the proposed solution we compare it with the traditional approach in which all devices use dedicated resources in an orthogonal multiple access (OMA) manner. We show that our proposed scheme can achieve the same reliability targets and support the same number of users as the conventional OMA techniques while utilizing lower amount of channel resources. In terms of the signaling overhead, NOMA-HARQ can reuse the scheduling mechanism implemented at the base stations. The only necessary extension is to introduce the capability to indicate the appropriate power to the transmitting users.

Lastly, we remark that although HARQ and NOMA have been analyzed

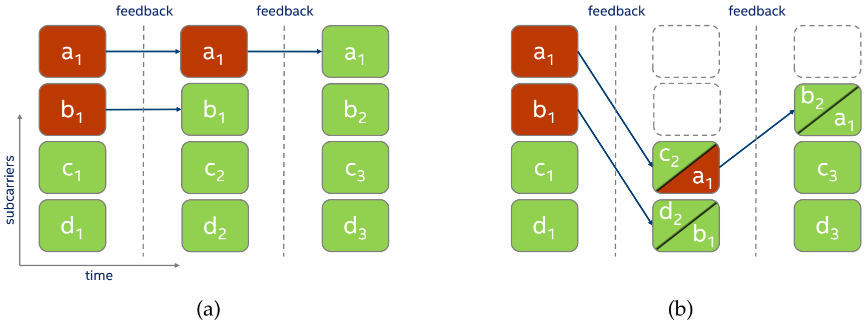


**Fig. E.1:** Comparison of the legacy retransmission mechanism (a) and the proposed NOMA-HARQ (b). In the example there are $N = 4$ transmitters $a, ..., d$. The maximum number of HARQ rounds is $L = 2$. The red color denotes decoding failure while green - success. In the traditional approach UE $a$ which happens to require 2 retransmissions would occupy in total 3 TF-blocks. Using NOMA-HARQ its replicas are transmitted together with packet $c_2$ first and then $b_2$. The resources saved this way can be used to e.g. admit more users or be opportunistically assigned to other type of traffic like eMBB

.

together before, they were considered in the downlink scenario [13], [14] (or a very similar, multicast D2D [15]), which has significantly different characteristics from the uplink NOMA. To the best of the authors' knowledge this contribution is the first to consider uplink traffic with reliability guarantees of multiple users in such a setup/framework.

The rest of the paper is organized as follows. In Section 2 we describe the system and signal model. In Section 3 we go into the details of the proposed scheduling strategy and describe the operation of the receiver. Finally we develop the power optimization procedure which enables the desired performance of the NOMA-HARQ. In Section 4 we explain the simulation setup and introduce the other techniques used as a benchmark for comparison. This is followed by presentation of the results and their discussion. Lastly, in Section 5 we offer final conclusions that close the paper.

## 2 System model

We consider a single cell serving $N$ URLLC-type users transmitting packets of the same, fixed size in the uplink. We assume that each packet is encoded with the same rate $R$ and transmitted over a block of $K$ contiguous time-frequency channel uses, constituting a unit we will refer to as time-frequency (TF) block. To accommodate the users, the base station (BS) can schedule them on up to $N$ distinct TF-blocks used in an orthogonal frequency division multiplexing (OFDM) fashion. Each UE as well as the base station are equipped with a single antenna. The channels between them are assumed to be constant throughout the TF-block[1] but change independently between different transmission attempts (Rayleigh block fading). In our model we consider a coordinated type of communication, in which every data transmission in the uplink is preceded by a scheduling message from the BS that informs the UE about the allocated TF-block and the power that should be used. In order to account for the latency constraints, we assume that each packet can be retransmitted at most $L$ times and the exact number is a system parameter depending on the specific use case and the latency budget.

In order to save resources and take advantage of the generally low probability of packet failures characteristic to URLLC scenarios, the users might be instructed to share their TF-blocks in a non-orthogonal manner. Let $\mathcal{I}^i$ denote the set of indices of the UEs that are scheduled to transmit using the $i$-th TF-block. Furthermore, let the cardinality of that set be equal to $M$. Then, the complex baseband signal received at the $k$-th channel use of the $i$-th TF-block reads

$$y_i(k) = \sum_{j \in \mathcal{I}^i} \sqrt{P_{i,j}} h_{i,j} x_{i,j}(k) + n_i(k) \tag{E.1}$$

where $P_{i,j} \in \mathbb{R}$ is the transmit power applied by user $j$ in TF-block $i$, $h_{i,j} \in \mathbb{C}$ is the channel gain of the $j$-th user over the $i$-th TF-block, $x_{i,j}(k) \in \mathbb{C}$ is a complex transmitted symbol and $n_i(k) \in \mathbb{C}$ is complex additive white Gaussian noise with zero mean and variance $\sigma^2$. Following the aforementioned Rayleigh block fading assumption, the channel coefficients are independent and identically distributed (i.i.d) zero mean circularly symmetric complex gaussian (ZMCSCG) variables with unit variance. Throughout the paper we assume that, prior to each transmission, BS and UEs know only the distribution of the channel gains. Upon reception of a packet, we assume the channel is perfectly estimated by the BS. Furthermore, in this paper users are not allowed to transmit on several TF-blocks simultaneously. Due to that we will omit the index $i$ in the remainder of this paper, whenever it doesn't create ambiguity and a single block is discussed.

One of the main enablers of NOMA is the possibility to perform successive interference cancellation (SIC) [16] by the receiver and iteratively decode signals of distinct UEs. Given the signal model in (E.1) and assuming UEs $1, 2, \ldots, M$ are jointly scheduled in a TF-block and then decoded in an increasing index order, the maximum instantaneous rate of user $j$ is given by

$$R_j = \log_2 \left( 1 + \frac{P_j \left| h_j \right|^2}{\sum_{k=j+1}^{M} P_k \left| h_k \right|^2 + \sigma^2} \right) \tag{E.2}$$

[1] I.e. we assume that the coherence time and bandwidth of the channel are significantly larger than, respectively, the TF-block's duration and bandwidth.

$$p_{er_j}^{(l)} = \begin{cases} \underbrace{1 - \left( \frac{P_j^{(l)}}{P_k^{(m)}\gamma_j^{(l)} + P_j^{(l)}} + \frac{P_k^{(m)}}{P_j^{(l)}\gamma_k^{(m)} + P_k^{(m)}} e^{-\sigma^2 \frac{\gamma_j^{(l)}\gamma_k^{(m)} + \gamma_k^{(m)}}{P_k^{(m)}}} \right) e^{-\frac{\gamma_j^{(l)}\sigma^2}{P_j^{(l)}}}}_{A}, & \text{if } \gamma_j^{(l)}\gamma_k^{(m)} \geq 1 \\ A - \left( \frac{P_k^{(m)}\gamma_j^{(l)}}{P_k^{(m)}\gamma_j^{(l)} + P_j^{(l)}} - \frac{P_k^{(m)}}{P_j^{(l)}\gamma_k^{(m)} + P_k^{(m)}} \right) e^{-\frac{\sigma^2}{1-\gamma_j^{(l)}\gamma_k^{(m)}} \left( \frac{\gamma_j^{(l)}\gamma_k^{(m)} + \gamma_k^{(m)}}{P_k^{(m)}} + \frac{\gamma_j^{(l)}\gamma_k^{(m)} + \gamma_j^{(l)}}{P_j^{l)}} \right)}, & \text{if } \gamma_j^{(l)}\gamma_k^{(m)} < 1 \end{cases} \tag{E.11}$$

and the corresponding packet can be decoded when

$$R_j \geq R \tag{E.3}$$

i.e. when instantaneous rate exceeds the rate $R$ with which the packet was transmitted. Please note, that the expression (E.2) holds true iff the packets of users $1, 2, \ldots, j-1$ were successfully decoded first.

# 3 NOMA-HARQ

The goal of NOMA-HARQ is to obtain a more efficient use of the uplink channel resources while still meeting the high reliability targets of URLLC. The main enablers of the technique are a non-orthogonal allocation scheme of users over the channel resources, along with an optimization of each user's transmit power. In our scheme each user requiring retransmission is paired with a user transmitting a new packet and together they are scheduled on the same TF-block. Consequently, in each uplink round the number of used blocks equals the number of new packets. In this contribution we will consider only the case where the single resource can be shared by at most 2 devices. An example of operation of such retransmission mechanism is demonstrated in Fig.E.1(b).

NOMA-HARQ is based on two optimization procedures: 1) Online power optimization: in each scheduling round, the BS computes the minimum power required by each user to attain their respective error targets. The computations are done based on their current transmission states and under the assumption of Rayleigh fading channels. The powers, as well as TF-block assignment, are then signaled to the UEs through feedback messages. 2) Offline optimization of error-probability targets: the error targets which should be pursued in each particular transmission/retransmission attempt are found, such that they minimize the expected transmit power of all users.

We begin this section by calculating the error probabilities of the users as a function of their transmit powers. These are later used as inputs to the online and offline optimizations in subsections 3.2 and 3.3.

## 3.1 Error probabilities

The probability that the packet of the $j$-th user after $l$ retransmission rounds cannot be decoded (assuming Chase Combining (CC) transmission mode) is given by:

$$p_{er_j}^{(l)} = \Pr\left\{\log_2\left(1+\sum_{i=0}^{l}\mathrm{SINR}_j^{(i)}\right) < R\right\} \tag{E.4}$$

where $\mathrm{SINR}_j^{(i)}$ is the effective (post-processing) SINR of the $i$-th packet replica sent by user $j$ defined as

$$\mathrm{SINR}_j^{(i)} = \frac{P_j^{(i)}\left|h_j^{(i)}\right|^2}{I+\sigma^2} \tag{E.5}$$

and $I$ denotes potential interference power terms coming from other UEs. The formula (E.4) can be further transformed as:

$$p_{er_j}^{(l)} = \Pr\left\{\mathrm{SINR}_j^{(l)} < \underbrace{2^R - 1 - \sum_{i=0}^{l-1}\mathrm{SINR}_j^{(i)}}_{\gamma_j^{(l)}}\right\} \tag{E.6}$$

where we will refer to the term $\gamma_j^{(l)}$ as a residual SINR at round $l$ which is the amount of "signal" missing until the packet can be decoded.

The presented solution can be easily modified to work with the Incremental Redundancy HARQ by redefining the way residual SINR is computed. In that case equation (E.4) becomes:

$$p_{er_j}^{(l)} = \Pr\left\{\sum_{i=0}^{l}\log_2\left(1+\mathrm{SINR}_j^{(i)}\right) < R\right\} \tag{E.7}$$

and consequently:

$$p_{er_j}^{(l)} = \Pr\left\{\mathrm{SINR}_j^{(l)} < \underbrace{\frac{2^R}{\prod_{i=0}^{l-1} 1+\mathrm{SINR}_j^{(i)}} - 1}_{\gamma_{IR_j}^{(l)}}\right\}. \tag{E.8}$$

In general, depending on the used strategy and a number of unsuccessful transmissions a TF-block will be either dedicated or a shared one. When the

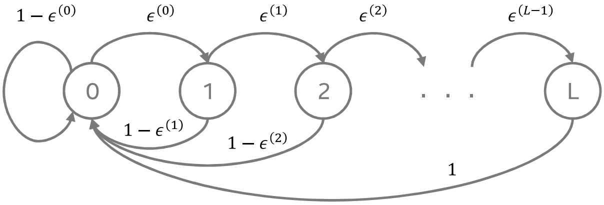


**Fig. E.2:** Markov chain of the transitions between states. The transition from last state $L$ to 0 is always 1 as it happens regardless of the success or failure of the packet.

TF-block is occupied by a single user only, then $I = 0$ in (E.5) and consequently (E.6) has a particularly simple form:

$$p_{er_j}^{(l)} = \Pr\left\{ \frac{P_j^{(l)} \left| h_j^{(l)} \right|^2}{\sigma^2} < \gamma_j^{(l)} \right\} = 1 - e^{-\frac{\gamma_j^{(l)}\sigma^2}{P_j^{(l)}}} \tag{E.9}$$

where we utilized the fact that $\left| h_j^{(l)} \right|^2$ are exponentially distributed due to the Rayleigh fading assumption. Moreover, the term $\gamma_j^{(l)}$ is a constant since it depends only on the SINRs from the previous unsuccessful rounds which are known to the receiver at the time of scheduling a retransmission.

In case the receiver decides to schedule two users in the same TF-block, the derivation becomes more complex as we also need to take into account the impact of successive interference cancellation. In our work we consider the receiver capable of performing SIC in an optimal order, which in practice could be realized as follows. First, the receiver tries to decode both packets. If both failed or both succeeded the procedure is finished for that round. If only one of the packets failed, then the receiver will make another decoding attempt but this time with the interference of the successful user perfectly canceled. Reader should note, that our approach is different from the ones typically encountered in related literature, where either fixed decoding order is assumed [12] or from the strongest to weakest channel gain [11]. These are clearly suboptimal when packets have different target rates, or different residual SNRs due to ongoing retransmissions.

Let us assume that user $j$ is sharing the TF-block with user $k$ who is currently at its $m$-th attempt. The error probability of user $j$ that stems from the optimally ordered SIC can be equivalently written as

$$
\begin{aligned}
p_{er_j}^{(l)} = \Pr & \left\{ \frac{P_j^{(l)} \left|h_j^{(l)}\right|^2}{\sigma^2} < \gamma_j^{(l)}, \frac{P_k^{(m)} \left|h_k^{(m)}\right|^2}{P_j^{(l)} \left|h_j^{(l)}\right|^2 + \sigma^2} > \gamma_k^{(m)} \right\} \\
+ \Pr & \left\{ \frac{P_j^{(l)} \left|h_j^{(l)}\right|^2}{P_k^{(m)} \left|h_k^{(m)}\right|^2 + \sigma^2} < \gamma_j^{(l)}, \frac{P_k^{(m)} \left|h_k^{(m)}\right|^2}{P_j^{(l)} \left|h_j^{(l)}\right|^2 + \sigma^2} < \gamma_k^{(m)} \right\}
\end{aligned} \tag{E.10}
$$

where the first term represents the error probability with interferer's signal decoded and canceled, while the second term corresponds to the case when SIC is not successful. The error probability of the other user $p_{er_k}^{(m)}$ is analogously obtained by interchanging the indices $j \leftrightarrow k$ and $l \leftrightarrow m$. The expression (E.10) can be obtained in the closed form and is given by (E.11).

## 3.2 Power optimization

Let us now assume that receiver expects the packet failure probability at round $l$ to be no larger than a certain target denoted as $\epsilon^{(l)}$. The goal is to find a minimum transmit power which guarantees that. In case of the dedicated TF-block the optimal value follows directly from (E.9) with $p_{er_j}^{(l)}$ set to $\epsilon^{(l)}$ and is equal:

$$
P_j^{(l)} = -\frac{\gamma_j^{(l)} \sigma^2}{\ln(1 - \epsilon^{(l)})}. \tag{E.12}
$$

When the TF-block is shared, finding the optimal powers requires solving an optimization problem:

$$
\underset{\{P_j^{(l)}, P_k^{(m)}\}}{\arg\min} \quad P_j^{(l)} + P_k^{(m)} \tag{E.13a}
$$

$$
\text{s.t.} \quad p_{er_j}^{(l)} \leq \epsilon^{(l)}, \tag{E.13b}
$$

$$
p_{er_k}^{(m)} \leq \epsilon^{(m)}. \tag{E.13c}
$$

The steps described above constitute an "online" optimization. It is performed by the receiver at each feedback stage in order to find the optimal powers for the next round of uplink transmissions. In the simulations presented in Section IV, a standard interior-point convex solver is used to solve the optimization problem. Although the constraint functions are not convex, there are two disjoint regions in which the local minimums can be found: $P_j^{(l)} > P_k^{(m)}$ and $P_j^{(l)} < P_k^{(m)}$. The global minimum is found by determining the minimum for each of the two disjoint regions and selecting the lower one.

To perform online optimization BS needs to know only the noise variance $\sigma^2$ and residuals $\gamma_j^{(l)}$, $\gamma_k^{(m)}$, which depend on the SINRs of past replicas of the packet. As per our assumption, those previous SINRs are known since the corresponding channel gains can be perfectly estimated (e.g. from pilot symbols) once the packet is received. The way it is formulated, online optimization is solved for a fixed set of error targets $\boldsymbol{\epsilon} = \left[\epsilon^{(0)}, \ldots, \epsilon^{(L)}\right]$. These are the subject of the next subsection.

## 3.3 Optimization of Error-Probability Targets

Clearly, the selection of a sequence $\boldsymbol{\epsilon}$ will have a large impact on the overall power efficiency of the system, hence the following offline optimization is necessary:

$$\underset{\{\epsilon^{(0)}, \ldots, \epsilon^{(L)}\}}{\text{arg min}} \quad E\left[\sum_{j=1}^{N}\left(P_j^{(0)} + \sum_{l=1}^{L} P_j^{(l)} \prod_{i=0}^{l-1} \epsilon^{(i)}\right)\right] \tag{E.14a}$$

s.t.

$$\prod_{l=0}^{L} \epsilon^{(l)} \leq \epsilon_{tar}, \tag{E.14b}$$

$$\sum_{i=\lfloor N/2 \rfloor+1}^{N} \binom{N}{i} \pi_0^{N-i} \left(1-\pi_0\right)^i \leq \epsilon_c \tag{E.14c}$$

where $\pi_0$ denotes the probability that a user's transmission corresponds to the transmission of a new packet. Because in our scheme we always combine retransmissions (occurring with probability $1-\pi_0$) with new packets, the constraint (E.14c) is there to ensure that the situations where the former outnumber the latter are reasonably rare. The constraint (E.14b) is straightforward and simply guarantees that the final error probability after $L$ retransmissions is below the required target. The objective (E.14a) itself is to minimize the average power each user spends on a packet, i.e. including potential retransmissions.

If we denote the device transmitting a new packet as being in state 0, device sending the first retransmission in state 1, etc., then the sequence of states of each UE can be approximated by a Markov chain as depicted in the state transition diagram in Fig.E.2. Following a standard solving procedure we can find the stationary probabilities $(\pi_0, \pi_1, \ldots, \pi_L)$ which represent a fraction of time a device spends in each particular state. For our problem only the first quantity and its complement are relevant as we can treat all the

**Table E.1:** Simulation parameters

| | |
|---|---|
| Number of UEs $N$ | 40 |
| Number of retransmissions $L$ | $1, 2$ |
| Final BLER $\epsilon_{tar}$ | $10^{-5}$ |
| Congestion probability $\epsilon_c$ | $10^{-5}$ |
| Transmission rate $R$ | variable, 0.5 - 3 bits/channel use |
| Channel Type | Rayleigh block fading |
| Channel estimation method | Perfect |

retransmission states jointly:

$$\pi_0 = \frac{1}{1 + \sum_{l=0}^{L-1} \prod_{i=0}^{l} \epsilon^{(i)}}. \tag{E.15}$$

Since each of the $N$ devices follows its own Markov chain independently, arriving at expression (E.14c) is straightforward and is in fact a complementary CDF of the binomial distribution $1 - F_{Binomial}\left(\lfloor N/2 \rfloor; N, 1 - \pi_0\right)$.

While computationally heavy, solution to problem (E.14) can be precomputed offline and doesn't need to be updated regularly as it depends mainly on the maximum allowed number of retransmissions $L$ and the final BLER target $\epsilon_{tar}$ which are system parameters.

# 4 Results

In the following section, we present the results obtained through system-level simulations. The parameters used to obtain them are gathered in Table E.1.

In addition to the introduced NOMA-HARQ approach we investigate the performance of two other techniques which both rely on a legacy OMA paradigm.[2] In those baseline schemes each transmission occurs on dedicated resources. Consequently, in each round all $N$ TF-blocks are being used, carrying either a new packet or a retransmission. The OMA schemes considered here can be further divided into sub-types:

- Fixed power: the simplest case where each packet has the same, fixed power (whether it is the first transmission or a retransmission), and its value is chosen such that the final error rate requirement is fulfilled.

[2] To the best of the authors' knowledge there are no direct competitors among other NOMA-based techniques proposed in the literature. Such solutions typically focus on maximizng the sum-rate and do not operate with fixed error targets, precluding any direct comparison with NOMA-HARQ.

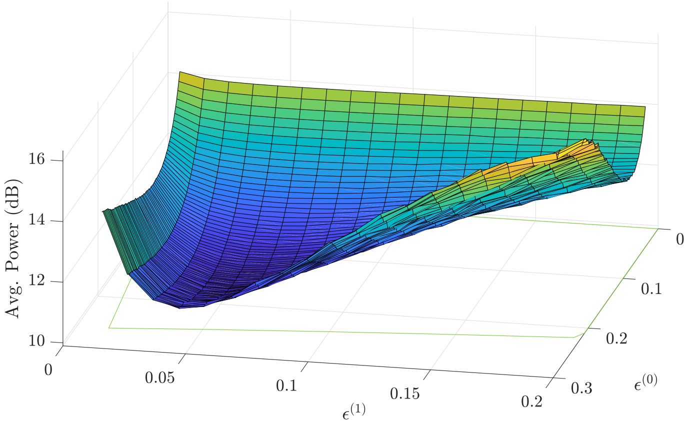


**Fig. E.3:** Average power as a function of $\epsilon^{(0)}$ and $\epsilon^{(1)}$ for $L = 2$, $R = 1$, $\epsilon_{tar} = 10^{-5}$ and $N = 40$. Example of the offline optimization where the minimum is achieved for $\epsilon^{(0)} = 0.16$, $\epsilon^{(1)} = 0.05$ (and consequently) $\epsilon^{(2)} = 0.00125$. The green outline on the bottom denotes feasible region imposed by (E.14c).

- Optimal: the transmit power is optimized according to the same procedure as described in the section 3, i.e. online and offline optimization is employed but the former requires only expression (E.12) as no sharing is involved.

Although we do not explicitly consider the latency in this contribution, we analyze the performance of the presented schemes under very limited number of allowed retransmissions. While URLLC is still in the standardization phase, it is realistic to assume that even with short packets (mini-slots spanning 7 OFDM symbols or less) and higher subcarrier-spacing the number of allowed retransmissions will have to be limited to no more than 1 or 2 so that the challenging latency requirements ranging from 0.5ms to 2ms can be met. Furthermore, due to the small payload sizes and required robustness of transmission we won't be typically interested in high rate regime.

To start the discussion, in Fig. E.3 we show the dependency between selected target error rates $\boldsymbol{\epsilon}$ and the average power spent on each packet. It is the visualization of the performed optimization (E.14) in the case of $L = 2$ allowed retransmissions. The feasible region ends around $\epsilon^{(0)} = 0.25$ as it starts to violate constraint (E.14c) but it does not impact the global optimum. Fig. E.3 illustrates that the optimization of the error targets tends to set more relaxed targets for early transmissions, in order to avoid unnecessary power expenditure, while the error targets become more demanding as we approach the maximum number of retransmissions. Because such late retransmissions occur very infrequently, the large power needed to attain their error targets does not significantly increase the average power used by the UEs.

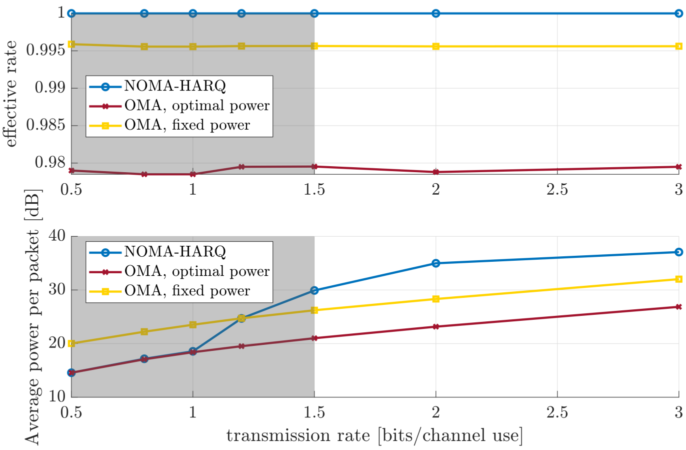


**Fig. E.4:** Average power spent per packet and effective rate comparison in the case of maximum $L = 1$ allowed retransmissions.

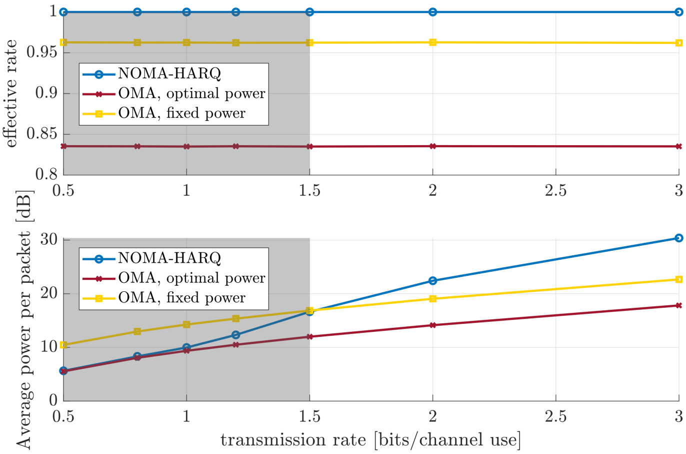


**Fig. E.5:** Average power spent per packet and effective rate comparison in the case of maximum $L = 2$ allowed retransmissions.

The core of the results is shown in Fig. E.4 and Fig. E.5. In addition to the average power metric we present also the effective rate performance of the discussed techniques, which we define as the ratio between the number of successfully decoded packets divided by the total number of used TF-blocks. Unlike NOMA-HARQ and optimal OMA, the provided results for fixed power OMA can be derived analytically and we discuss this further in the Appendix A.

As we can see, in case of single retransmission the effective rate gains with NOMA-HARQ are rather limited. This is justified as the effective rate is mostly determined by the initial $\epsilon^{(0)}$ which has to be quite low already in the first attempt. Compared to the approach based on dedicated resources and optimal power the savings in terms of bandwidth (scheduled TF-blocks) reach 2%.

With respect to the average power NOMA-HARQ proves to work best in the low rate regime (marked in grey on both figures). In fact, up to a certain point it can provide the same level of reliability while requiring less energy than the fixed power OMA and initially even approaching the optimal OMA.

In Fig. E.5 we can observe much better the attractiveness of the proposed technique. As the number of allowed retransmissions rises to 2 the value of initial target $\epsilon^{(0)}$ becomes less stringent giving more headroom for improvement. Compared with the fixed power and optimal OMA the presented NOMA-HARQ allows to save respectively 4% and 18% of the bandwidth resources. In addition to that, the range of rates over which NOMA-HARQ outperforms fixed power OMA in terms of average power increases to 1.5 bits/ch.u.

In Fig. E.6 we look into the power performance of the schemes from a different perspective and, for completeness, in Fig. E.7 we provide the corresponding error targets. Instead of focusing on average power per packet, we analyze the average power used to transmit state 0, 1 and 2 packets. For low rates NOMA-HARQ uses almost exactly the same power as the optimal OMA at every stage. The divergence between NOMA-HARQ and its OMA counterpart appears as the transmissions' rate is increased beyond 1. Worth noticing is the fact that even for high rates significant difference is present only at later stages (during retransmissions) and not for the first attempt. The reason for that is two-fold. Firstly, NOMA-HARQ has only slightly lower error target $\epsilon^{(0)}$ than optimal OMA (cf. E.7). Secondly, when coupled with some other device's retransmission, the optimal powers following from (E.13) in most cases will be such that the new packet is the "weaker" one. As a result, $P_j^{(0)}$ of NOMA-HARQ is low and close to that of optimal OMA, which is important since initial transmission is the one having the biggest impact on the average power per packet. We also remark that the error targets vary only slightly with respect to the selected rate, simplifying the implementation of NOMA-HARQ in practical systems.

It is important to realize why there is a region of rates over which NOMA-HARQ performs particularly well and the explanation for that lies in solution (E.11). Whenever the product of residual SINRs is below 1 the error probability for both UEs is reduced and consequently achieving a certain target rate requires less power. Motivated by that finding, in Fig. E.8 we analyze the achievable outage probability of one-shot, low-rate transmissions as a function of their transmit power. Similarly as before, we want to compare the

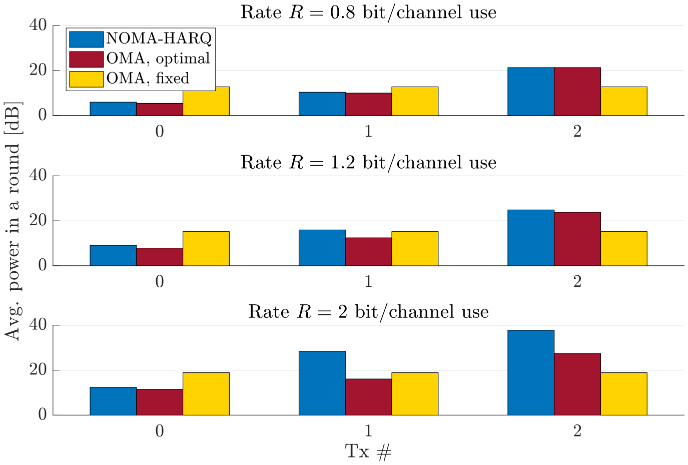


**Fig. E.6:** Average power spent at each round with $L = 2$ retransmissions.

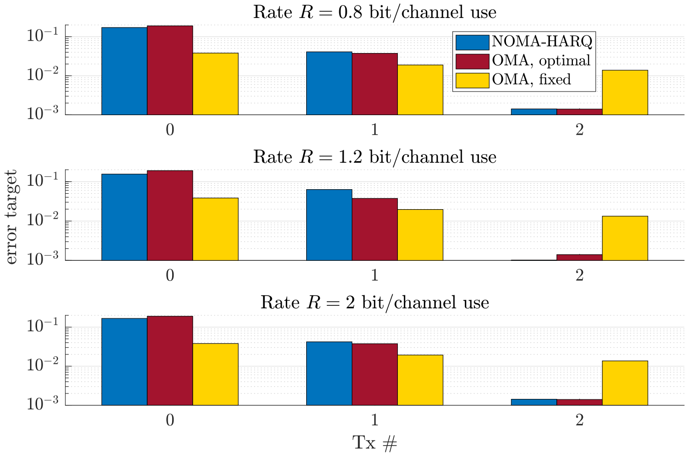


**Fig. E.7:** Optimal error targets for each round when $L = 2$ retransmissions. Note that for fixed power OMA the error targets are not set in advance and are the consequence of the interplay between fixed power, distribution of channel gains and distribution of residual SNRs.

performance of the single device using dedicated resource and two devices sharing the TF-block. The outage probability of one UE is given by (E.9) and for two users we find it by solving a mini-max problem:

$$\min_{\{P_1, P_2\}} \quad \max_{i=1,2} \left(p_{er_i}\right) \tag{E.16a}$$

$$\text{s.t.} \quad P_1 + P_2 = P_t. \tag{E.16b}$$

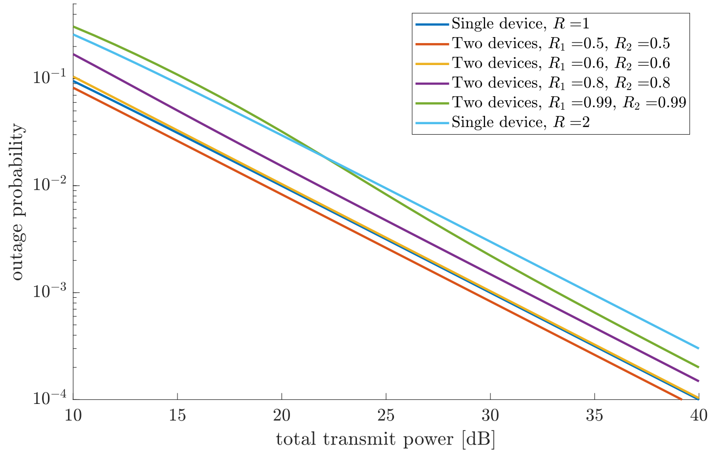


**Fig. E.8:** Achievable outage probability as a function of transmit power.

i.e. we minimize the worst case outage probability, under the condition that the sum of two powers is equal to that of a device transmitting alone.

As shown in Fig. E.8, with NOMA approach two UEs can share physical resources while having the same sum-rate and using same amount of power as the traditional OMA user and yet achieve lower outage probability (red and blue lines). If we decide to keep the same error rate as the single device, then the rate of both coexisting users can be increased to 0.6 giving a total boost of 20% in terms of sum-rate. Similarly, we can compare one user transmitting with rate 2 and two users transmitting with rates close to 1 (consequence of the eq. (E.11)). The visible crossing point around 22ldB is due to the fact that below this amount there is not enough power for NOMA to split it optimally between the two users.

# 5 Conclusions

In this publication we propose a novel approach where HARQ mechanism is assisted by NOMA to free up the portion of the bandwith that would otherwise be used to carry retransmissions, hence resulting in a more spectrally efficient system. The contribution relies on the optimization techniques which ensure reliable communication while keeping the average power minimized. In the paper we confront the presented NOMA-HARQ with its competitors that rely on legacy OMA. Based on the provided results we conclude that for URLLC applications with limited number of retransmissions NOMA-HARQ can offer considerable gains in terms of spectral efficiency. Moreover, whenever use cases requiring robust, low rate transmissions are considered, NOMA-HARQ is capable of outperforming naive OMA approach also in terms of average utilized power. Compared to classical HARQ operation, the spectral and power efficiency benefits of NOMA-HARQ come only at the expense of extending the standard scheduling messages sent by the BS with signaling of the UE's power.

# A Appendix

Since channel realizations are independent between different attempts, the sum of SNRs after $L+1$ transmissions is a sum of $L+1$ exponentially distributed random variables with the same scale parameter equal to the selected power $P_{fixed}$. This in turn is a gamma distribution with shape $k = L+1$ and scale $\theta = P_{fixed}$. Finding the appropriate power is then a simple numerical problem which requires finding the smallest $P_{fixed}$ such that

$$F_{gamma}\left(2^R - 1; L+1, P_{fixed}\right) \leq \epsilon_{tar}. \tag{E.17}$$

With the $P_{fixed}$ determined, the average number of TF-blocks used per packet reads

$$\alpha = 1 + \sum_{l=1}^{L} F_{gamma}\left(2^R - 1; l, P_{fixed}\right). \tag{E.18}$$

Finally, the average power spent per packet is $\alpha P_{fixed}$ and the effective rate is $\frac{1}{\alpha}$.

# Acknowledgment

This work has been supported by the cooperative project Virtuoso, which was funded by: Innovation Fund Denmark, Intel Mobile Communications, Telenor, Keysight Technologies, and Aalborg University.

# References

[1] 3GPP, "Service requirements for the 5G system; Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.261, 2018, v16.5.0.

[2] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity*, 2014, pp. 146–151.

[3] S. Nagata, L. H. Wang, and K. Takeda, "Industry Perspectives: Latency reduction toward 5G," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 2–4, 2017.

[4] 3GPP, "Study on latency reduction techniques for LTE," 3rd Generation Partnership Project (3GPP), TR 36.881, 2016, v14.0.0.

[5] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[6] P. Popovski, J. J. Nielsen, C. Stefanovic, E. d. Carvalho, E. Strom, K. F. Trillingsgaard, A.-S. Bana, D. M. Kim, R. Kotaba, J. Park, and R. B. Sorensen, "Wireless Access for Ultra-Reliable Low-Latency Communication: Principles and Building Blocks," *IEEE Network*, vol. 32, no. 2, pp. 16–23, 2018.

[7] S. Parkvall, E. Dahlman, P. Frenger, P. Beming, and M. Persson, "The evolution of WCDMA towards higher speed downlink packet data access," in *IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings (Cat. No.01CH37202)*, vol. 3, 2001, pp. 2287–2291 vol.3.

[8] A. Anand and G. de Veciana, "Resource allocation and HARQ optimization for URLLC traffic in 5G wireless networks," *CoRR*, vol. abs/1804.09201, 2018. [Online]. Available: http://arxiv.org/abs/1804.09201

[9] S. Sesia, I. Toufik, and M. Baker, *LTE-The UMTS Long Term Evolution: From Theory to Practice*, 2nd ed. John Wiley & Sons Ltd, 2011.

[10] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Communications Magazine*, vol. 53, no. 9, pp. 74–81, 2015.

[11] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink Nonorthogonal Multiple Access in 5G Systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, 2016.

[12] J. Choi, "On Power and Rate Allocation for Coded Uplink NOMA in a Multicarrier System," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2762–2772, 2018.

[13] ——, "On HARQ-IR for Downlink NOMA Systems," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3576–3584, 2016.

[14] Y. Xu, D. Cai, F. Fang, Z. Ding, C. Shen, and G. Zhu, "Outage Analysis and Power Allocation for HARQ-CC Enabled NOMA Downlink Transmission," in *2018 IEEE Global Communications Conference (GLOBECOM)*, 2018, pp. 1–6.

[15] Z. Shi, S. Ma, H. ElSawy, G. Yang, and M.-S. Alouini, "Cooperative HARQ-Assisted NOMA Scheme in Large-Scale D2D Networks," *IEEE Transactions on Communications*, vol. 66, no. 9, pp. 4286–4302, 2018.

[16] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-Orthogonal Multiple Access (NOMA) for Cellular Future Radio Access," in *2013 IEEE 77th Vehicular Technology Conference (VTC Spring)*, 2013, pp. 1–5.

# Paper F

## How URLLC can Benefit from NOMA-based Retransmissions

Radosław Kotaba, Carles Navarro Manchón, Tommaso Balercia and Petar Popovski

The paper has been submitted in the
*IEEE Transactions on Wireless Communications* Vol. 20, No. 3, pp. 1684–1699,
2021.

# Abstract

*Among the new types of connectivity unleashed by the emerging 5G wireless systems, Ultra-Reliable Low Latency Communication (URLLC) is perhaps the most innovative, yet challenging one. Ultra-reliability requires high levels of diversity, however, the reactive approach based on packet retransmission in HARQ protocols should be applied carefully to conform to the stringent latency constraints. The main premise of this paper is that the NOMA principle can be used to achieve highly efficient retransmissions by allowing concurrent use of wireless resources in the uplink. We introduce a comprehensive solution that accommodates multiple intermittently active users, each with its own HARQ process. The performance is investigated under two different assumptions about the Channel State Information (CSI) availability: statistical and instantaneous. The results show that NOMA can indeed lead to highly efficient system operation compared to the case in which all HARQ processes are run orthogonally.*

***Keywords***— HARQ, NOMA, radio resource management, uplink, URLLC

# 1 Introduction

The fifth generation (5G) wireless networks are slowly becoming a reality. While historically the primary motivation behind each new generation was to increase data rates, coverage and other metrics related to the quality of experience of the users, 5G promises to be more than just an incremental improvement over previous technologies [1, 2]. This shift is driven by a growing popularity and rapid advancements in the area of Internet of Things (IoT) which represents a different, non-human-centric communication paradigm. Among those new, emerging applications, especially prominent are those that fall into the category of ultra-reliable low-latency communications (URLLC). Examples of such use cases include: smart cities, factory automation (Industry 4.0) [3], and tactile Internet (involving remote motion control, telesurgery, etc.) [4]. To enable those demanding applications, the underlying network will need to provide MAC-layer end-to-end latencies from 0.5 to few milliseconds and reliability (defined as the probability of successful delivery of the packet within the stipulated latency) above $99,999\%$ [5].

Designing an efficient URLLC system capable of meeting the aforementioned requirements poses a significant challenge, especially considering the fundamental tradeoffs between latency, reliability, spectral efficiency, and power consumption [6]. While it has been shown that on their own legacy systems are either not able to operate in URLLC regime [7], or become prohibitively inefficient [8], many of the concepts they use are still valid and can be adapted to this new paradigm. Diversity-providing mechanisms are particularly crucial, since they are an unavoidable necessity when facing stringent reliability requirements.

One such mechanism is hybrid automatic repeat request (HARQ), which provides diversity in a reactive way upon reporting of an error by the receiver. Its flexibility and the potential to offer significant gains have been thoroughly studied both theoretically [9] and in practical scenarios [10] which led to the implementation of HARQ in the third generation system HSPA and onwards. While applying HARQ in URLLC is challenging due to the stringent latency constraint, we note that the alternative for reaching high reliability through one-shot transmission [11] is very inefficient in terms of power and, whenever feasible, some form of HARQ is highly desirable. As shown in [12], even with latency budget as low as 1 ms, the new 5G features including: shortened transmission time intervals (TTIs), higher subcarrier spacing and improved processing times will allow for at least one retransmission opportunity.

While generally beneficial, especially as a mechanism to enhance reliability, HARQ in URLLC should be designed in a lean way and avoid inefficiencies. First, as the amount of time-frequency resources in the system is finite, the need to accommodate both new packets and retransmissions increases the probability of queuing which is especially detrimental for URLLC. Second, as the system preserves the previous unsuccessful copies of the packet, retransmission of the full payload can be wasteful. Meanwhile, practical systems prefer to work with fixed-size resources where adapting the size of the retransmissions is not possible.

The shortcomings of HARQ can be mitigated with the help of non-orthogonal multiple access (NOMA). This technique involves transmitting multiple packets over the same time-frequency resources thereby intentionally introducing interference. Due to its ability to accommodate more users and reduce latency, NOMA has been identified by researchers as one of the enablers of URLLC [13]. For a comprehensive overview of this topic and a discussion on different existing variants of NOMA reader is directed to [14] [15]. Our motivation for using NOMA is the fact that it can address the HARQ inefficiencies described earlier and allow efficient use of the time-frequency resources. More specifically, with NOMA-HARQ the transmissions and retransmissions can be sent non-orthogonally to increase the throughput of the system and avoid queuing during the periods of congestion.

## 1.1 Related work

On their own, both HARQ and NOMA topics have been extensively covered in the literature. In [16], the authors optimize the average power of HARQ with finite number of retransmissions and a given outage probability target. The Chase combining (CC) variant is assumed, Rayleigh fading channel and a single bit feedback. The incremental redundancy (IR) type HARQ is studied in [17], where the aim is to maximize the throughput for a given reliability constraint. This is achieved through rate adaptation, however the assumption of a full buffer used there might not be suitable for all URLLC applications. In [18] and [19], the authors investigated HARQ explicitly in the URLLC context by considering transmission of short packets (finite blocklength regime) over additive white Gaussian noise (AWGN) channel. Moreover, in their optimization problems authors consider the impact of the feedback delay and overall energy budget.

The relevant work on NOMA include the following. In [20] the authors consider URLLC use case in the downlink, and propose an energy-efficient resource allocation for a two-user heterogeneous NOMA system. While the literature on uplink NOMA is not as extensive as on its downlink counterpart, some interesting contributions can be found in [21, 22], although they do not explicitly address URLLC. The former provides insights into the achievable sum-rate and outage probability with a given transmit power, while the latter discusses rate and power allocation scheme that ensures required probability of error.

As far as solutions combining both HARQ and NOMA are concerned, the literature is even more scarce. Some of the reported works include [23–25], but except for [25], they do not consider uplink scenario which entails radically different system model. The fundamental difference is that, in the downlink, the power ratio between the non-orthogonal signals is fully controlled by the base station (BS) and does not depend on the specific channels of the users. In the uplink, on the other hand, both slow and fast fading have an impact on the ratio in addition to the allocated power. This involves a larger degree of uncertainty for uplink when trying to control the SINR of the non-orthogonal transmissions, which significantly complicates the problem compared to the downlink case. To the best of the authors knowledge, none of the contributions on NOMA and HARQ deal with the comprehensive, multi-user, uplink scenario where the amount of resources is finite and the effects of queuing are considered.

## 1.2 Contributions

In this work we investigate the performance of uplink OMA and NOMA systems employing HARQ mechanism. Considered framework involves very limited number of retransmission opportunities and tight reliability constraints which are meant to conform to the URLLC use case and hence provide useful insights into the design of practical systems. As a main contribution of this paper, we develop a comprehensive solution involving power allocation and packet scheduling that efficiently accommodates multiple intermittently active URLLC users, each with its own HARQ process, over a finite pool of resources. We achieve this by decoupling the two problems. First, we formulate the power allocation problem as a minimization of the average transmit power subject to the reliability constraints. This is done by finding optimal error targets per each HARQ round. Next, a joint scheduling problem is considered, where we develop a simple heuristic that allows to make a decision which packets should be prioritized in case of insufficient resources based on the optimal error targets and transmit powers determined in the earlier step.

The solution outlined above is developed in two variants, based on OMA and NOMA principle. Furthermore, for each of them we propose two different approaches depending on the type of channel state information (CSI) available: *Statistical CSI*, where only the distribution of channel realizations is known and *Instantaneous CSI*, where additionally the channel conditions for the immediate uplink transmission are known in advance. In the former case, we study and compare the performance of CC and IR HARQ techniques assuming asymptotic (infinite) blocklength. In the Instantaneous CSI case, where the channel at hand becomes AWGN, we develop the methodology and analyze OMA- and NOMA-HARQ in incremental redundancy mode under the finite-blocklength assumption. The two CSI scenarios are meant to provide bounds on how the channel knowledge can impact the performance. The proposed approaches are evaluated by means of Monte Carlo simulations, revealing that our NOMA schemes can effectively deal with more than twice as much URLLC load as their OMA counterparts using the same amount of channel resources. This increase in the system capacity can be achieved with only a slight increase in transmit power, and in some cases (when the traffic intensity exceeds the servicing capabilities of OMA) even more efficiently than OMA.

This contribution extends the prior study [26] by introducing significant changes. Most notably, as the activation of users is no longer deterministic and the total user population is larger than the amount of channel resources queuing issues need to be taken into account. The signal model now includes the effect of distant-dependent large scale fading, which impacts the power assignment in NOMA, as some UEs become preferable to the others. Furthermore, unlike in [26], we do not restrict the pairing in NOMA to be only between new packets and retransmissions. Instead, we generalize the approach and allow the packets to be scheduled non-orthogonally in whichever way that minimizes the total power spent.

The rest of the paper is organized as follows. In Section 2 we describe the system and signal model. In Section 3 we go into the details of optimal error targets and power allocation for OMA and NOMA with statistical CSI. In Section 4 we extend the discussion to the known channel case and finite blocklength communication. In Section 5 we discuss briefly the scheduling and resource allocation technique. In Section 6 we present the simulation results together with their thorough discussion. Lastly, in Section 7 we offer final conclusions that close the paper.

## 2 System model

In this work, we consider a single cell serving the uplink traffic of $N$ devices running URLLC applications. We assume that the packets of each UE are of the same, fixed size and span $K$ channel uses, i.e. symbols. Moreover, they carry the same amount of $B$ information bits leading to equal rates which we denote as $R = B/K$ [bits/symbol]. The $K$ channel uses that constitute a packet occupy a contiguous block of time-frequency resources which we will interchangeably refer to as TF-block or slot. A single TF-block is considered smaller than the coherence bandwidth/time and distinct slots experience independent Rayleigh fading. The number of available TF-blocks is limited to $W$ per uplink phase, and the base station's (BS) goal is to best distribute them between UEs' transmissions. The generation of new packets at each device $j$ is intermittent and occurs with probability $b$. Whenever new packet appears, UE sends a scheduling request (SR) for that packet in the next available uplink phase and consequently receives the grant from the BS with instructions regarding time-frequency resource allocation and appropriate transmit power. It is further assumed that this step is error-free and happens in parallel with the usual exchange of packets that are carrying payload, i.e. there are dedicated resources for SRs and in a single uplink phase UE can send both the previously scheduled packets as well as a new request. These assumptions are reasonable in the URLLC context considering the stringent latency requirements. As such, we can view the scheduling handshake procedure as transparent as it simply creates a constant offset between the arrival of the new packet at UE's buffer and the moment it is transmitted. Hence, in the remainder of this paper we will simply say that in each uplink phase device will transmit a new packet with probability $b$.

Due to the latency requirement of URLLC, we assume that once a new packet is generated it can be transmitted only during the next $L+1$ uplink phases and is dropped otherwise. Consequently, unsuccessful packets can be retransmitted during that window to increase the reliability (up to $L$ times if every opportunity is used). Two variants of the HARQ mechanism are considered for this: Chase combining (CC) and incremental redundancy (IR).

Following the NOMA principle, in this work we admit the possibility of users sharing the same resources. Let us denote by $\mathcal{I}^i$ the set of indices of the UEs transmitting over $i$-th TF-block. The complex baseband signal received over its $K$ channel uses can be written as

$$\mathbf{y}_i = \sum_{j \in \mathcal{I}^i} \sqrt{P_{i,j}} g_{i,j} \mathbf{x}_{i,j} + \mathbf{n}_i \tag{F.1}$$

where $P_{i,j} \in \mathbb{R}$ is the transmit power of user $j$ in TF-block $i$, $g_{i,j} \in \mathbb{C}$ is the channel between $j$-th UE and the BS over the $i$-th TF-block, $\mathbf{x}_{i,j} \in \mathbb{C}^K$ are the complex transmitted symbols assumed to be Gaussian distributed with zero mean and unit variance and $\mathbf{n}_i \in \mathbb{C}^K$ is complex additive white Gaussian noise with zero mean and variance $\sigma^2$. The channel coefficients are given by $g_{i,j} = \frac{h_{i,j}}{\sqrt{d_j^{\alpha}}}$, where $h_{i,j}$ is the Rayleigh fading component, which is independent and identically distributed (i.i.d) zero mean circularly symmetric complex Gaussian (ZMCSCG) random variable with unit variance, while $d_j^{\alpha}$ is a pathloss term accounting for the distance between UE $j$ and the BS. The realizations of $h_{i,j}$ change between different transmissions while the distance $d_j$ remains constant for a particular user.

In this work we consider two different scenarios that provide the bounds on the performance of presented HARQ schemes.

#### Statistical CSI

Similarly to the work presented in [26] we assume here that at the time of scheduling new transmissions the base station has only a statistical knowledge about the future channel realizations $h_{i,j}$, i.e. that they are i.i.d. ZMCSCG. BS knows however the distances $\mathbf{d}$ of all users and hence knows the variance of $\mathbf{g}$.

#### Instantaneous CSI

In this scenario we assume that BS knows the CSI of the next transmission at the time of performing the scheduling, i.e. it knows the channel coefficients $\mathbf{g}$ for the immediate uplink stage, but not the ones coming afterwards[1].

## 2.1 Base Station operation

A simplified diagram explaining the principle of operation of the receiver is shown in Fig. F.1. As discussed earlier, UEs generate new packets independently with probability $b$, so the resulting total number of scheduling requests received by the BS is given by a binomial distribution with $N$ trials and success probability $b$. The new packets are referred to as being in state/round 0, while all those that arrived earlier and failed the decoding (or were not transmitted) belong to any of the other $1, \ldots, L$ rounds. The BS considers all the packets that are currently in the system and performs their joint scheduling, which involves determining the appropriate assignment of TF-blocks and transmit powers. The exact procedure governing this

[1] In practice, obtaining the CSI involves auxiliary procedures that can deteriorate the reliability. By neglecting these we gain an insight into the upper bound of the performance in such scenario.

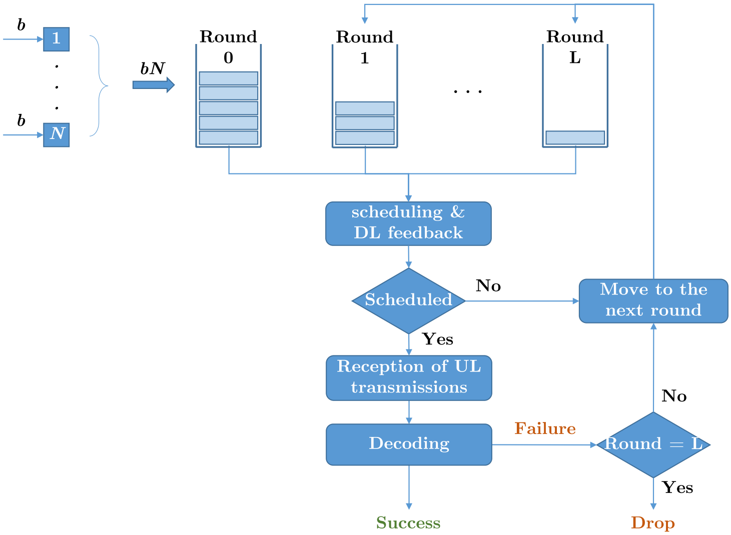


**Fig. F.1:** Base station operation

step is described in detail in section 5. The information regarding scheduling and power allocation is then signaled to all concerned UEs so that they can perform coordinated transmission in the upcoming uplink phase. Note that, when the number of resources $W$ is finite, it might not be possible to schedule all the packets, in which case they are moved directly to their next round as if they failed or were transmitted with power 0. In the decoding step, if NOMA is employed, it is assumed that the receiver is capable of successive interference cancellation (SIC) and depending on the use case we will consider either optimal or fixed decoding order.

# 3 HARQ with Statistical CSI

## 3.1 OMA-HARQ

Let us start by analyzing a simpler approach where the base station is allowed to schedule uplink transmissions only in an orthogonal manner, dedicating one TF-block for each packet. The SNR of the packet received from user $j$, conditioned on its power $P_j$ and distance from the BS $d_j$, is distributed exponentially according to the pdf

$$f_e\left(x; \frac{P_j}{d_j^\alpha \sigma^2}\right) = \frac{d_j^\alpha \sigma^2}{P_j} e^{-\frac{x d_j^\alpha \sigma^2}{P_j}}. \tag{F.2}$$

Taking into account prior unsuccessful transmissions ans assuming CC is used, the decoding failure probability after $l$-th attempt (counting from 0 as the initial one) is given by [9]

$$p_{er,cc_j}^{(l)} = \Pr\left\{ \log_2\left(1 + \sum_{i=0}^{l} \mathrm{SNR}_j^{(i)}\right) < R \right\} \tag{F.3}$$

where $\mathrm{SNR}_j^{(i)}$ is the SNR of $j$-th UE's packet in its $i$-th attempt. When IR-HARQ is used, then

$$p_{er,ir_j}^{(l)} = \Pr\left\{ \sum_{i=0}^{l} \log_2\left(1 + \mathrm{SNR}_j^{(i)}\right) < R \right\}. \tag{F.4}$$

The two expressions can be rearranged to depend only on the last packet realization since all the previous SNRs are already known

$$p_{er,cc_j}^{(l)} = \Pr\left\{ \mathrm{SNR}_j^{(l)} < 2^R - 1 - \sum_{i=0}^{l-1} \mathrm{SNR}_j^{(i)} = \gamma_{cc_j}^{(l)} \right\} \tag{F.5}$$

$$p_{er,ir_j}^{(l)} = \Pr\left\{ \mathrm{SNR}_j^{(l)} < \frac{2^R}{\prod_{i=0}^{l-1}\left(1 + \mathrm{SNR}_j^{(i)}\right)} - 1 = \gamma_{ir_j}^{(l)} \right\}. \tag{F.6}$$

To simplify the notation we introduce the terms $\gamma_{cc_j}^{(l)}$ and $\gamma_{ir_j}^{(l)}$ denoting a "residual SNR" which is the amount of signal power needed until the packet can be decoded (at $l = 0$ simply equal to $2^R - 1$). Throughout this paper we will typically omit the $_{cc}/_{ir}$ subscript since each method is discussed in a dedicated section making it clear which definition is used.

By combining eq. (F.2) with either of the two (F.5), (F.6) the error probability is obtained:

$$p_{er_j}^{(l)} = 1 - e^{-\frac{\gamma_j^{(l)} d_j^{\alpha} \sigma^2}{P_j^{(l)}}}. \tag{F.7}$$

It further follows from (F.7) that the minimum power required to achieve certain target error $p_{er_j}^{(l)} = \epsilon_j^{(l)}$ is

$$P_j^{(l)} = -\frac{\gamma_j^{(l)} d_j^{\alpha} \sigma^2}{\ln(1 - \epsilon_j^{(l)})}. \tag{F.8}$$

Because the BS's goal is to spend (on average) as little power on a packet as possible while providing certain reliability guarantees, we define the optimization problem (P1) given at the top of the page, where $\Theta_j^{(l)} = \frac{\epsilon_{tar}}{\prod_{i=0}^{l-1} \epsilon_j^{(i)}}$

*Recursive, decomposed per user power optimization*

$$\Psi_j^{(l,L)}(\gamma_j^{(l)},\Theta_j^{(l)}) = \min_{\epsilon_j^{(l)}} \quad P_j^{(l)} + \int_0^{\gamma_j^{(l)}} f_e\left(x_l;\frac{P_j^{(l)}}{d_j^{\alpha}\sigma^2}\right)\Psi_j^{(l+1,L)}\left(\gamma_j^{(l+1)},\frac{\Theta_j^{(l)}}{\epsilon_j^{(l)}}\right)dx_l$$
$$\text{s.t.} \quad \prod_{i=l}^{L}\epsilon_j^{(i)} \leq \Theta_j^{(l)} \tag{P1}$$

is the remaining error budget resulting from the previous transmission attempts and the overall target is $\epsilon_{tar}$ (such as $10^{-5}$ in URLLC). The problem (P1) can be summarized as follows. For a given packet, currently at round $l$, BS needs to decide on its next error target $\epsilon_j^{(l)}$ that will minimize the expected power moving forwards. The objective (cost) is composed of two terms. First, the power $P_j^{(l)}$ spent in the immediate round, which is directly related to the chosen error target via (F.8). Second, the expected additional power that will be spent if the packet fails. Note that $\epsilon_j^{(l)}$ impacts the second part in two ways: it determines the remaining error budget $\Theta_j^{(l+1)}$ and, through $P_j^{(l)}$, the distribution of the SNR of the current transmission $x_l$ that affects the new residual SNR $\gamma_j^{(l+1)}$. Depending on the HARQ mode, the relationship between $x_l$ and $\gamma_j^{(l+1)}$ is captured by either (F.5) or (F.6). In general, the problem (P1) is difficult as it involves the recursive term $\Psi_j^{(l+1,L)}$ which contains $\Psi_j^{(l+2,L)}$, etc., that all require finding an optimal target. We will now consider the two special cases that arise when CC or IR is used.

### Chase Combining

In case of CC mode of HARQ, the optimization problem is greatly simplified, which is captured by the following lemma:

**Lemma 1.** *When using Chase Combining, the individual, per-stage error targets that minimize the expected power depend only on the remaining error budget $\Theta_j^{(l)}$. Due to the lack of dependency on other parameters, in particular rate $R$ and residual SNRs, the recursive problem* (P1) *becomes equivalent to* (F.9)*, where we use the convention that $\prod_{i=m}^{n} x_i = 1$ when $m > n$.*

The proof of Lemma 1 can be found in Appendix A. By minimizing $\Psi_j^{(0,L)}(2^R-1,\epsilon_{tar})$ using the definition (F.9) stated in the Lemma 1, the BS can determine all targets $\boldsymbol{\epsilon} = \left[\epsilon^{(0)},\ldots,\epsilon^{(L)}\right]$ in advance. While the analytical

$$\Psi_j^{(l,L)}(\gamma_j^{(l)},\Theta_j^{(l)}) = \min_{\epsilon_j^{(l)},\dots,\epsilon_j^{(L)}} \quad \left(-\gamma_j^{(l)} d_j^{\alpha}\sigma^2\right)\sum_{i=l}^{L}\frac{1}{\ln(1-\epsilon_j^{(i)})}\prod_{k=l}^{i-1}\frac{\ln(1-\epsilon_j^{(k)})+\epsilon_j^{(k)}}{\ln(1-\epsilon_j^{(k)})} \tag{F.9a}$$

$$\text{s.t.} \qquad \prod_{i=l}^{L}\epsilon_j^{(i)} = \Theta_j^{(l)} \tag{F.9b}$$

$$\Psi_j^{(L-1,L)}(\gamma_j^{(L-1)},\Theta_j^{(L-1)}) = \min_{\epsilon_j^{(L-1)}} \quad P_j^{(L-1)} + \int_0^{\gamma_j^{(L-1)}} f_e\left(x_L; \frac{P_j^{(L-1)}}{d_j^{\alpha}\sigma^2}\right)\left(-\frac{\frac{\gamma_j^{(L-1)}-x_L}{1+x_L}d_j^{\alpha}\sigma^2}{\ln(1-\epsilon_j^{(L)})}\right)dx_L \tag{F.10}$$

approach is not tractable, numerical solutions can be obtained rather easily. Moreover, since the final target error rate $\epsilon_{tar}$ and the maximum number of retransmissions $L$ are typically system-wide parameters with limited number of configurations, the sequence $\boldsymbol{\epsilon}$ do not require frequent updates and is identical for all UEs.

### Incremental Redundancy

When the IR-type HARQ is used, determining optimal error targets is much more complex. In general their values do depend on the current residual SNR and should be recomputed after each failed transmission. Consequently, it is not possible to simplify the problem in the same way as in CC and compute all targets at once for arbitrary $(l, L)$. When $l = L-1$, the problem can be turned into a univariate, unconstrained optimization (by merit of $\epsilon_j^{(L)} = \frac{\epsilon_{tar}}{\prod_{i=0}^{L-1}\epsilon_j^{(i)}}$) and reads as in (F.10) above, where the update to the residual SNR in incremental redundancy $\gamma_{ir_j}^{(l+1)} = \frac{\gamma_{ir_j}^{(l)}-\mathrm{SNR}_j^{(l)}}{1+\mathrm{SNR}_j^{(l)}}$ follows from the definition in (F.6). The integral doesn't have a closed form, however a relatively simple approximation can be obtained by substituting exponential function with its first-order Taylor expansion around 0 i.e. $e^{\frac{\ln(1-\epsilon_j^{(L-1)})x}{\gamma_j^{(L-1)}}} \approx \left(1+\frac{\ln(1-\epsilon_j^{(L-1)})}{\gamma_j^{(L-1)}}x\right)$.

The approximated objective function becomes then

$$
\begin{aligned}
&-\frac{\gamma_j^{(L-1)} d_j^{\alpha} \sigma^2}{\ln(1-\epsilon_j^{(L-1)})} + \frac{\ln(1-\epsilon_j^{(L-1)}) d_j^{\alpha} \sigma^2}{\gamma_j^{(L-1)} \ln(1-\epsilon_j^{(L)})} \Bigg( \ln(1-\epsilon_j^{(L-1)}) \\
&+ \frac{(\gamma_j^{(L-1)}+1)(\gamma_j^{(L-1)} - \ln(1-\epsilon_j^{(L-1)})) \ln(\gamma_j^{(L-1)}+1)}{\gamma_j^{(L-1)}} \\
&+ \frac{\gamma_j^{(L-1)}(\ln(1-\epsilon_j^{(L-1)})-2)}{2} \Bigg)
\end{aligned}
\tag{F.11}
$$

Since in this work we will consider only scenarios with at most $L = 2$ retransmissions (in line with the low latency requirement) we adopt the following approach:

1. For the few limited configurations characterized by transmission rate $R$ and $\epsilon_{tar}$ the optimal $\epsilon_j^{(0)}$, which is a solution to $\Psi_j^{(0,2)}\ (2^R - 1, \epsilon_{tar})$ as defined in (P1), is found through an exhaustive search (performed offline). To do this, we sweep through its possible values, fixing $\epsilon_j^{(0)}$, and then calculating the remaining expected power by solving and integrating (F.11) over a $[0, \gamma_j^{(0)}]$ range and with $\Theta_j^{(1)} = \frac{\epsilon_{tar}}{\epsilon_j^{(0)}}$. Note that since the initial $\gamma_j^{(0)} = 2^R - 1$ is identical for all UEs, so is the optimal $\epsilon_j^{(0)}$.

2. After the first transmission, users who failed will end up with different residual SNRs. For each of them, the optimal error target for the upcoming retransmission is obtained separately by minimizing (F.11). Since $\Psi_j^{(1,2)}\ (\gamma_j^{(1)}, \frac{\epsilon_{tar}}{\epsilon_j^{(0)}})$ is a univariate unconstrained problem with a closed form, it is relatively simple to obtain the solution numerically.

3. In case second retransmission is necessary $\epsilon_j^{(2)} = \frac{\epsilon_{tar}}{\epsilon_j^{(0)} \epsilon_j^{(1)}}$ as follows from the constraint.

## 3.2 NOMA-HARQ

As an enhancement of the OMA scheme we explore an approach in which the base station is allowed to schedule multiple UEs over the same channel resources, i.e. making the access non-orthogonal. This can be useful especially in two instances 1) when due to the inherent randomness of new packet arrivals combined with decoding errors the system enters a period

$$p_{er_j}^{(l)} = \begin{cases} \underbrace{1 - \left( \frac{S_j^{(l)}}{S_k^{(m)}\phi_k^{(m)} + S_j^{(l)}} + \frac{S_k^{(m)}}{S_j^{(l)}\phi_j^{(l)} + S_k^{(m)}} e^{-\sigma^2 \frac{\phi_j^{(l)}+1}{S_k^{(m)}}} \right) e^{-\frac{\sigma^2}{S_j^{(l)}}},}_{A} & \text{if } \phi_j^{(l)}\phi_k^{(m)} \geq 1 \\ A - \left( 1 - \frac{S_j^{(l)}}{S_k^{(m)}\phi_k^{(m)}+S_j^{(l)}} - \frac{S_k^{(m)}}{S_j^{(l)}\phi_j^{(l)}+S_k^{(m)}} \right) e^{-\frac{\sigma^2}{1-\phi_j^{(l)}\phi_k^{(m)}}\left( \frac{\phi_j^{(l)}+1}{S_k^{(m)}} + \frac{\phi_k^{(m)}+1}{S_j^{l)}} \right)}, & \text{if } \phi_j^{(l)}\phi_k^{(m)} < 1 \end{cases} \tag{F.13}$$

of congestion and is forced to queue packets 2) when the residual SNR of an unsuccessful packet is very low and assigning dedicated resources to a retransmission would be wasteful.

The usage of NOMA is facilitated by the introduction of SIC mechanism, which allows to remove the signal of the decodable user thus removing the interference it causes to the other UE. As such, the error probability consists of two terms originating from two mutually exclusive events: error probability with interferer's signal decoded and canceled, and the case when SIC is not successful. In our model we assume that the receiver performs optimal order decoding, i.e. if any of the two packets can be decoded in the presence of interference, the other one becomes interference-free.

Let the UE $j$, who is transmitting for the $l$-th time, share the TF-block with the UE $k$, who is currently at its $m$-th attempt, and let $Q_j^{(l)} = \frac{P_j^{(l)}\left|h_j^{(l)}\right|^2}{d_j^{\alpha}}$, $Q_k^{(m)} = \frac{P_k^{(m)}\left|h_k^{(m)}\right|^2}{d_k^{\alpha}}$ denote their received powers. Then

$$\begin{aligned} p_{er_j}^{(l)} =& \Pr\left\{ \frac{Q_j^{(l)}}{\sigma^2} < \gamma_j^{(l)}, \frac{Q_k^{(m)}}{Q_j^{(l)}\zeta_j + \sigma^2} > \gamma_k^{(m)} \right\} \\ &+ \Pr\left\{ \frac{Q_j^{(l)}}{Q_k^{(m)}\zeta_k + \sigma^2} < \gamma_j^{(l)}, \frac{Q_k^{(m)}}{Q_j^{(l)}\zeta_j + \sigma^2} < \gamma_k^{(m)} \right\} \end{aligned} \tag{F.12}$$

and the error probability of the second user of the TF-block $p_{er_k}^{(m)}$ is obtained by simply interchanging the indices $j \leftrightarrow k$ and $l \leftrightarrow m$. The coefficients $\zeta_j$ and $\zeta_k$ denote the interference reduction coefficients which will be explained later on. The expression (F.12) can be obtained in the closed form as given in (F.13), where $S_j^{(l)} = \frac{P_j^{(l)}}{\gamma_j^{(l)}d_j^{\alpha}}$ and $\phi_j^{(l)} = \gamma_j^{(l)}\zeta_j$. The derivation of (F.13) is discussed in Appendix B.

Recall that in the OMA case, the first step was to find the optimal error target $\epsilon_j^{(l)}$ which then could be plugged into (F.8) to determine the transmit power for the next transmission. In the NOMA setting $\epsilon_j^{(l)}$ and $\epsilon_k^{(m)}$ that minimize the expected power per packet of each respective user would have to be found jointly which is significantly more difficult. Due to its high computational complexity, in this work we will omit this process and instead use the same targets as for OMA. This can be further justified by analyzing the results and findings presented in [26] which show that the optimal error targets for OMA and NOMA are in fact very similar.

With $\epsilon_j^{(l)}, \epsilon_k^{(m)}$ given and the error probabilities defined as in (F.13), the transmit powers are assigned to users by solving the following optimization problem:

$$\underset{P_j^{(l)}, P_k^{(m)}}{\arg\min} \quad P_j^{(l)} + P_k^{(m)} \tag{F.14a}$$

$$\text{s.t.} \qquad p_{er_j}^{(l)} \leq \epsilon_j^{(l)}, \quad p_{er_k}^{(m)} \leq \epsilon_k^{(m)} \tag{F.14b}$$

The solution is found using an interior-point convex solver. While the constraint functions are not convex, the domain can be divided into two disjoint regions: $\frac{P_j^{(l)}}{d_j^\alpha} > \frac{P_k^{(m)}}{d_k^\alpha}$ and $\frac{P_j^{(l)}}{d_j^\alpha} < \frac{P_k^{(m)}}{d_k^\alpha}$. The global minimum is determined by finding the local minimum of each and selecting the lower one.

### Chase Combining

When using NOMA with Chase Combining additional assumption is required for the expression (F.12) to be valid. Note that the total SINR of a packet can be written as a simple sum of the SINRs of its individual copies only when the interference in each of them is uncorrelated. For that reason we ensure in our simulations that throughout its $L+1$ transmissions a packet is never paired more than once with the same packet of other user. This is rarely an issue and does not impact reliability, only the scheduling process explained later on.

In certain cases the procedure described in [26] can be refined by utilizing the previous copies of the interfering packet to partially cancel its contribution in the current transmission even before applying the SIC. This is reflected in (F.12) by the reduction coefficients $\zeta_j$ and $\zeta_k$. The details on how to obtain them can be found in Appendix C.

### Incremental Redundancy

In addition to CC-HARQ, we investigate the NOMA approach with IR. Since in IR each packet is composed of different symbols, it can be assumed that all of them experience independent interference. As a result, there is no need for the additional constraint on the scheduling that was required for CC. At the same time, since the additional interference reduction in CC was achieved by combining previous signals containing the same packet it is no longer possible to use this feature with IR [2] and $\zeta_j, \zeta_k = 1$.

# 4 Instantaneous CSI and finite blocklength

The preceding analysis pertained to the case of Rayleigh fading channel whose realizations are unknown until after the reception of the packet (through perfect estimation) and a priori only their distribution is known. However, due to the low end-to-end latency of the URLLC communication it is of interest to investigate also the case where the channel coherence time is large enough that the BS can treat the channel during subsequent uplink transmission as known. Unlike in statistical CSI case where the dominant source of errors is fading [27], here the finite blocklength effects become crucial. Since the channel effectively becomes AWGN, the decoding errors are caused solely by noise which is especially prominent in short packets. Hence, to study the case of instantaneous CSI we resort to finite blocklength analysis [28] and, for tractability reasons, limit the scope to just the case of IR-HARQ.

## 4.1 Finite blocklength OMA-HARQ

As previously, let us start with a simpler case of dedicated resources. The average information density contained in $l$-th transmission of the packet from user $j$ can be written as [28]

$$i_j^{(l)} = \frac{1}{K} \sum_{n=1}^{K} \left[ \ln \left( 1 + \frac{P_j^{(l)} \left| h_j^{(l)} \right|^2}{d_j^{\alpha} \sigma^2} \right) + \frac{\left| y_{j,n}^{(l)} \right|^2}{\frac{P_j^{(l)} \left| h_j^{(l)} \right|^2}{d_j^{\alpha}} + \sigma^2} - \frac{\left| y_{j,n}^{(l)} - \sqrt{\frac{P_j^{(l)}}{d_j^{\alpha}}} h_j^{(l)} x_{j,n}^{(l)} \right|^2}{\sigma^2} \right] \tag{F.15}$$

[2] An equivalent technique could be attempted with incremental redundancy, however the exact procedure and resulting gains are difficult to assess. To suppress the current interfering packet the previous (unsuccessful) packets would have to be soft-decoded and then re-encoded with mother code rate to "guess" the next corresponding symbols in the buffer.

$$\Psi_j^{(l,L)}\left(\left|h_j^{(l)}\right|^2,\mu_j^{(l-1)},\nu_j^{(l-1)}\right)=\min_{P_j^{(l)}}\; P_j^{(l)}+\epsilon_j^{(l)}\int_0^{\infty}e^{-z_{l+1}}\Psi_j^{(l+1,L)}\left(z_{l+1},\mu_j^{(l)},\nu_j^{(l)}\right)dz_{l+1} \tag{F.17a}$$

$$\text{s.t.}\quad F_{\mathcal{N}}\left(R\ln 2;\mu_j^{(L)},\nu_j^{(L)}\right)\leq\epsilon_{tar} \tag{F.17b}$$

---

where the sum involves all received and transmitted symbols $y_{j,n}^{(l)}$ and $x_{j,n}^{(l)}$ respectively. In (F.15) the difference of the last two terms is a Laplacian random variable with zero mean and variance equal to $\frac{2P_j^{(l)}\left|h_j^{(l)}\right|^2}{P_j^{(l)}\left|h_j^{(l)}\right|^2+d_j^{\alpha}\sigma^2}$. As shown in [28] a sum of $K$ such Laplacian random variables can be well approximated by a Gaussian random variable with zero mean and $K$ times higher variance. Hence, the average information density contained in a codeword of size $K$ follows:

$$\hat{i}_j^{(l)}\sim\mathcal{N}\left(\ln\left(1+\frac{P_j^{(l)}\left|h_j^{(l)}\right|^2}{d_j^{\alpha}\sigma^2}\right),\frac{2P_j^{(l)}\left|h_j^{(l)}\right|^2}{K\left(P_j^{(l)}\left|h_j^{(l)}\right|^2+d_j^{\alpha}\sigma^2\right)}\right)\quad. \tag{F.16}$$

Since the codewords in IR can be treated as independent, the total information density provided by $l$ subsequent transmissions is also a Gaussian random variable with mean $\mu_j^{(l)}=\sum_{i=0}^{l}\ln\left(1+\frac{P_j^{(i)}\left|h_j^{(i)}\right|^2}{d_j^{\alpha}\sigma^2}\right)$ and variance $\nu_j^{(l)}=\frac{1}{K}\sum_{i=0}^{l}\frac{2P_j^{(i)}\left|h_j^{(i)}\right|^2}{P_j^{(i)}\left|h_j^{(i)}\right|^2+d_j^{\alpha}\sigma^2}$.

Again, the ultimate goal is to minimize the expected total power per packet, however the optimization problem is considerably different. Since the immediate channel realization is known, it is clear that the optimal transmit power (and the corresponding optimal error probability) is a function of both the instantaneous channel gain and the statistics of the future channel realizations. Moreover, after the failed attempt, the receiver is not able to determine the exact residual information density[3] as it depends on the particular realizations of the noise which are unknown. To determine the transmit power for the packet at round $l$ belonging to user $j$, the BS needs to solve the recursive optimization problem which can be framed as in (F.17), where

[3] In fact, the residual information density $R\ln 2-\sum_{i=0}^{l}\hat{i}_j^{(i)}$ is a random variable following truncated Gaussian distribution restricted to $[0,R\ln 2]$.

$F_{\mathcal{N}}(\cdot;\cdot,\cdot)$ is the CDF of Gaussian distribution, $\epsilon_j^{(l)} = \frac{F_{\mathcal{N}}\left(R\ln 2;\mu_j^{(l)},\nu_j^{(l)}\right)}{F_{\mathcal{N}}\left(R\ln 2;\mu_j^{(l-1)},\nu_j^{(l-1)}\right)}$ is the failure probability at round $l$ and the integration is over the possible channel gains in the next uplink phase. Note that both $\epsilon_j^{(l)}$ and the recursive term $\Psi_j^{(l+1,L)}\left(z_{l+1},\mu_j^{(l)},\nu_j^{(l)}\right)$ depend on the latest $P_j^{(l)}$ and $\left|h_j^{(l)}\right|^2$ as they are included in $\mu_j^{(l)}$ and $\nu_j^{(l)}$.

As was the case earlier, the general closed form expression for the objective function (F.17a) is difficult to obtain. In the last attempt, i.e. $l = L$, the optimal power follows directly from the constraint (F.17b). More specifically, $P_j^{(L)} = \rho\frac{d_j^{\alpha}}{\left|h_j^{(L)}\right|^2}$, where $\rho$ is the solution to the equation

$$\frac{1}{2}\left(1+\operatorname{erf}\left(\frac{R\ln 2-\ln\left(1+\frac{\rho}{\sigma^2}\right)-\mu_j^{(L-1)}}{\sqrt{\frac{4\rho}{K(\rho+\sigma^2)}+\nu_j^{(L-1)}}}\right)\right)=\epsilon_{tar} \qquad \text{(F.18)}$$

with erf$(\cdot)$ being the error function. For $l = L-1$, the optimal power $P_j^{(L-1)}$ can be determined by minimizing the objective (F.17a) which in that case becomes

$$P_j^{(L-1)}+\frac{F_{\mathcal{N}}\left(R\ln 2;\mu_j^{(L-1)},\nu_j^{(L-1)}\right)}{F_{\mathcal{N}}\left(R\ln 2;\mu_j^{(L-2)},\nu_j^{(L-2)}\right)}\,\rho\int_0^{\infty}\frac{1}{z_L}e^{-z_L}dz_L. \qquad \text{(F.19)}$$

The integral, which represents the expected value of $1/\left|h_j^{(L)}\right|^2$ (inverse exponential distribution), does not converge. To remedy that, we assume that any packet which is in a deep enough fade during its last $L$-th transmission will be dropped. Since the overall target error rate is $\epsilon_{tar}$ we select $\epsilon_{drop} < \epsilon_{tar}$ and find the point through inverse CDF $\left|h_{fade}\right|^2 = F_e^{-1}\left(\epsilon_{drop};1\right)$ which will be the lower limit for the integration. Although minimization of (F.19) requires solving recursively (F.18) as well (choice of $P_j^{(L-1)}$ determines $\rho$), it can still be done quite efficiently numerically.

For $l \le L-2$ the optimization becomes even more complex as it adds another level of recursion and since in this work we consider at most $L = 2$ retransmissions, $\Psi_j^{(0,2)}$ can become a bottleneck. A better idea is to precompute $P_j^{(0)}$ as a function of $\left|h_j^{(0)}\right|^2$ offline which we show in Fig.F.2. In practice, during our simulations the following approach is used:

1. For the newly arrived packet experiencing $\left|h_j^{(0)}\right|^2$, the optimal power

$P_j^{(0)}$ is determined by interpolation on Fig. F.2(a).

2. If the packet fails during initial transmission, the optimal power $P_j^{(1)}$ is determined by minimizing (F.19). If the packet ended up in round 1 due to being postponed (e.g. because of the deep fade) then $P_j^{(1)}$ can be interpolated from Fig. F.2(b). Note that this is a consequence of $\Psi_j^{(0,1)}$ being equivalent to $\Psi_j^{(1,2)}$ with $P_j^{(0)} = 0$.

3. If the packet fails for the second time, then $P_j^{(2)}$ is obtained by solving (F.18).

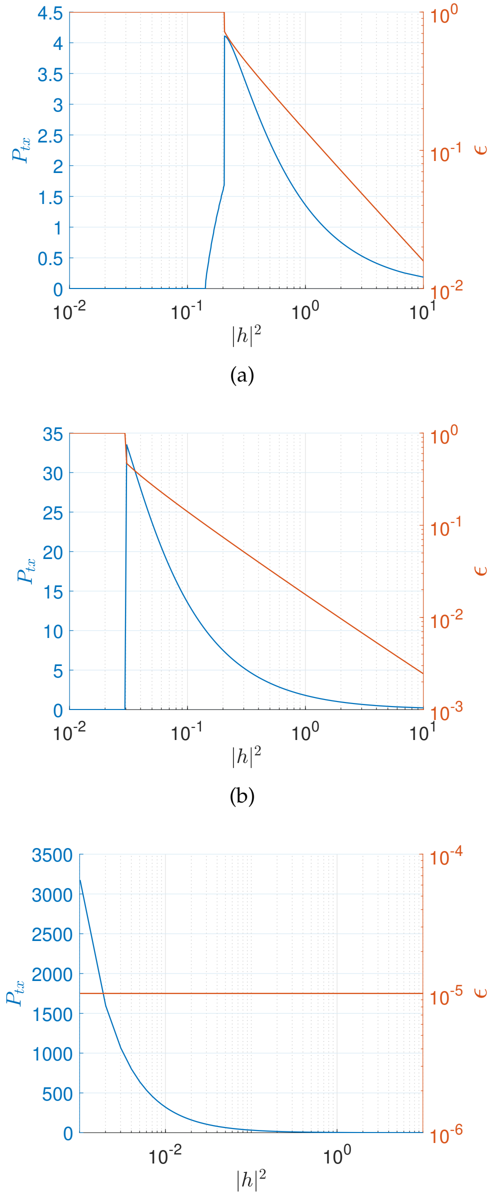
Lastly, we note that the optimal power obtained by solving (F.17) can be 0 (cf. Fig. F.2). This means the BS consciously chooses to postpone the packet based on unfavourable $\left|h_j^{(l)}\right|^2$.

## 4.2 Finite blocklength NOMA-HARQ

Lastly, we move on to discuss the application of NOMA-HARQ in the finite blocklength scenario. Let $P_j^{(l)}$ and $P_k^{(m)}$ be the transmit powers of the two packets which are to be scheduled in the same TF-block, $Q_j^{(l)}$ and $Q_k^{(m)}$ denote the respective received powers defined similarly as in section 3.2, and let the packet from user $j$ be attempted to decode first[4]. The resulting error probabilities are

$$p_{er_j}^{(l)} = F_{\mathcal{N}}\left(R\ln 2\,;\, \ln\left(1 + \frac{Q_j^{(l)}}{Q_k^{(m)} + \sigma^2}\right) + \mu_j^{(l-1)},\right. \\ \left. \frac{2Q_j^{(l)}}{K\left(Q_j^{(l)} + Q_k^{(m)} + \sigma^2\right)} + \nu_j^{(l-1)}\right) Z_j^{(l-1)} \tag{F.20}$$

[4] Accounting for optimal SIC ordering becomes relevant when there is an uncertainty about the relationship between two received powers. Clearly this is not the case when channel realizations are known in advance.



(a)

(b)

(c)

**Fig. F.2:** The optimal transmit power and resulting error probability as a function of current channel gain realization for $R = 1$. The values are normalized to $\sigma^2 = 1$ and $d^\alpha = 1$. The three figures correspond to the power used during initial transmission assuming (a) $L = 2$, (b) $L = 1$, (c) $L = 0$ allowed retransmissions.

$$
\begin{aligned}
p_{er_k}^{(m)} = \left(1 - p_{er_j}^{(l)}\right) F_{\mathcal{N}} \Bigg( & R \ln 2 \,; \ln\left(1 + \frac{Q_k^{(m)}}{\sigma^2}\right) + \mu_k^{(m-1)}, \\
& \frac{2Q_k^{(m)}}{K\left(Q_k^{(m)} + \sigma^2\right)} + \nu_k^{(m-1)} \Bigg) Z_k^{(m-1)} \\
+ p_{er_j}^{(l)} F_{\mathcal{N}} \Bigg( & R \ln 2 \,; \ln\left(1 + \frac{Q_k^{(m)}}{Q_j^{(l)} + \sigma^2}\right) + \mu_k^{(m-1)}, \\
& \frac{2Q_k^{(m)}}{K\left(Q_k^{(m)} + Q_j^{(l)} + \sigma^2\right)} + \nu_k^{(m-1)} \Bigg) Z_k^{(m-1)}
\end{aligned}
\tag{F.21}
$$

where $Z_j^{(l)} = \frac{1}{F_{\mathcal{N}}\left(R \ln 2; \mu_j^{(l)}, \nu_j^{(l)}\right)}$. We make note here of the slight abuse of our usage of the terms $\mu$ and $\nu$. These are meant to represent the means and variances of the information density obtained in earlier rounds and are clearly a function of the SINR of the signal of interest. Since in the NOMA approach earlier replicas of the packet could have also been scheduled non-orthogonally one should remember to include the appropriate interference terms when calculating $\mu$ and $\nu$.

To find the appropriate powers for the two UEs we follow similar heuristic as before. Let $P_j^{(l)^\star}$ and $P_k^{(m)^\star}$ be the optimal OMA powers of users $j$ and $k$ for the upcoming round as given by (F.17). Consequently, in the OMA setting, these powers would result in the error probabilities $\epsilon_j^{(l)^\star}$ and $\epsilon_k^{(m)^\star}$ respectively. Then, the goal is to find appropriate $P_j^{(l)}$ and $P_k^{(m)}$ for the NOMA transmissions such that the individual OMA error targets are met:

$$
\underset{P_j^{(l)}, P_k^{(m)}}{\arg\min} \quad P_j^{(l)} + P_k^{(m)} \tag{F.22a}
$$

$$
\text{s.t.} \quad p_{er_j}^{(l)} \leq \epsilon_j^{(l)^\star}, \quad p_{er_k}^{(m)} \leq \epsilon_k^{(m)^\star} \tag{F.22b}
$$

The rationale behind using the same error targets for NOMA as for OMA is again the tractability of the problem. Finding even a single pair of optimal NOMA targets has high computational complexity, and in the instantaneous CSI case they would have to be computed for each pair of values $\left(\left|h_j^{(l)}\right|^2, \left|h_k^{(m)}\right|^2\right)$. The solution to (F.22) is found numerically.

# 5 Scheduling

The last element missing before we move on to the results is the matter of scheduling. As already mentioned, in this work we consider a system having a finite amount of resources, namely $W$ TF-blocks. These might not be enough to accommodate all the packets of the active users which inevitably leads to queuing and requires defining a scheduling policy. Because of the complexity of the problem whose optimal solution would require taking into account both current packets and future arrivals and since scheduling is not the primary topic of this work, we decide to settle for a heuristic approach which will be now described.

## 5.1 OMA scheduling

When the total number of packets in the system $T$ is lower than the amount of resources $W$ the scheduling decision is straightforward. Furthermore, in the instantaneous CSI case, the BS can already at this point decide that some of the packets should be postponed based on their poor channel conditions. We also remark that there is no limit regarding how many of them a single UE can send in one UL phase as long as there are available resources and is instructed to do so by the BS[5].

In case the number of packets $T$ exceeds $W$, the BS performs an intermediate step and decides which of them should be postponed. To do this we adopt the following procedure:

1. The priority is given to the packets currently in their last $L$-th round. If the number $T_C$ of such critical packets exceeds $W$ the ones which require the least power are transmitted and the remaining are dropped. All the non-critical $T - T_C$ packets are postponed, i.e. they are moved to the next round. Note that dropping the packets is the last resort since it compromises the overall reliability.

2. If $T_C < W$, the remaining TF-blocks are used to transmit some of the non-critical packets. For each of them BS calculates two values: current expected OMA power $\Psi_j^{(l,L)}$, and the expected power assuming this round was skipped $\Psi_j^{(l+1,L)}|P_j^{(l)} = 0$. Note that since $P_j^{(l)} = 0 \implies \epsilon_j^{(l)} = 1$ the second value entails more aggressive future error targets that will account for the lower number of retransmission opportunities. The packets that will be scheduled are those with the highest

[5] Because at each step user generates a new packet with probability $b$ and, since the allowed number of retransmissions is $L$, each user can be storing in its buffer up to $L + 1$ packets at any given time (each at a different round).

difference $(\Psi_j^{(l+1,L)}|P_j^{(l)} = 0) - \Psi_j^{(l,L)}$. The rationale is to prioritize the packets which are the most "expensive" to postpone. Depending on the CSI scenario, the expected powers are obtained based on either (P1) or (F.17).

Once the set of packets that will be sent is established, their optimal transmit powers are determined as outlined in the appropriate section (statistical/instantaneous CSI, CC/IR).

## 5.2 NOMA scheduling

The overall procedure for deciding which packets to postpone is similar to the OMA case with the following caveats:

1. Since pairing allows to accommodate twice as many packets, queuing starts only when $T > 2W$. Again, the priority is given to the packets at round $L$ and if $T_C > 2W$ the ones requiring highest power are dropped.
2. When deciding which of the non-critical packets to transmit immediately and which to postpone the procedure is identical as for OMA, i.e. the calculation of the expected powers is also based on the equations derived for OMA. While this is a suboptimal approach, it is not clear how to compare the current and future NOMA powers since that would require the a priori knowledge of which users will be active next and what will be the exact pairing now and in the future. Instead, we resort to a simple heuristic that if the packet is "expensive" to postpone in OMA terms, then it is also the case for NOMA, especially since the latter always require some extra power.

The next step is to determine the pairing. Let us denote the number of pairs to be created as $q$. When $T \geq 2W$ then $q = W$, however when $T < 2W$ we can consider two cases for our simulations. a) Power conservative (PC) approach which will form pairs only if necessary, i.e. when $T > W$, leading to $q = \max(T - W, 0)$. Consequently the usage of resources is maximized. b) Resource conservative (RC) approach where as many pairs as possible are made resulting in $q = \left\lfloor \frac{T}{2} \right\rfloor$. The usage of resources is minimized at the cost of higher power.

Once it is decided which packets will be transmitted and how many pairs are needed, the appropriate matching of the users is determined:

1. First, for each pair of packets we calculate the optimal NOMA powers according to either (F.14) or (F.22). Then, we compare them with the optimal OMA powers of each of the user to determine the difference $\left(P_{j,NOMA}^{(l)} + P_{k,NOMA}^{(m)}\right) - P_{j,OMA}^{(l)} - P_{k,OMA}^{(m)}$, which is the extra cost of

scheduling the two packets together rather than on dedicated resources. Note that some of the pairs cannot be formed. The possible reasons include: earlier joint transmission (only applicable to CC-HARQ), optimization (F.14) or (F.22) did not converge or the two packets belong to the same UE.

2. Once the costs for all pairs are known, $q$ pairs that produce the lowest combined cost are selected. This step is a variation of the maximum weight matching problem which can be solved by Blossom algorithm [29].

### 5.3 Complexity

In general, the most costly operation in terms of complexity is the numerical optimization that determines the transmit powers for a pair of UEs (eq. (15) for Statistical CSI case and eq. (23) for the Instantaneous one). The main issue is that in order to find the optimal scheduling, we have to perform that operation for every possible pair of packets, so that we can later on select the subset of pairs which produces the lowest combined power. This requires $t(t+1)/2$ independent optimizations, where $t = \min(T, 2W)$ is the number of considered packets (note that it cannot exceed $2W$, since the surplus is postponed). Interestingly, the next step (selecting the best pairs using maximum weight matching algorithm) is not a bottleneck, even though it has a higher order of complexity $\mathcal{O}(t^3)$ [29]. This is because in our case $t$ is relatively small so it becomes a matter of $\mathcal{O}(t^3)$ basic operations as opposed to $\mathcal{O}(t^2)$ optimizations.

The other auxiliary procedures, such as finding the optimal error targets are not contributing significantly to the overall complexity. Moreover, the initial error targets, whose computation is the most time consuming, can always be determined offline. However, if the latency and hence complexity is the main bottleneck, then the Chase Combining variant might be preferable as it allows to compute all targets offline.

Lastly, we remark, that our goal is not so much to present an algorithm fully ready for practical implementation, but instead to analyze the potential of the NOMA-HARQ concept when the optimal powers are obtained. With this potential being proven in the optimal case, we expect that one can relatively easily devise suboptimal solutions that achieve close-to-optimal performance while having feasible computational requirements.

## 6 Results

The parameters used for simulations are gathered in Table F.1. The power of noise is given per symbol and is calculated as $\sigma^2 = N_0 + 10\log_{10} B_w$,

**Table F.1:** Simulation parameters

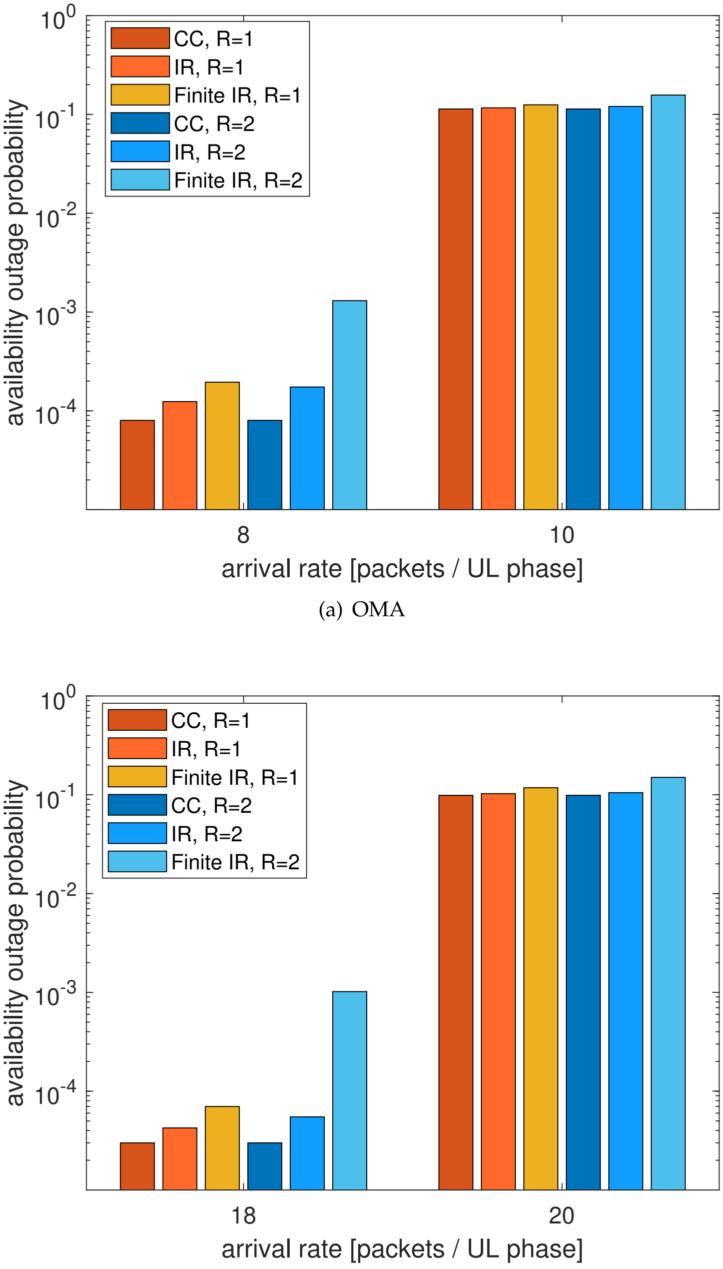
| | |
|---|---|
| Number of UEs $N$ | 40 |
| Number of TF-blocks $W$ | 10 |
| Number of retransmissions $L$ | 2 |
| Min. distance $D_{min}$ | 20 m |
| Max. distance $D_{max}$ | 120 m |
| Pathloss exponent $\alpha$ | 2 |
| Noise power $\sigma^2$ | $-129.1$ dBm |
| Number of symbols $K$ | 50 |
| Final BLER $\epsilon_{tar}$ | $10^{-5}$ |
| Deep fade threshold $\epsilon_{drop}$ | $10^{-6}$ |
| Activation probability $b$ | $b \in [0.05, 0.5]$ |
| Transmission rate $R$ | $R \in [0.5, 2.5]$ bits/symbol |
| Channel Type | Rayleigh block fading |
| Channel estimation method | Perfect |

**Table F.2:** Optimal error targets in statistical CSI HARQ and $\epsilon_{tar} = 10^{-5}$

| | |
|---|---|
| CC | $\epsilon^{(0)} = 0.189$, $\epsilon^{(1)} = 0.0374$, $\epsilon^{(2)} = 0.0014$ |
| IR, $R = 0.5$ | $\epsilon^{(0)} = 0.2$ |
| IR, $R = 1.0$ | $\epsilon^{(0)} = 0.215$ |
| IR, $R = 1.5$ | $\epsilon^{(0)} = 0.2255$ |
| IR, $R = 2.0$ | $\epsilon^{(0)} = 0.2485$ |
| IR, $R = 2.5$ | $\epsilon^{(0)} = 0.262$ |

where $N_0 = -173.9$dBm/Hz is a typical power spectral density at $298K$ and $B_w = 30$kHz was chosen as the symbol bandwidth. The distances between the BS and individual UEs are drawn from the uniform distribution[6], i.e. $d_j \sim U(D_{min}, D_{max})$. Throughout this section and in the legends of the figures we will refer to the Chase combining and incremental redundancy HARQ utilizing statistical CSI knowledge as CC and IR respectively, while the IR-HARQ with instantaneous CSI and finite blocklength as Finite IR. Furthermore, we will distinguish three access methods: OMA, power conserva-

[6] It could be claimed, that the user positions being uniformly distributed over a two-dimensional disk would be more realistic than having a uniformly distributed distance to the BS. Simulation results not reported here for space constraints show, however, that in both cases the same relative trends can be observed. The average transmit power is higher in the 2-D case (due to the higher number of UEs that are far), but the offset is essentially flat.



(a) OMA

(b) NOMA

**Fig. F.3:** The availability outage probability for (a) OMA and (b) NOMA.

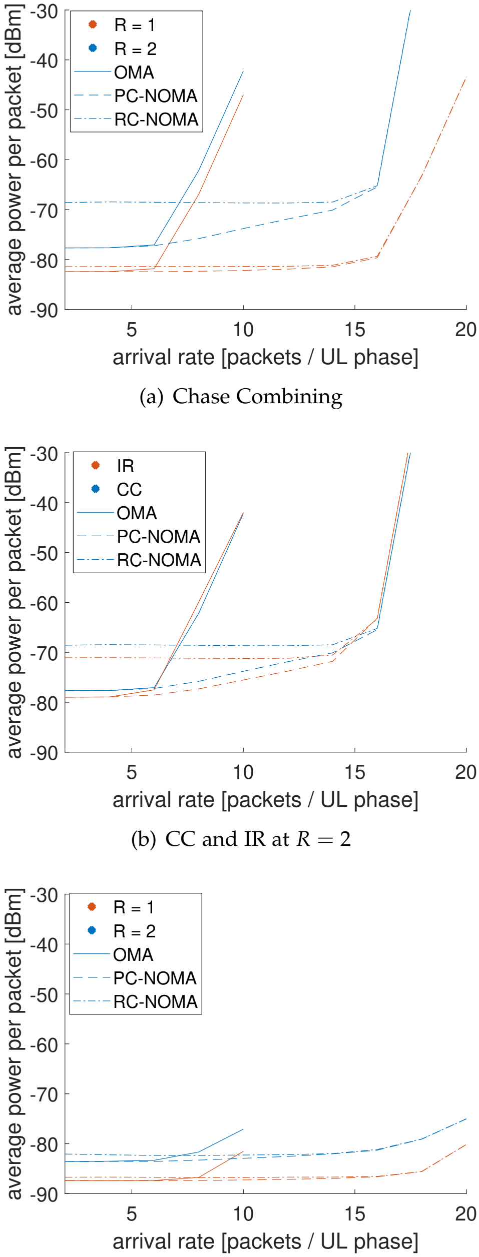
tive NOMA (PC-NOMA) and resource conservative NOMA (RC-NOMA).

In Table F.2 we provide the optimal error targets calculated for IR and CC. In general, the initial error target $\epsilon^{(0)}$ in IR is more relaxed and, unlike in CC, it increases with the transmission rate. Consequently, when using IR, the higher the rate, the more the system will rely on retransmissions to achieve the required reliability.

Let us start by looking into the most fundamental difference between OMA and NOMA which is reflected in their availability outage performance. We define availability outage as the state in which BS is forced to drop packets (i.e. timeslots where $T_C > 2W$). This way we make a clear distinction between availability and reliability similarly to [30]. Note that all packets which are not dropped have their reliability requirements fulfilled, as this is ensured by the power optimization and selection step. Fig. F.3 depicts the availability of OMA and NOMA system as a function of the mean number of new packets per uplink phase $bN$. For arrival rates which are below the shown values ($bN < 8$ for OMA and $bN < 18$ for NOMA) the availability outage probability becomes much lower than the transmission outage probability of $10^{-5}$. Conversely, arrival rates higher than $W$ and $2W$ result in an unstable system. In terms of availability PC-NOMA and its RC variant perform almost identically, hence, for brevity, only the former is presented. This is due to the fact that availability becomes an issue only as the mean number of arrivals approaches the system bandwidth, at which point PC and RC methods become equivalent since $T \geq 2W$ most of the time[7]. In this example the introduction of NOMA allows to support URLLC traffic of more than two times higher intensity compared to the baseline OMA. For a given arrival rate, the differences in availability outage between the three methods are a consequence of their distinct error targets for the initial transmission $\epsilon_j^{(0)}$, which are most demanding for CC, and least for Finite IR. Furthermore, they also increase with rate $R$ (except for CC). Since retransmissions add up to an already high number of new packets, when using CC the probability of driving the system into availability outage is lowest.

In Fig. F.4 the average power spent per packet (i.e. including retransmissions) as a function of arrival rate is investigated in different configurations. Note that in these and other figures the results for OMA are only shown until $bN = 10$, since at higher intensities the system is in a state of almost permanent availability outage. In Fig. F.4(a) the mode used is CC while the two sets of curves (red and blue) correspond to different transmission rates $R$. For very low arrival rates ($bN \in [2, 4]$) OMA and PC-NOMA are equivalent. As the arrival rate increases, the PC-NOMA approach quickly becomes much more efficient than the baseline scheme. This leads to one of the main

[7] Note that the total number of packets $T$ is a sum of new arrivals, postponed packets and those that failed previous transmission.



(a) Chase Combining

(b) CC and IR at $R = 2$

(c) Finite IR

**Fig. F.4:** Average power spent per packet. Figures (a) and (b) correspond to Statistical CSI case, while (c) to Instantaneous CSI.

takeaways of this work: in a latency-constrained system with high reliability requirements, the largest power penalty comes from the necessity to queue the packets. While scheduling them in a non-orthogonal way introduces penalty of its own, it is in fact less detrimental than having to make up for the lost transmission opportunities with more aggressive error targets. By comparing the difference between PC-NOMA (dashed) and RC-NOMA (dot-dashed) we can see that this is especially the case for low transmission rates $R$ (1dB of difference between red set of curves and 9dB for $R = 2$).

In Fig. F.4(b) CC (blue) and IR (red) at $R = 2$ are compared. Application of the latter method allows to further improve the performance by lowering the average power by  1.5dB in case of OMA/PC-NOMA and  2.5dB with RC-NOMA. We note that towards higher arrival rates CC gains an upper hand over IR since its slightly lower initial error targets make it less likely to queue the packets. Although the difference is minor, it reveals that obtaining a truly optimal solution would require adapting the error targets based on the current state of the buffer $T$ and knowledge of the arrival rate as well[8].

Lastly, Fig. F.4(c) depicts the results corresponding to the finite block-length scenario with known channel. The availability of instantaneous CSI allows to greatly decrease the mean power compared to the statistical CSI case. In the low to moderate traffic range ($bN \leq 14$) savings reach 4.5dB at $R = 1$ and 11dB at $R = 2$. Furthermore, as the arrival rate grows the increase in required power is much slower in the Finite IR case than for the statistical CSI counterparts.

In Fig. F.5 we investigate in more detail the average power per packet metric by looking at the performance of users grouped in different zones around the BS. As an example we take the Chase Combining case at $R = 2$ and low, medium and high arrival rate ($bN = [2, 8, 18]$). Most notably, as the intensity of traffic increases, the burden is shifted to the users close to the BS. The reason is twofold. The first cause is again related to queuing which typically introduces lower penalty for UEs closer to the BS[9]. Another cause is specific to NOMA, which in order to work requires that one packet has higher received power than the other. Since raising the power of UEs that are close is cheaper, typically they will be the ones asked to boost it (this behavior can be observed for RC-NOMA from the beginning). Moreover, in a PC-NOMA at low to moderate arrival rates, only few pairs are needed so they are often created among UEs positioned closer to the BS, while the furthest users are assigned the remaining TF-blocks in an orthogonal manner. Similar effects as

[8] However, as noted the room for improvement is not large and would add significant complexity to an already difficult problem. Last but not least, the information about the arrival rate in many scenarios might not be readily available.

[9] As described in Section 5.1 the process of deciding which UEs to postpone is slightly more complex and ultimately depends also on the residual SNR/information density and remaining error target. Nevertheless, packets from UEs which are positioned further away are less likely to be queued.

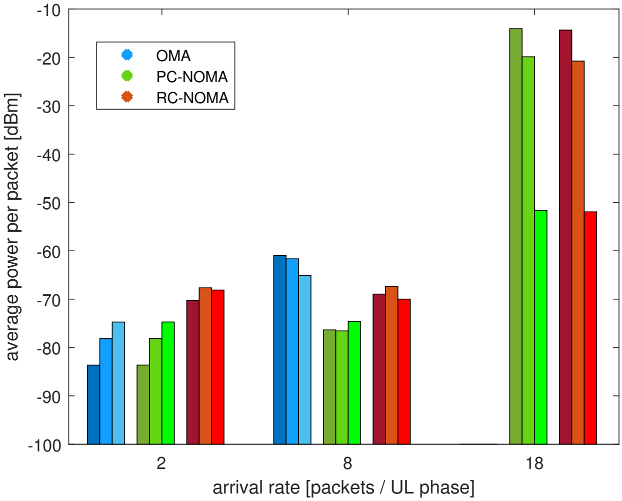
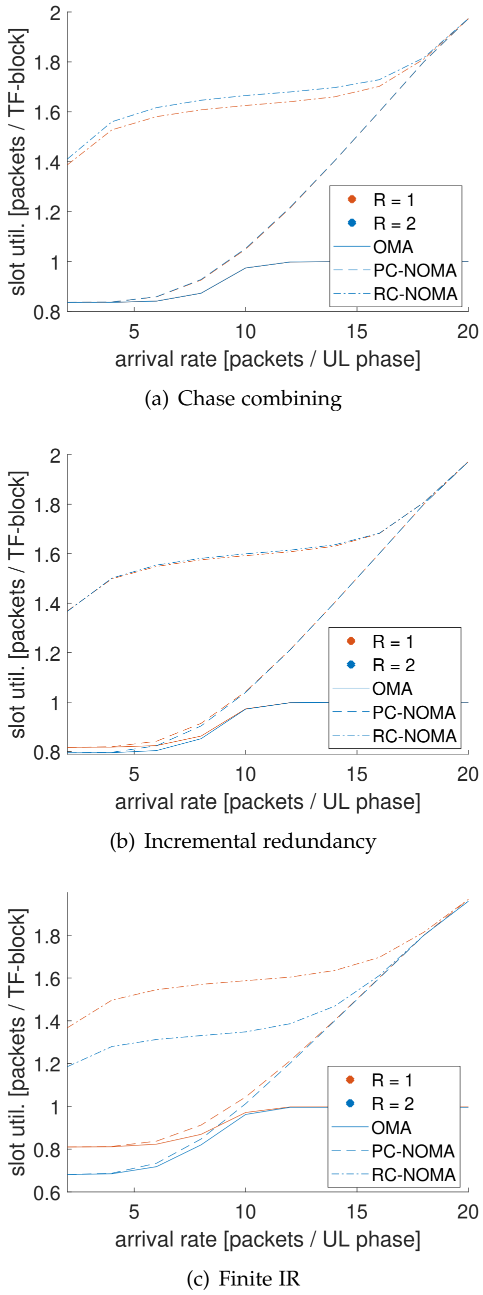


**Fig. F.5:** The average power per packet divided by zones. The three consecutive columns with different shades of the same color correspond to the average power per packet in a close (dark), middle and far (bright) zone around the BS. For the scenario considered here these are 20-53,33 meters, 53,33 - 86,66 meters, 86,66 - 120 meters.

those described have been observed also for lower transmission rates $R$ and in finite blocklength scenarios. Furthermore, the conclusions drawn in this work remain valid for other cell sizes.

Another set of results is provided in Fig. F.6. We define the slot utilization as the total number of successfully decoded packets from all UEs divided by the total number of used TF-blocks. The dependency of slot utilization on retransmission mode IR/CC and rate follow the same discussion as earlier for Fig. F.3. The higher the initial error targets, the more retransmissions are needed thus degrading the performance. Between PC-NOMA and RC-NOMA, the more aggressive pairing strategy can clearly offer significant gains. The reader is encouraged to analyze this especially in conjunction with Fig.F.4. Observe that for low rate $R = 1$ and low-to-medium traffic RC-NOMA almost doubles the resource efficiency of PC-NOMA with very little penalty to the average power (around 1dB). For higher transmission rates the increase in average power is more significant so the choice between PC and RC variant becomes a matter of trade-off.

Lastly, in Fig. F.7 we fix the average arrival rate of new packets to $bN = 8$ and instead vary the transmission rate $R$. The spectral efficiency presented in F.7(b) is obtained as the product of slot utilization and $R$. The noticeable jump in power of RC-NOMA with statistical CSI above $R > 1$ is in line with the observations first made in [26]. This behavior can be explained by inspecting the result (F.13), which contains a special term that decreases

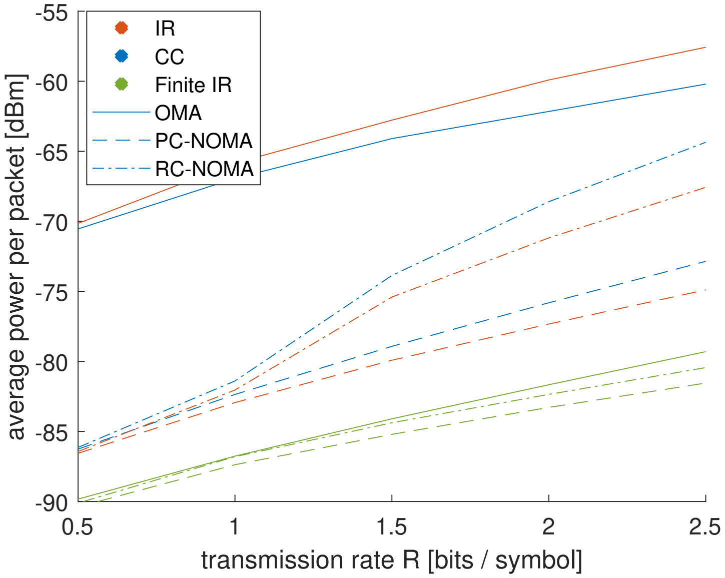


(a) Chase combining

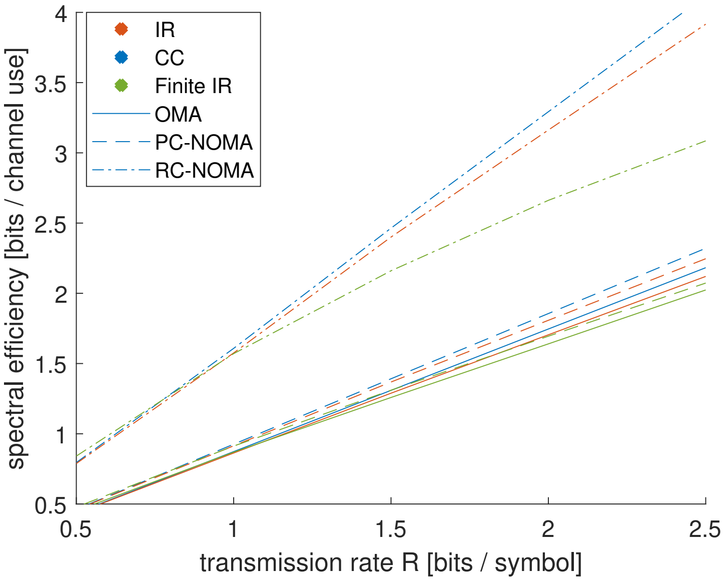
(b) Incremental redundancy

(c) Finite IR

**Fig. F.6:** Slot utilization of the studied access methods as a function of the arrival rate. Figures (a) and (b) correspond to Statistical CSI case, while (c) to Instantaneous CSI.



(a)



(b)

**Fig. F.7:** Average power per packet and spectral efficiency as a function of transmission rate at $bN = 8$ [packets / UL phase].

the error probability whenever $\gamma_j^{(l)}\gamma_k^{(m)}\zeta_j\zeta_k < 1$. Since $\gamma_j^{(0)} = 2^R - 1$ and $a < b \iff \gamma_j^{(a)} \geq \gamma_j^{(b)}$, then the condition $\gamma_j^{(l)}\gamma_k^{(m)}\zeta_j\zeta_k < 1$ is always true for $R \leq 1$. The similar jump in PC-NOMA is not observed at this arrival rate due to the fact that with $bN$ only equal to 8, pairs are still relatively infrequent. Moreover, most of the time pairing between two new packets can be avoided. Instead, it is possible to transmit them on dedicated slots, while only the ones with $\gamma_j^{(l)}, \gamma_k^{(m)} < 2^R - 1$ are combined so that $\gamma_j^{(l)}\gamma_k^{(m)}\zeta_j\zeta_k < 1$.

# 7 Conclusions

In this work we have proposed and investigated the performance of the system which combines NOMA and HARQ mechanisms to efficiently serve uplink URLLC traffic. Two distinct scenarios were discussed: one where only statistical CSI is available, and another where additionally also the instantaneous channel realizations are known. In each case we have defined an optimization problem that aims to minimize the average power spent per packet under a given latency (reflected by the maximum number of retransmissions) and reliability constraint. The schemes were evaluated in a multi-user scenario with fixed amount of channel resources and varying traffic intensity to investigate the impact of queuing on the overall reliability, power and resource efficiency. Our findings show that the introduction of NOMA is especially promising in two cases. First (RC-NOMA), the technique can be used to increase the total capacity of the system up to two times at a low-to-moderate cost in terms of power. Second (PC-NOMA), it can be implemented as an emergency mechanism in situations where due to higher traffic demand using traditional OMA would lead to prohibitively high power or even complete availability outage. The latter case is especially interesting as it shows that, in a latency-constrained system with given reliability requirements, the typical power penalty associated with NOMA is significantly smaller than the one arising from queuing the packets. Lastly, by investigating each scheme in two CSI cases, we provide some insights into the bounds on achievable performance in practical scenarios. Especially prominent is how the availability of instantaneous CSI can greatly reduce the transmit power needed for achieving the reliability targets.

# A Appendix

## Proof of Lemma 1

*Proof.* The proof is split into two parts. The first claim is proven by induction as follows.

### The induction step

Assume that there exists a certain round $l$ where the optimal error $\epsilon_j^{(l)}$ minimizing the average power $\Psi_j^{(l)}(\gamma_j^{(l)}, \Theta_j^{(l)})$ depends only on the remaining final error target $\Theta_j^{(l)}$, such that $\Psi_j^{(l)}(\gamma_j^{(l)}, \Theta_j^{(l)}) = \gamma_j^{(l)} d_j^{\alpha} \sigma^2 \widetilde{\Psi}_j^{(l)}(\Theta_j^{(l)})$. If this is the case, then the optimization problem at an earlier round $l-1$ becomes

$$\begin{aligned} \underset{\epsilon_j^{(l-1)}}{\arg\min} \quad & -\frac{\gamma_j^{(l-1)} d_j^{\alpha} \sigma^2}{\ln(1-\epsilon_j^{(l-1)})} \\ & + \int_0^{\gamma_j^{(l-1)}} f_e\left(x; \frac{P_j^{(l-1)}}{d_j^{\alpha}\sigma^2}\right) (\gamma_j^{(l-1)} - x) d_j^{\alpha} \sigma^2 \widetilde{\Psi}_j^{(l)} \left(\frac{\Theta_j^{(l-1)}}{\epsilon_j^{(l-1)}}\right) dx \end{aligned} \tag{F.23}$$

where the update $\gamma_j^{(l)} = \gamma_j^{(l-1)} - SNR_j^{(l-1)}$ is specific to CC and follows from (F.5). The objective function, which requires only simple integration can be obtained in the closed form

$$-\gamma_j^{(l-1)} d_j^{\alpha} \sigma^2 \left(\frac{1}{\ln(1-\epsilon_j^{(l-1)})} - \frac{\ln(1-\epsilon_j^{(l-1)}) + \epsilon_j^{(l-1)}}{\ln(1-\epsilon_j^{(l-1)})} \widetilde{\Psi}_j^{(l)} \left(\frac{\Theta_j^{(l-1)}}{\epsilon_j^{(l-1)}}\right)\right) \tag{F.24}$$

It is clear from the expression (F.24) which has a form $af(x)$, that the $\epsilon_j^{(l-1)}$ which minimizes it depends only on $\Theta_j^{(l-1)}$.

### The basis step

Since $\epsilon_j^{(L)} = \Theta_j^{(L)} = \frac{\epsilon_{tar}}{\prod_{i=0}^{L-1} \epsilon_j^{(i)}}$ used in the last possible transmission is fully determined by earlier attempts, the first non-trivial term corresponds to $\Psi_j^{(L-1)}(\gamma_j^{(L-1)}, \Theta_j^{(L-1)})$. The objective function there, which we denote for short $P_j^{avg}$, reads

$$\begin{aligned} P_j^{avg} &= P_j^{(L-1)} + \int_0^{\gamma_j^{(L-1)}} f_e\left(x_{L-1}; \frac{P_j^{(L-1)}}{d_j^{\alpha}\sigma^2}\right) \left(-\frac{(\gamma_j^{(L-1)} - x_{L-1}) d_j^{\alpha} \sigma^2}{\ln(1-\epsilon_j^{(L)})}\right) dx_{L-1} \\ &= \left(-\gamma_j^{(L-1)} d_j^{\alpha} \sigma^2\right) \left(\frac{1}{\ln(1-\epsilon_j^{(L-1)})} + \frac{\ln(1-\epsilon_j^{(L-1)}) + \epsilon_j^{(L-1)}}{\ln(1-\epsilon_j^{(L)}) \ln(1-\epsilon_j^{(L-1)})}\right) \end{aligned} \tag{F.25}$$

While solving $\frac{dP_j^{avg}}{d\epsilon_j^{(L-1)}} = 0$ requires numerical method it is again clear that the result is independent of $\gamma_j^{(L-1)}$, $d_j^\alpha$ or $\sigma^2$.

Applying the induction to the basis step proves sequentially that in all rounds $L-1, \ldots, 1, 0$ the optimal error target depends only on the current error budget. As for the second claim of the lemma, notice that when the optimal error targets do not depend on the residual SNRs, it means that for each round $l$ they must have a single, well-defined value, which can be computed in advance. This is because fixing $\epsilon_j^{(l)}$ leads to a chain of uniquely determined values $\epsilon_j^{(l)} \rightarrow \frac{\Theta_j^{(l)}}{\epsilon_j^{(l)}} \xrightarrow{opt} \epsilon_j^{(l+1)} \rightarrow \frac{\Theta_j^{(l)}}{\epsilon_j^{(l)}\epsilon_j^{(l+1)}} \xrightarrow{opt} \ldots \xrightarrow{opt} \epsilon_j^{(L)}$. By writing the problem (P1) in its explicit form and using the fact that error targets do not depend on the residual SNRs and hence on the variables of integration it is possible to eventually arrive at (F.9). The derivation is relatively simple albeit quite tedious. Although calculations involve multiple nested integrals, all integrands are of the form either $ae^x$ or $axe^x$ and display a regular structure. □

# B Appendix

Here, we will show the derivation of (F.13) from (F.12). First, let us shorten the notation by introducing following quantities: $X \sim f_e(x; s)$ where $s = \frac{P_j^{(l)}}{d_j^\alpha}$ is the exponentially distributed received power from user $j$ and similarly $Y \sim f_e(y; p)$ where $p = \frac{P_k^{(m)}}{d_k^\alpha}$ corresponds to user $k$. Also, since only a single packet from each user is considered we can drop the superscripts $(l)$ and $(m)$ moving forward. The first probability component in (F.12) now reads:

$$\Pr\left\{\frac{X}{\sigma^2} < \gamma_j, \frac{Y}{X\zeta_j + \sigma^2} > \gamma_k\right\} = \int_0^{\gamma_j\sigma^2} \left(\int_{\gamma_k(x\zeta_j+\sigma^2)}^{\infty} \frac{1}{p} e^{-\frac{y}{p}} dy\right) \frac{1}{s} e^{-\frac{x}{s}} dx \tag{F.26}$$

while the second term

$$\Pr\left\{\frac{X}{Y\zeta_k + \sigma^2} < \gamma_j, \frac{Y}{X\zeta_j + \sigma^2} < \gamma_k\right\} = \int_0^{\infty} \left(\int_{\frac{x}{\gamma_j\zeta_k} - \frac{\sigma^2}{\zeta_k}}^{\gamma_k(x\zeta_j+\sigma^2)} \frac{1}{p} e^{-\frac{y}{p}} dy\right) \frac{1}{s} e^{-\frac{x}{s}} dx \tag{F.27}$$

Notice that when $x < \gamma_j\sigma^2$, the lower limit of the inner integral in (F.27) is negative and outside of the support of the exponential distribution. Hence we can write (F.27) instead as:

$$\int_0^{\gamma_j\sigma^2}\left(\int_0^{\gamma_k\left(x\zeta_j+\sigma^2\right)}\frac{1}{p}e^{-\frac{y}{p}}dy\right)\frac{1}{s}e^{-\frac{x}{s}}dx+\int_{\gamma_j\sigma^2}^{\infty}\left(\int_{\frac{x}{\gamma_j\zeta_k}-\frac{\sigma^2}{\zeta_k}}^{\gamma_k\left(x\zeta_j+\sigma^2\right)}\frac{1}{p}e^{-\frac{y}{p}}dy\right)\frac{1}{s}e^{-\frac{x}{s}}dx \tag{F.28}$$

The expression (F.26) and the first term in (F.28) complement each other so their sum becomes

$$\int\limits_0^{\gamma_j\sigma^2}\left(\int\limits_0^{\infty}\frac{1}{p}e^{-\frac{y}{p}}dy\right)\frac{1}{s}e^{-\frac{x}{s}}dx=\int\limits_0^{\gamma_j\sigma^2}\frac{1}{s}e^{-\frac{x}{s}}dx=F_e\left(\gamma_j\sigma^2;s\right) \tag{F.29}$$

The second component of (F.28) is slightly more involved. First, let us focus on the relationship between the limits of its second integral. After rearranging the terms we obtain:

$$x\left(\gamma_k\zeta_j-\frac{1}{\gamma_j\zeta_k}\right)\geq-\gamma_k\sigma^2-\frac{\sigma^2}{\zeta_k}. \tag{F.30}$$

Since the right side is negative and $x>0$, then (F.30) is always true whenever $\gamma_k\zeta_j-\frac{1}{\gamma_j\zeta_k}>0$ leading to no additional constraint on $x$. However, when $\gamma_k\zeta_j-\frac{1}{\gamma_j\zeta_k}$ is negative, or equivalently $\gamma_j\gamma_k\zeta_j\zeta_k<1$, then the upper limit on $x$ appears:

$$x\leq\frac{\gamma_j\gamma_k\zeta_k\sigma^2+\gamma_j\sigma^2}{1-\gamma_j\gamma_k\zeta_j\zeta_k} \tag{F.31}$$

which is a valid limit since $\frac{\gamma_j\gamma_k\zeta_k\sigma^2+\gamma_j\sigma^2}{1-\gamma_j\gamma_k\zeta_j\zeta_k}>\frac{\gamma_j\sigma^2}{1-\gamma_j\gamma_k\zeta_j\zeta_k}>\gamma_j\sigma^2$. The missing integral yields

$$\begin{aligned}&\int_{\gamma_j\sigma^2}^{C}\left(\int_{\frac{x}{\gamma_j\zeta_k}-\frac{\sigma^2}{\zeta_k}}^{\gamma_k\left(x\zeta_j+\sigma^2\right)}\frac{1}{p}e^{-\frac{y}{p}}dy\right)\frac{1}{s}e^{-\frac{x}{s}}dx\\&=\frac{1}{s}\int_{\gamma_j\sigma^2}^{C}e^{\frac{\sigma^2}{\zeta_k p}}e^{-x\frac{s+\gamma_j\zeta_k p}{\gamma_j\zeta_k ps}}-e^{-\frac{\gamma_k\sigma^2}{p}}e^{-x\frac{\gamma_k\zeta_j s+p}{ps}}dx\\&=\frac{\gamma_j\zeta_k p}{s+\gamma_j\zeta_k p}\left(e^{-\frac{\gamma_j\sigma^2}{s}}-e^{-C\frac{s+\gamma_j\zeta_k p}{\gamma_j\zeta_k ps}+\frac{\sigma^2}{\zeta_k p}}\right)\\&\quad-\frac{p}{p+\gamma_k\zeta_j s}\left(e^{-\frac{\gamma_j\sigma^2}{s}}e^{-\gamma_k\sigma^2\frac{1+\gamma_j\zeta_j}{p}}-e^{-C\frac{\gamma_k\zeta_j s+p}{ps}-\frac{\gamma_k\sigma^2}{p}}\right).\end{aligned} \tag{F.32}$$

When $\gamma_j\gamma_k\zeta_j\zeta_k>1$ the second and fourth term in (F.32) disappear since $\lim\limits_{C\to\infty}e^{-C\frac{s+\gamma_j\zeta_k p}{\gamma_j\zeta_k ps}+\frac{\sigma^2}{\zeta_k p}}=0$ and $\lim\limits_{C\to\infty}e^{-C\frac{\gamma_k\zeta_j s+p}{ps}-\frac{\gamma_k\sigma^2}{p}}=0$. Otherwise,

$C=\frac{\gamma_j\gamma_k\zeta_k\sigma^2+\gamma_j\sigma^2}{1-\gamma_j\gamma_k\zeta_j\zeta_k}$ and after some simplification we obtain that $e^{-C\frac{s+\gamma_j\zeta_k p}{\gamma_j\zeta_k ps}+\frac{\sigma^2}{\zeta_k p}} = e^{-C\frac{\gamma_k\zeta_j s+p}{ps}-\frac{\gamma_k\sigma^2}{p}} = e^{-\frac{\sigma^2}{1-\gamma_j\gamma_k\zeta_j\zeta_k}\left(\gamma_j\frac{\gamma_k\zeta_k+1}{s}+\gamma_k\frac{\gamma_j\zeta_j+1}{p}\right)}$. The total error probability is then the sum of (F.29) and (F.32).

# C Appendix

Let us consider a received signal over a single TF-block given by

$$\mathbf{y}' = h_1^{(a)}\mathbf{x}_1 + h_2^{(b)}\mathbf{x}_2 + \mathbf{n}_1 \tag{F.33}$$

and let us assume that in one of the previous uplink phases the interferer (UE 2) already had an unsuccessful transmission attempt of the packet so the BS has stored

$$\mathbf{y}'' = h_2^{(b-1)}\mathbf{x}_2 + h_3^{(c)}\mathbf{x}_3 + \mathbf{n}_2 \tag{F.34}$$

where $h_1^{(a)}$, $h_2^{(b)}$, $h_2^{(b-1)}$ and $h_3^{(c)}$ denote the complex channel coefficients and the transmit power and path loss coefficients of each user were omitted for simplicity. Instead of attempting to decode $\mathbf{x}_1$ directly from $\mathbf{y}'$ which would yield SINR equal to $\frac{\left|h_1^{(a)}\right|^2}{\left|h_2^{(b)}\right|^2+\sigma^2}$ the receiver can consider signal $\mathbf{y}' - q\mathbf{y}''$ which yields SINR $\frac{\left|h_1^{(a)}\right|^2}{\left|h_2^{(b)}-qh_2^{(b-1)}\right|^2+|q|^2\left(\left|h_3^{(c)}\right|^2+\sigma^2\right)+\sigma^2}$. The expression is maximized for $q = \frac{h_2^{(b)}h_2^{(b-1)*}}{\left|h_2^{(b-1)}\right|^2+\left|h_3^{(c)}\right|^2+\sigma^2}$ in which case the the SINR becomes $\frac{\left|h_1^{(a)}\right|^2}{\left|h_2^{(b)}\right|^2\frac{\left|h_3^{(c)}\right|^2+\sigma^2}{\left|h_2^{(b-1)}\right|^2+\left|h_3^{(c)}\right|^2+\sigma^2}+\sigma^2}$. It's easy to notice that, compared to (F.33), the power of the interfering component $\left|h_2^{(b)}\right|^2$ is now scaled down by a factor

$$\zeta_2 = \frac{\left|h_3^{(c)}\right|^2+\sigma^2}{\left|h_2^{(b-1)}\right|^2+\left|h_3^{(c)}\right|^2+\sigma^2} = \left(1+\frac{\left|h_2^{(b-1)}\right|^2}{\left|h_3^{(c)}\right|^2+\sigma^2}\right)^{-1}. \tag{F.35}$$

The amount is directly related to the SINR that UE 2 experienced in its past replica (F.34).

Note that the operation described above has this particularly simple form only when the signal $\mathbf{y}''$ used to reduce the interference is uncorrelated with the symbols $\mathbf{x}_1$, but this is ensured already since in CC we do not allow $\mathbf{x}_1$ and $\mathbf{x}_2$ to be paired together twice.

# References

[1] 3GPP, "NR;NR and NG-RAN Overall description; Stage-2," 3rd Generation Partnership Project (3GPP), TS 38.300, 2019, v15.7.0.

[2] P. Popovski, "Ultra-reliable communication in 5G wireless systems," in *1st International Conference on 5G for Ubiquitous Connectivity*, 2014, pp. 146–151.

[3] P. Schulz, M. Matthe, H. Klessig, M. Simsek, G. Fettweis, J. Ansari, S. A. Ashraf, B. Almeroth, J. Voigt, I. Riedel, A. Puschmann, A. Mitschele-Thiel, M. Muller, T. Elste, and M. Windisch, "Latency Critical IoT Applications in 5G: Perspective on the Design of Radio Interface and Network Architecture," *IEEE Communications Magazine*, vol. 55, no. 2, pp. 70–78, 2017.

[4] Q. Zhang, J. Liu, and G. Zhao, "Towards 5g enabled tactile robotic telesurgery," *CoRR*, vol. abs/1803.03586, 2018. [Online]. Available: http://arxiv.org/abs/1803.03586

[5] 3GPP, "Service requirements for the 5G system; Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.261, 2018, v16.5.0.

[6] B. Soret, P. Mogensen, K. I. Pedersen, and M. C. Aguayo-Torres, "Fundamental tradeoffs among reliability, latency and throughput in cellular networks," in *2014 IEEE Globecom Workshops (GC Wkshps)*, 2014, pp. 1391–1396.

[7] S. Nagata, L. H. Wang, and K. Takeda, "Industry Perspectives: Latency reduction toward 5G," *IEEE Wireless Communications*, vol. 24, no. 3, pp. 2–4, 2017.

[8] G. Pocovi, B. Soret, K. I. Pedersen, and P. Mogensen, "MAC layer enhancements for ultra-reliable low-latency communications in cellular networks," in *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*, 2017, pp. 1005–1010.

[9] G. Caire and D. Tuninetti, "The throughput of hybrid-ARQ protocols for the Gaussian collision channel," *IEEE Transactions on Information Theory*, vol. 47, no. 5, pp. 1971–1988, 2001.

[10] S. Parkvall, E. Dahlman, P. Frenger, P. Beming, and M. Persson, "The evolution of WCDMA towards higher speed downlink packet data access," in *IEEE VTS 53rd Vehicular Technology Conference, Spring 2001. Proceedings (Cat. No.01CH37202)*, vol. 3, 2001, pp. 2287–2291 vol.3.

[11] A. Anand and G. de Veciana, "Resource Allocation and HARQ Optimization for URLLC Traffic in 5G Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2411–2421, 2018.

[12] N. H. Mahmood, A. Karimi, G. Berardinelli, K. I. Pedersen, and D. Laselva, "On the Resource Utilization of Multi-Connectivity Transmission for URLLC Services in 5G New Radio," in *2019 IEEE Wireless Communications and Networking Conference Workshop (WCNCW)*, 2019, pp. 1–6.

[13] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, no. 10, pp. 1834–1853, 2018.

[14] Z. Wu, K. Lu, C. Jiang, and X. Shao, "Comprehensive Study and Comparison on 5G NOMA Schemes," *IEEE Access*, vol. 6, pp. 18 511–18 519, 2018.

[15] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A Survey of Non-Orthogonal Multiple Access for 5G," *IEEE Communications Surveys Tutorials*, vol. 20, no. 3, pp. 2294–2323, 2018.

[16] T. V. K. Chaitanya and E. G. Larsson, "Optimal Power Allocation for Hybrid ARQ with Chase Combining in i.i.d. Rayleigh Fading Channels," *IEEE Transactions on Communications*, vol. 61, no. 5, pp. 1835–1846, 2013.

[17] L. Szczecinski, S. R. Khosravirad, P. Duhamel, and M. Rahman, "Rate Allocation and Adaptation for Incremental Redundancy Truncated HARQ," *IEEE Transactions on Communications*, vol. 61, no. 6, pp. 2580–2590, 2013.

[18] A. Avranas, M. Kountouris, and P. Ciblat, "Energy-Latency Tradeoff in Ultra-Reliable Low-Latency Communication With Retransmissions," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 11, pp. 2475–2485, 2018.

[19] ——, "Throughput Maximization and IR-HARQ Optimization for URLLC Traffic in 5G Systems, year=2019, volume=, number=, pages=1-6, doi=10.1109/ICC.2019.8761154," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*.

[20] Y. Xu, C. Shen, T.-H. Chang, S.-C. Lin, Y. Zhao, and G. Zhu, "Transmission Energy Minimization for Heterogeneous Low-Latency NOMA Downlink," *IEEE Transactions on Wireless Communications*, vol. 19, no. 2, pp. 1054–1069, 2020.

[21] N. Zhang, J. Wang, G. Kang, and Y. Liu, "Uplink Nonorthogonal Multiple Access in 5G Systems," *IEEE Communications Letters*, vol. 20, no. 3, pp. 458–461, 2016.

[22] J. Choi, "On Power and Rate Allocation for Coded Uplink NOMA in a Multicarrier System," *IEEE Transactions on Communications*, vol. 66, no. 6, pp. 2762–2772, 2018.

[23] ——, "On HARQ-IR for Downlink NOMA Systems," *IEEE Transactions on Communications*, vol. 64, no. 8, pp. 3576–3584, 2016.

[24] Y. Xu, D. Cai, F. Fang, Z. Ding, C. Shen, and G. Zhu, "HARQ-CC enabled NOMA designs with outage probability constraints," *CoRR*, vol. abs/1911.01167, 2019. [Online]. Available: http://arxiv.org/abs/1911.01167

[25] J. Choi, "H-ARQ Based Non-Orthogonal Multiple Access with Successive Interference Cancellation," in *IEEE GLOBECOM 2008 - 2008 IEEE Global Telecommunications Conference*, 2008, pp. 1–5.

[26] R. Kotaba, C. N. Manchon, N. M. K. Pratas, T. Balercia, and P. Popovski, "Improving Spectral Efficiency in URLLC via NOMA-Based Retransmissions," in *ICC 2019 - 2019 IEEE International Conference on Communications (ICC)*, 2019, pp. 1–7.

[27] E. Dosti, U. L. Wijewardhana, H. Alves, and M. Latva-aho, "Ultra reliable communication via optimum power allocation for type-i ARQ in finite block-length," *CoRR*, vol. abs/1701.08617, 2017. [Online]. Available: http://arxiv.org/abs/1701.08617

[28] D. Buckingham and M. C. Valenti, "The information-outage probability of finite-length codes over AWGN channels," in *2008 42nd Annual Conference on Information Sciences and Systems*, 2008, pp. 390–395.

[29] J. Edmonds, "Paths, Trees, and Flowers," *Canadian Journal of Mathematics*, vol. 17, p. 449–467, 1965.

[30] 3GPP, "Service requirements for the 5G system; Stage 1," 3rd Generation Partnership Project (3GPP), TS 22.261, 2019, v16.10.0.

# Paper G

## How to Identify and Authenticate Users in Massive Unsourced Random Access

Radosław Kotaba, Anders E. Kalør, Petar Popovski, Israel Leyva-Mayorga, Beatriz Soret, Maxime Guillaud and Luis G. Ordóñez

The paper has been published in the
*IEEE Communications Letters* Vol. 25, No. 12 pp. 3795–3799, 2021.

*The layout has been revised.*

# Abstract

*Identification and authentication are two essential features for traditional random access protocols. In ALOHA-based random access, the packets usually include a field with a unique user address. However, when the number of users is massive and relatively small packets are transmitted, the overhead of including such field becomes restrictive. In unsourced random access (U-RA), the packets do not include any address field for the user, which maximizes the number of useful bits that are transmitted. However, by definition an U-RA protocol does not provide user identification. This paper presents a scheme that builds upon an underlying U-RA protocol and solves the problem of user identification and authentication. In our scheme, the users generate a message authentication code (MAC) that provides these functionalities without violating the main principle of unsourced random access: the selection of codewords from a common codebook is i.i.d. among all users.*

***Keywords***— massive access, unsourced random access

# 1 Introduction

One of the hallmarks of the fifth generation (5G) wireless systems and beyond is massive Internet of Things (IoT) connectivity [1]. A scenario for massive IoT access features a large number of devices (typically in the order of thousands) connected to a Base Station (BS), each being sporadically active and sending short data packets (e.g., a few kilobytes or bytes). This sporadic activation entails that the set of devices trying to access at a given instant is unknown, thereby requiring random access protocols.

In the classical ALOHA model for random access [2], a packet is the smallest, atomic unit of information. The analyses in massive access scenarios are usually performed with an infinite population, where the number of users is $N \to \infty$. However, in order to examine the fundamental performance bounds of massive access protocols, one needs to look into the structure of the packet. This is where the assumption $N \to \infty$ leads to a paradox: to make user identification possible, a field with a unique user address of $\approx \log_2 N$ bits must be included in a packet of finite and relatively short length. To deal with this paradox, two information-theoretic approaches have been introduced.In the many access channel [3] the number of users is given as a function of the codeword length, which allows to preserve identification capabilities even as both tend to infinity.

Differently from this, [4] addresses the problem of $N \to \infty$ with finite blocklength (FBL) packets by assuming that a packet does not contain the address of the sender. This makes the access scheme *unsourced*, and leads to the case in which all users share the same codebook. While U-RA was initially proposed as a theoretically elegant scheme, it can also be justified by the desire to simplify the receiver and reduce the communication overhead. This is particularly important for short IoT packets where the address field can constitute a large portion of the packet [5].

The unsourced, uncoordinated nature of the problem and the FBL effects have implications in the design of practical low-complexity coding schemes, which has been the focus of several works. Bounds of the performance of finite-length codes were derived in the initial paper by Polyanskiy [4], and later generalized to the quasi-static fading channel [6]. The basic unsourced random access was extended to the case with a large number of antennas in [7], and the impact of correlated activations was studied in [8].

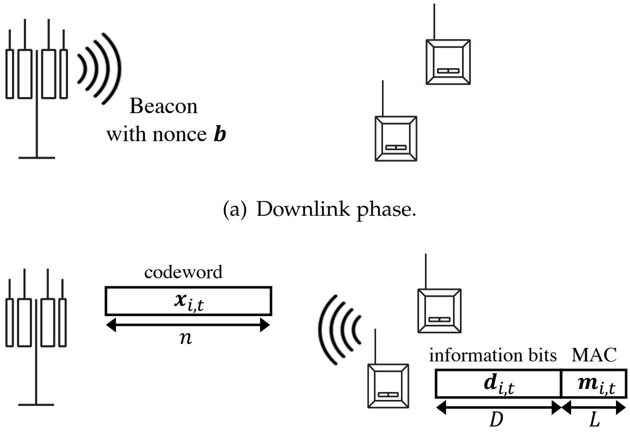
Despite its benefits in terms of efficiency, U-RA keeps the question of user identification (and, consequently, user authentication) open. In this paper, we aim to answer the following: *assuming that a given protocol for unsourced random access is available as a black box, how can it be extended to support user identification and authentication?* Rather than deferring this question to the higher layers or additional transmissions, in this contribution we present a scheme that enables those functionalities at the lower layers, in a way that is consistent with the paradigm of U-RA, i.e., when users share the same codebook. In that sense, the main contribution of our scheme is that it enables the identification and authentication of users over U-RA; the potential performance gains compared to sourced random access is of secondary importance.

The key idea is to generate and append a message authentication code (MAC)[1] to the packets (rather than an explicit address), which enables the identification and authentication of the users while complying with the main assumptions of U-RA. For this, we employ a two-step procedure as illustrated in Fig. G.1. First, the BS broadcasts a beacon with a *nonce* to the users prior to data transmission. A nonce is an arbitrary number generated periodically by the BS that is allowed to be used only once by each node to prevent replay of messages. Then, each active user generates a MAC based on the nonce, a secret key known only by the user and the BS (e.g., pre-shared using Universal Subscriber Identity Module (USIM) as in LTE [9]), and the data to be sent; this field is appended to the packet and transmitted as shown in Fig. G.1(b).

# 2 System model

We study the massive random access scenario as described by Polyanskiy [4], where $N \to \infty$ users communicate through a time-slotted channel with a single BS. Although the proposed scheme works without modifications with a (potentially massive) MIMO BS, we assume a single antenna BS to simplify the presentation. At each time slot, $K$ out of the $N$ users are active and send

[1] To avoid confusion between this term and the widely-used acronym for medium access control, the latter is avoided throughout the paper.



(a) Downlink phase.

(b) Uplink phase.

**Fig. G.1:** The two-step procedure.

messages $\mathcal{W} = W_1, W_2, \ldots, W_K$ in the uplink, where $W_i$ is drawn independently and uniformly at random from the message set $\mathcal{M} = \{1, 2, \ldots, M\}$. For typical massive IoT scenarios, $K$ will be in the range of 50 to a few hundreds [10]. All users share the same encoder $f : [M] \to \mathcal{X}^n$, and use it to construct the codewords $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K$ as $\mathbf{x}_i = f(W_i)$, which are subject to the constraint $\|\mathbf{x}_i\|_2^2 \leq nP$, where $P$ is the average energy per symbol. The codewords are transmitted over a permutation-invariant and memoryless multiple access channel $P_{Y|X_1^K} : \mathcal{X}^{n \times K} \to \mathcal{Y}^n$, i.e., it satisfies $P_{Y|X_1^K}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_K) = P_{Y|X_1^K}(\mathbf{y}|\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(K)})$ for any $\mathbf{y} \in \mathcal{Y}^n$ and $\mathbf{x}_1, \ldots, \mathbf{x}_K \in \mathcal{X}^{n \times K}$, and any permutation $\pi$.

We assume that the BS periodically broadcasts a beacon in the downlink as depicted in Fig. G.1. The beacon includes the necessary information for the users to synchronize, to obtain the main configuration parameters, and to estimate and invert the channel. To keep the presentation simple and aligned with [4], we assume that channel inversion is perfect, so that fading can be neglected[2] and the uplink transmissions are only affected by additive white Gaussian noise, denoted by $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma \mathbf{I}_n)$. Consequently, the resulting Gaussian multiple access channel model at a given time slot is

$$\mathbf{y} = \sum_{i=1}^{K} \mathbf{x}_i + \mathbf{z} \tag{G.1}$$

At the BS, the decoder $g : \mathcal{Y}^n \to [M]^K$ outputs an unordered list of $K$

[2] We note that the users who cannot perform inversion due to poor channel conditions can simply remain inactive, which leads to the problem that is structurally the same.

messages from $\mathcal{M}$. In line with the U-RA literature [4], we assume that $K$ is fixed and known to the decoder. We note that this assumption allows the codebook to be designed based on $K$, which does not reflect a true random access scenario. In practice, the codebook would have to be designed based on the expected maximum (or average) number of active users instead. Similarly, in practical implementations the decoder, rather than outputting a fixed number of messages, might rely on separate activity detection [10].

An error occurs whenever the $g(\mathbf{y})$ does not contain a transmitted message, or if multiple users transmit the same message. More specifically, an error for user $i$ is defined as $E_i = \{W_i \notin g(\mathbf{y})\} \cup \{W_i = W_j \text{ for some } j \neq i\}$. Note that since we assume the decoder always outputs $K$ messages, it implies that for each error $E_i$, the list $g(\mathbf{y})$ must contain a message which was not transmitted by any of the devices. We shall refer to this set $g(\mathbf{y}) \setminus \mathcal{W}$ as decoder false positives. Denoting by $k_{\mathrm{TP}}$ the number of genuine (true positive) messages and by $k_{\mathrm{FP}}$ the number of false positives in the set $g(\mathbf{y})$, we have that $k_{\mathrm{TP}} + k_{\mathrm{FP}} = K$.

# 3 Identification and authentication in unsourced random access protocols

The key idea behind the proposed scheme is to generate MAC that enables identification and authentication of the users and that can be applied to U-RA protocols. The MAC $\mathbf{m}_i = \{0,1\}^L$ is generated by user $i$ based on its data $\mathbf{d}_i \in \{0,1\}^D$ of size $D$, its secret key $\mathbf{k}_i$, and a nonce $\mathbf{b}$. The secret key is fixed and only known by the corresponding user and the BS, e.g., pre-shared using USIM as in LTE [9]. The MAC length $L$ is fixed and independent from the other parameters.

Our scheme is divided into phases as shown in Fig. G.1. At the beginning of each round, the BS generates a nonce and broadcasts it to all the devices. The nonce is a sequence or pseudo-random number that changes in each round but is otherwise public. Once the nonce is received, a given user $i$ generates the MAC $\mathbf{m}_i$ based on the data bits it wants to transmit $\mathbf{d}_i$, its secret key $\mathbf{k}_i$, and the nonce $\mathbf{b}$, i.e. $\mathbf{m}_i = h(\mathbf{d}_i, \mathbf{k}_i, \mathbf{b})$, where $h(\cdot)$ is designed to be computationally hard to invert and have low collision probability (i.e., the output is approximately uniform for any input distribution). The user appends the MAC to the data to create a packet and transmits it, as shown in Fig. G.1(b). At the BS, the packets are first decoded to extract $[\widehat{\mathbf{d}}_i, \widehat{\mathbf{m}}_i]$ tuples. For each, the message authenticity can be verified and the identity of the sender determined by computing the MACs of the data part $h(\widehat{\mathbf{d}}_i, \mathbf{k}_j, \mathbf{b})$ with different secret keys $\mathbf{k}_j$ and comparing them with the MAC in the received packet $\widehat{\mathbf{m}}_i$. If a match is found, the authenticator declares the user with the

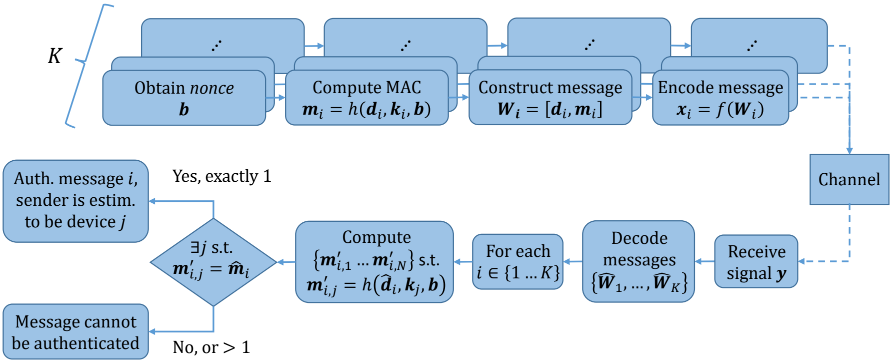

$K$
Obtain *nonce* $\boldsymbol{b}$
Compute MAC $\boldsymbol{m}_i = h(\boldsymbol{d}_i, \boldsymbol{k}_i, \boldsymbol{b})$
Construct message $\boldsymbol{W}_i = [\boldsymbol{d}_i, \boldsymbol{m}_i]$
Encode message $\boldsymbol{x}_i = f(\boldsymbol{W}_i)$
Channel
Receive signal $\boldsymbol{y}$
Decode messages $\{\widehat{\boldsymbol{W}}_1, \dots, \widehat{\boldsymbol{W}}_K\}$
For each $i \in \{1 \dots K\}$
Compute $\{\boldsymbol{m}'_{i,1} \dots \boldsymbol{m}'_{i,N}\}$ s.t. $\boldsymbol{m}'_{i,j} = h(\widehat{\boldsymbol{d}}_i, \boldsymbol{k}_j, \boldsymbol{b})$
$\exists j$ s.t. $\boldsymbol{m}'_{i,j} = \widehat{\boldsymbol{m}}_i$
Yes, exactly 1
Auth. message $i$, sender is estim. to be device $j$
No, or > 1
Message cannot be authenticated

**Fig. G.2:** Block diagram of the proposed scheme including message generation and subsequent decoding, authentication and identification.

matching key to be the potential transmitter. The full scheme is depicted in Fig. G.2.

While a nonce is commonly used to prevent replay attacks, in our scheme it has the additional function of randomizing the MAC. That is, without a nonce, a particular piece of data and secret key from a given device would always produce the same MAC, which violates the assumption that all codewords are equally likely. Typical methods to generate the MAC include, e.g., symmetric key cryptography as in AES-CMAC (RFC 4493), used in LoRaWAN, or a HMAC (RFC 2104). Any of these methods can be applied to our scheme, so the MAC is computationally challenging to guess without the secret key.

Note that in our scheme *cryptographic errors*, which we define as any instance where the matching MAC is generated by a key that does not belong to the actual sender, can occur. They are possible since: 1) the generated MAC might not be a unique identifier for the user (unlike the actual address) and 2) the BS must generate many MACs with different secret keys to find the one that matches the one in the received packet.

Therefore, several tradeoffs arise. The first one is between the length of the metadata and the amount of *cryptographic errors*, where in the extreme case with no metadata (i.e. neither MAC nor address) identification and authentication cannot be provided. Meanwhile, longer packets entail higher energy. Another tradeoff involves the computational complexity and probability of cryptographic errors that both increase with the number of devices supported by the system[3].

[3] It could be argued that the scheme is not practical as $N \to \infty$. However, in practice good performance was observed for $N$ as large as $10^6$ and $K > 100$.

# 4 Cryptographic errors: collisions, false positives and misidentifications

The probability of decoder false positives describes only the physical layer performance of U-RA. The full characterization of the proposed scheme has to take into account also potential cryptographic errors, erroneous acceptance of false positives, and misidentification events. For the purpose of this evaluation, we assume ideal MACs that are uniformly distributed, i.e., the probability that a given $(data, key, nonce)$ tuple produces a specific MAC of length $L$ is $p = 2^{-L}$.

## 4.1 Exhaustive search

We first consider authentication using exhaustive search, where all keys are tried on each message. We start by studying the per-user cryptographic error probabilities. A genuine message $W'$ with data $\mathbf{d}'$ transmitted by user $u'$ will fail to be authenticated whenever any of the keys from users $u \neq u'$ produces the same MAC when applied to $\mathbf{d}'$. We refer to those events as *type 1* errors. Since there are $N-1$ other keys, the type 1 event happens with probability

$$p_{\mathrm{t1}} = 1 - (1-p)^{N-1} \tag{G.2}$$

Because we assume that each user transmits at most one message per round, an error occurs also when the key of user $u'$ produces a valid MAC for any of the other decoded messages in $g(\mathbf{y}) \setminus \{W'\}$. Given that there are $K-1$ other decoded messages, this *type 2* error happens with probability

$$p_{\mathrm{t2}} = 1 - (1-p)^{K-1}. \tag{G.3}$$

Taking into account both types of errors, the probability that a genuine message is successfully authenticated is

$$p_{\mathrm{s\_auth}} = (1-p)^{N+K-2}. \tag{G.4}$$

Another type of event is when a false positive message produced by the decoder is erroneously authenticated. While (G.4) is conditional on the fact that there is at least one key that authenticates the message, here we cannot assume that. Since the keys from the genuine messages cannot be used again without causing type 2 error, there are $N - k_{\mathrm{TP}}$ keys that can potentially decode the false positive message without resulting in a collision. Since each of these keys accepts the message with probability $p_{\mathrm{s\_auth}}$, the probability of accepting a false positive message from the decoder is

$$p_{\mathrm{fp\_auth}} = (N - k_{\mathrm{TP}})\, p \cdot p_{\mathrm{s\_auth}} = (N - k_{\mathrm{TP}}) p (1-p)^{N+K-2}. \tag{G.5}$$

Note that the authenticator is generally unable to determine whether a message that *fails* to be authenticated belongs to the set of decoder true positive or decoder false positive messages. The only exception to this is the special case in which no key is able to decode a given message, which can only happen for false positive messages. The probability that this happens for a given false positive message is $p_{\text{d\_fp}} = (1-p)^N$.

## 4.2 Heuristic search

We now turn our attention to the heuristic search, in which the authenticator tries keys only until it finds a matching key. While more efficient, this approach cannot detect type 1 and type 2 errors defined above, and thus the probability of erroneously authenticating a message increases.

Providing exact analytical expressions for the heuristic case proves to be difficult, due to the dependency on the order in which packets are authenticated, the number of decoder false positives and true positives, and how they are interleaved. To that end, we will provide only approximations, noting that they are very close to the true values. We shall assume without loss of generality that the decoded messages are authenticated in the order $\widehat{W}_1, \widehat{W}_2, \dots, \widehat{W}_K$. Furthermore, we will neglect the events where the sender of message $\widehat{W}_j$ becomes incorrectly identified as the sender of one of the previous messages $\widehat{W}_1, \dots, \widehat{W}_{j-1}$, which happens with very low probability[4].

We first consider the probability of correctly authenticating a genuine message. In the heuristic search case the successful authentication of message $W_j$ can happen even if there are cryptographic collisions, as long as the correct user happens to be tested first. For a set of $i$ successfully authenticating keys, this happens with probability $1/i$. By marginalizing over the number of keys *additional* to the genuine key we obtain

$$p_{\text{s\_auth},j} = \sum_{i=0}^{N_j-1} \binom{N_j-1}{i} p^i (1-p)^{N_j-1-i} \left(\frac{1}{1+i}\right), \tag{G.6}$$

where $N_j$ is the number of remaining keys which is the total number of keys, $N$, minus those that have authenticated any of the previous messages. $N_j$ is nonincreasing, and $N_j \geq N - j + 1$ since the authenticator may have been unable to authenticate some of the previous $j-1$ messages. As already mentioned, in the heuristic approach the detection of collisions (type 1 and type 2 errors) is not possible, which can result in misidentification, i.e., attributing a genuine message to the wrong user. The probability of misidentifying the

[4] Note that we do not neglect misidentification events in general, but only the case where specific user authenticates a specific message, which is tied to the probability $p$ and hence very low.

$j$-th message is the probability that one or more of the $N_j - 1$ non-genuine keys authenticate the message before the correct one:

$$p_{\text{mis\_id},j} = \sum_{i=1}^{N_j-1} \binom{N_j - 1}{i} p^i (1-p)^{N_j-1-i} \left(1 - \frac{1}{1+i}\right) = 1 - p_{\text{s\_auth},j}. \quad \text{(G.7)}$$

On the other hand, if the message is a false positive, the probability of accepting it is equal to the probability of having at least one key which produces a matching MAC:

$$p_{\text{fp\_auth},j} = 1 - (1-p)^{N_j}. \quad \text{(G.8)}$$

We note that from the point of view of the receiver there is no difference between misidentification and false positive authentication, hence, the total error probability should include both. For a given packet, which is genuine with probability $p_{\text{TP}}$ and a false positive with probability $p_{\text{FP}}$ we obtain

$$p_{\text{mis\_auth},j} = p_{\text{TP}} p_{\text{mis\_id},j} + p_{\text{FP}} p_{\text{fp\_auth},j}. \quad \text{(G.9)}$$

Lastly, let us remark that when $N \gg K$, we have that $N_j \approx N$. By making this substitution in (G.6) - (G.9), we obtain a rigorous upper bound on the probability of each type of error. Furthermore, they become independent of packet number and allow us to drop the subscript $j$ which simplifies the comparison between the exhaustive and heuristic approach.

In Fig. G.3 we show the probability of successful authentication and probability of mis-authentication as a function of the total number of devices $N$ for the exhaustive and heuristic search. In addition to the small gain in terms of success probability, the latter method allows to reduce the complexity as, on average, it requires only half of the MAC checks (assuming the probability of transmission is uniform across the devices). This is at the cost of an increased probability of mis-authentication. Since the eq. (G.6) and (G.9) used to produce the solid red curves are approximations that neglect some of the effects mentioned earlier, we provide also the results obtained through numerical simulations. Clearly, the differences are very minor making the approximations a viable tool.

## 4.3 Spoofing attacks

It is of interest to consider what happens when an attacker sends a forged message with the intent of getting it accepted by the authenticator. Without private keys the attacker is not able to compute the correct MAC for the spoofed *data* and current *nonce* so it has to generate MAC bits at random. As such, from the cryptographic point of view, the message acts as a false positive, and the transmitter cannot target a specific device (i.e. it cannot choose

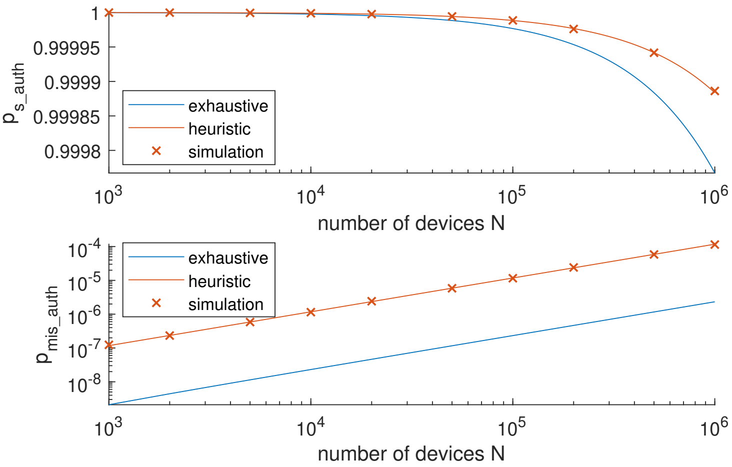


**Fig. G.3:** Probability of successful authentication and probability of mis-authentication as a function of the total number of devices $N$. The number of messages is $K = 100$, $p_{TP} = 0.99$, $p_{FP} = 0.01$, MAC length $L = 32$bits.

whom it is impersonating). However, from the physical layer point of view this is an actual transmitted codeword, and as such subject to probability of decoding $p_{\text{TP}}$, so the total probability of successful spoof is

$$p_{\text{s\_spoof}} = p_{\text{TP}} \left(1 - (1-p)^N\right). \tag{G.10}$$

This is to be compared to the traditional frame structure where the source address is included in the packet. In that case, the authenticator only tries the single MAC associated to that user, and the spoof attack is successful with probability $p_{\text{TP}} p$.

# 5 Results

We start by looking into the physical layer performance. The results were obtained based on the random coding bound given in [4, eq. (3)-(10)]. The codeword length (number of symbols) was chosen to be $n = 2^{15} = 32768$. In Fig. G.4 we depict the achievable error probability as a function of the energy per codeword $nP$. The values are shown for a range of packet sizes $B$ and for $K = 50$ and $K = 150$. In line with the assumption that two users selecting the same message is considered an error (c.f. Section 2), each curve has a floor at $\binom{K}{2}/M = \binom{K}{2}/2^B$ (visible only for 32 bits). In general, the higher B and K are, the steeper the curves become and the transition from almost certain error $p_{\text{FP}} \approx 1$ to very high reliability (such as $p_{\text{FP}} < 10^5$) becomes increasingly abrupt. This is even better explained with Fig. G.5 which shows the energy

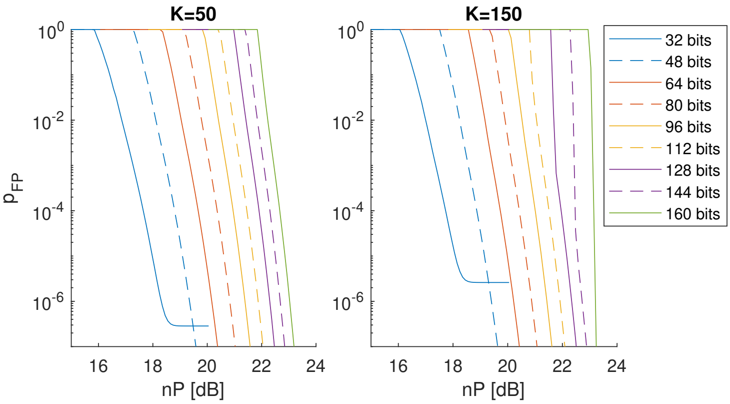


**Fig. G.4:** Achievable physical layer error probability as a function of the energy per codeword and packet size $B$. Two different values of $K$ are considered.

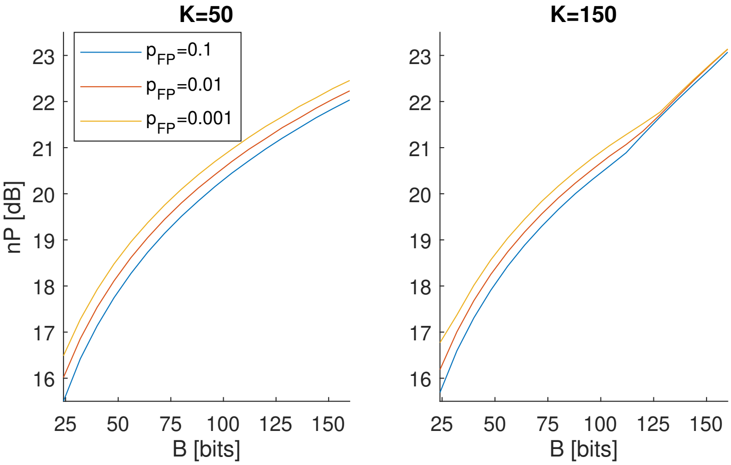


**Fig. G.5:** Minimum energy required to achieve fixed physical layer error probability as a function of the number of information bits $B$.

per codeword as a function of $B$ for fixed error rates. Firstly, as the packet size increases, less energy is needed to decrease $p_{\text{FP}}$. For example, with $K = 50$, when $B = 25\,$b improving error rate from $10^{-1}$ to $10^{-3}$ requires 1 dB, while with $B > 100\,$b the same shift requires less than 0.5 dB. Secondly, there is a point where the system turns from being noise-limited to interference-limited (curves merging). Such a transition occurs for lower packet sizes the more simultaneous messages $K$ there are.

In Fig. G.6 we combine all the earlier insights and look into the total probability of mis-authentication that takes into account both the physical and cryptographic layer performance. These results are obtained assuming a

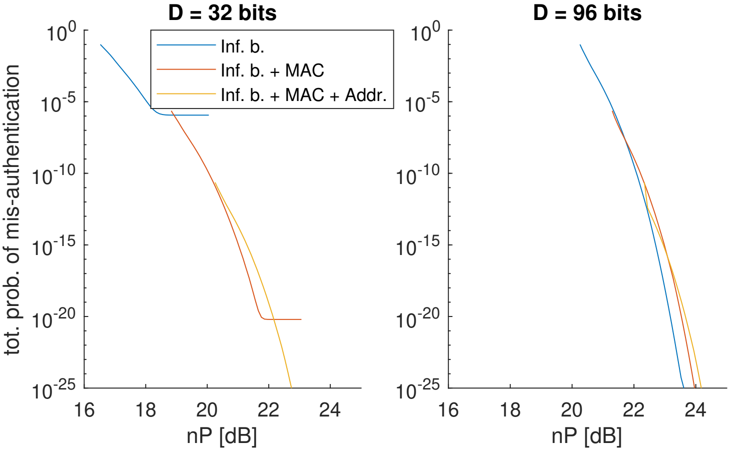


**Fig. G.6:** Total probability of mis-authentication as a function of energy, taking into account both physical and cryptographic layer. The number of messages is $K = 100$ and $N = 10^5$.

population of $N = 10^5$ users and $K = 100$ messages. In the plots, the blue line represents a packet consisting solely of the information bits, i.e. $B = D$. Since there is no additional means of authentication, every decoded packet is accepted and hence $p_{\text{miss\_auth}} = p_{\text{FP}}$. The red line denotes our proposed scheme in which the packet consists of information bits and a MAC, that is, in total $B = D + L$, where $L = 32\,\text{b}$ is fixed. The values reported here correspond to the exhaustive search variant, hence, the total probability of mis-authentication is $p_{\text{miss\_auth}} = p_{\text{FP}} p_{\text{fp\_auth}}$ with $p_{\text{fp\_auth}}$ given by (G.5). Lastly, the yellow curve represents the classic packet structure, where the address is also included (here $A = 32\,\text{b}$ as well) which yields $B = D + L + A$. In such case, the receiver checks only one key corresponding to the given address, hence we have that $p_{\text{miss\_auth}} = p_{\text{FP}} p$. It is important to keep in mind that the most basic mode of operation (blue) does not provide any way of identifying the users, and as such it is not directly comparable with the other two. Furthermore, it might not provide sufficient level of reliability when the packets are very short, which is due to the floor on $p_{\text{FP}}$. What might be surprising, is that the classic packet structure actually performs slightly worse than our proposed scheme (at least until $10^{-20}$ level which should be more than enough). This is because even though the probability of MAC collision is significantly lower, the packet needs to be larger to accommodate the address, which requires higher energy.

# 6 Conclusions

In this paper we proposed a method to introduce identification and authentication capabilities to algorithms that follow the framework of unsourced random access. Our scheme adds very limited amount of metadata to the communication, which is especially important for short IoT packets. Furthermore, as a consequence of not including explicit user identification, the packets are fully anonymous to everyone except the BS, which opens the door to new use cases and applications. This is in contrast to traditional protocols, where only the message content is assumed to be secret while the identities are public. The extra functionalities come at the cost of increased processing at the receiver. However, our results show that by avoiding the address we can simultaneously improve the spectral efficiency, and, for a given given energy per codeword, decrease the overall mis-authentication probability compared to the case where the address is included in the packet.

# References

[1] C. Bockelmann, N. K. Pratas, G. Wunder, S. Saur, M. Navarro, D. Gregoratti, G. Vivier, E. De Carvalho, Y. Ji, v. Stefanović, P. Popovski, Q. Wang, M. Schellmann, E. Kosmatos, P. Demestichas, M. Raceala-Motoc, P. Jung, S. Stanczak, and A. Dekorsy, "Towards Massive Connectivity Support for Scalable mMTC Communications in 5G Networks," *IEEE Access*, vol. 6, pp. 28 969–28 992, 2018.

[2] N. Abramson, "THE ALOHA SYSTEM: another alternative for computer communications," in *Proceedings of the November 17-19, 1970, fall joint computer conference*. ACM, pp. 281-285, 1970.

[3] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian Many-Access Channels," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3516–3539, 2017.

[4] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527.

[5] G. Durisi, T. Koch, and P. Popovski, "Toward Massive, Ultrareliable, and Low-Latency Wireless Communication With Short Packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, Sept 2016.

[6] S. S. Kowshik and Y. Polyanskiy, "Fundamental Limits of Many-User MAC With Finite Payloads and Fading," *IEEE Transactions on Information Theory*, vol. 67, no. 9, pp. 5853–5884, 2021.

[7] A. Fengler, G. Caire, P. Jung, and S. Haghighatshoar, “Massive MIMO unsourced random access,” *CoRR*, vol. abs/1901.00828, 2019. [Online]. Available: http://arxiv.org/abs/1901.00828

[8] K. Stern, A. E. Kalør, B. Soret, and P. Popovski, “Massive Random Access with Common Alarm Messages,” in *2019 IEEE International Symposium on Information Theory (ISIT)*, 2019, pp. 1–5.

[9] ETSI, “Security architecture and procedures for 5G system,” TS 133 501 v16.6.0, 2021.

[10] A. Decurninge, I. Land, and M. Guillaud, “Tensor-Based Modulation for Unsourced Massive Random Access,” *IEEE Wireless Communications Letters*, vol. 10, no. 3, pp. 552–556, 2021.

# Paper H

## Unsourced Random Access With Authentication and Joint Downlink Acknowledgements

Radosław Kotaba, Anders E. Kalør, Petar Popovski, Israel Leyva-Mayorga, Beatriz Soret, Maxime Guillaud and Luis G. Ordóñez

The paper has been submitted to
*55th Asilomar Conference on Signals, Systems, and Computers*, 2021.

## Abstract

*In massive access scenarios, an unknown subset of a very large number of wireless devices transmit small packets to the base station (BS) in an uncoordinated manner. These scenarios have been a hot research topic in the context of 5G/Beyond-5G as they pose many challenges and require the design of highly efficient and lightweight communication protocols. This has inspired the paradigm of unsourced random access (U-RA) as a way to simplify multi-user decoding and reduce overhead in the uplink. However U-RA, being mainly a physical layer approach, lacks the ability to identify and authenticate users, which needs to be taken into account. Another important functionality is the transmission of acknowledgements (ACK) in the downlink. The naïve method of sending a dedicated message to each user may not be feasible in the massive random access scenario, thus a solution is to provide jointly encoded ACKs, which can achieve much higher efficiency at the cost of introducing false positives. In this paper we consider a system that combines U-RA with joint ACKs transmitted in the downlink. We focus on the systematic description and analysis of the false positives, as well as the design options, and the associated trade-offs among reliability, rate of retransmissions and power efficiency.*

# 1 Introduction

Random access protocols constitute a central component in the support of massive Internet of Things (IoT), where a large number of devices sporadically transmit small packets to a single base station (BS). The combination of the small packet sizes and the fact that the set of active users are difficult to predict makes grant-based transmission inefficient, and motivates the use of protocols in which the users transmit their data in an uncoordinated fashion.

Due to the small packet sizes, the metadata, such as the source address and message authentication code (MAC), take up a significant fraction of the total packet length and, in particular, the number of bits required for the address increases with the total number of users. However, a central assumption in the widely used ALOHA model of random access [1] is that the total number of users tends to infinity. This leads to the apparent paradox that the packet length needs to be infinitely large in order for the users to be identifiable.

Two information-theoretic models of random access have recently been introduced to address this paradox. The many access channel [2] circumvents the problem by specifying the number of users as a function of the codeword length, so that identification is possible as both tend to infinity. On the other hand, in Polyanskiy's model [3] the users share the same codebook, which precludes user identification. As a result, the model has been termed Unsourced Random Access (URA). Although the shared codebooks in URA

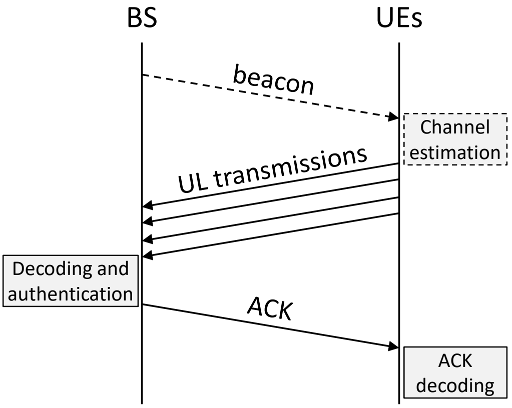


**Fig. H.1:** The diagram of the considered scheme. If used, the beacon broadcasted by the BS allows UEs to estimate and invert their channels.

were initially proposed to ensure that the model was theoretically elegant, it has also received much practical interest as the lack of source identity allows to simplify both the receiver and the transmitters, and to reduce the communication overhead [4].

Despite the large interest in Polyanskiy's model, user identities (and user authentication) are of high interest in many practical systems. Motivated by this, in previous work we proposed a method for the base station to identify and authenticate users that transmit using a URA-based protocol [5]. In this paper, we extend our previous work by considering fading channels and by introducing a joint feedback message broadcasted by the BS in the downlink, which allows the transmitting users to validate their transmissions. In addition, we propose to encode the feedback message using a Bloom filter [6], which reduces the number of bits required for the message at the cost of a small fraction of false positives. We show that this reduction ultimately leads to an increased probability that the users, following a transmission in the uplink, correctly decode and deduce the outcome of their transmissions compared to the case in which the feedback message is encoded by simply concatenating the acknowledgements of individual users.

The paper is organized as follows. In Sec. 2 we define the system model for uplink, authentication, and the downlink. Sec. 3 analyzes the uplink and authentication phases and the downlink is analyzed in Sec. 4. Numerical results are presented in Sec. 5, and conclusions are offered in Sec. 6.

# 2 System model

We consider a massive access scenario with a large number of users $N$ and a single base station. The channel, which is shared among all $N$ users, is divided into recurring *random access opportunities*. Each random access opportunity comprises a total of $n$ channel uses, which are split into $n_{\text{ul}}$ uplink channel uses followed by $n_{\text{dl}} = n - n_{\text{ul}}$ downlink channel uses for message acknowledgments. For simplicity, we will assume that the random access opportunities are independent, i.e., the activity of the users does not depend on the feedback that they receive. The authentication occurs between the uplink and downlink phases.

## 2.1 Uplink phase

We assume that the uplink phase follows the scheme proposed by Polyanskiy [3]. Accordingly, in each random access opportunity a random subset $\mathcal{U}$ of the $N$ users, comprising a fixed number of users $K$, transmit their messages $\mathcal{W} = \{W_k\}_{k=1}^{K}$. The messages are drawn independently and uniformly from the set of messages $\mathcal{M} = \{1, 2, \ldots, M\}$. All users share the same encoder

$$f : \mathcal{M} \to \mathcal{X}^{n_{\text{ul}}}, \tag{H.1}$$

where $\mathcal{X}$ is the set of (complex) signals that can be transmitted during a single channel use. Each codeword $\mathbf{x}_k$ produced by the encoder is subject to the energy constraint $\|\mathbf{x}_k\|^2 \leq n_{\text{ul}}P$ where $P$ is the average energy per symbol. The users use the encoder as $\mathbf{x}_k = f(W_k)$ to produce the codewords $\{\mathbf{x}_k\}_{k=1}^{K}$, which are transmitted over a memoryless channel

$$P_{Y|X_1^K} : \mathcal{X}^{n_{\text{ul}} \times K} \to \mathcal{Y}^{n_{\text{ul}}}, \tag{H.2}$$

which is also permutation-invariant, i.e., $P_{Y|X_1^K}(\mathbf{y}|\mathbf{x}_1, \ldots, \mathbf{x}_K) = P_{Y|X_1^K}(\mathbf{y}|\mathbf{x}_{\pi(1)}, \ldots, \mathbf{x}_{\pi(K)})$ for any $\mathbf{y} \in \mathcal{Y}$, $\{\mathbf{x}_k\}_{k=1}^{K}$ and permutation $\pi$.

We consider the block-fading SIMO uplink channel where the BS is equipped with $M_A$ antennas. The signal received at the $m$-th antenna is

$$\mathbf{y}_m = \sum_{k=1}^{K} s_k \sqrt{g_k} h_{k,m} \mathbf{x}_k + \mathbf{z}_m. \tag{H.3}$$

Here, $s_k$ is the amplitude and phase controlled by user $k$, $g_k$ is the path loss, and $h_{k,m} \sim \mathcal{CN}(0,1)$ is the fading coefficient between the $k$-th user and the $m$-th antenna. The noise samples $\mathbf{z}$ are i.i.d and drawn from $\mathcal{CN}(\mathbf{0}, N_{0,ul})$.

Based on the received signals $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_{M_A}]$, the decoder $g\colon \mathcal{Y}^{M_A \times n_{\text{ul}}} \to \mathcal{M}^K$ outputs an unordered set $\widehat{\mathcal{W}} = \{\widehat{W}_k\}_{k=1}^{K}$ of exactly $K$ decoded messages.

Due to the influence of the channel, the set of decoded messages is not necessarily equal to the set of transmitted messages. Thus, we declare an error if a user transmitting message $W_k$ is not among the set of decoded messages. Similarly, we also declare an error if two or more users transmit the same message, This leads to the definition of an error event for user $k \in \mathcal{U}$ as $E_k = \{W_k \notin g(\mathbf{y})\} \cup \{W_k = W_i \text{ for some } i \neq k\}$. This error can be seen as a decoder *false negative* in the sense that a transmitted message was not outputted by the decoder. Because the decoder always outputs $K$ messages, a decoder false negative error implies a *decoder false positive* error, i.e., a message that was not transmitted by any user.

## 2.2 Authentication phase

Each of the decoded messages in $\widehat{\mathcal{W}}$ produced by the decoder undergoes an authentication step to determine the identity the senders $\widetilde{\mathcal{U}} \in \{[N]^{K'} | K' \leq K\}$. Formally, the authentication is a function

$$\tilde{f} : \mathcal{M}^K \to \{[N]^{K'} | K' \leq K\}. \tag{H.4}$$

Note that there may be some messages that cannot be authenticated, and as a result the authentication phase can output less (but not more) than $K$ identities. We define two types of errors for this function. First, a false negative error occurs if a transmitting user $k \in \mathcal{U}$ is not in the list of identities $\widetilde{\mathcal{U}}$, i.e., $\widetilde{E}_k^{\mathrm{FN}} = \{k \notin \widetilde{\mathcal{U}}\}$. The other error is a false positive and occurs when a user $\widehat{k} \notin \mathcal{U}$ that was *not* transmitting is in the list of identities, $\widetilde{E}_{\widehat{k}}^{\mathrm{FP}} = \{\widehat{k} \in \widetilde{\mathcal{U}}\}$.

## 2.3 Downlink phase

Following the uplink and authentication phases, from which the BS has obtained a set of $K'$ identities $\widetilde{\mathcal{U}}$, it broadcasts a common feedback message in the downlink, which informs the transmitting users about $\widetilde{\mathcal{U}}$. The feedback message is encoded using the encoder

$$\bar{f} : \{[N]^{K'} | K' \leq K\} \to \bar{\mathcal{X}}^{n_{\mathrm{dl}}}, \tag{H.5}$$

which is subject to power constraint $\|\bar{\mathbf{x}}\|_2^2 = n_{\mathrm{dl}}\bar{P}$. The feedback is transmitted over a memoryless broadcast channel

$$P_{\bar{Y}_1^K | \bar{X}} : \bar{\mathcal{X}}^{n_{\mathrm{dl}}} \to \bar{\mathcal{Y}}^{n_{\mathrm{dl}}}, \tag{H.6}$$

to the set of $K$ users that transmitted during the uplink phase (we assume that the other users are inactive during the entire random access opportunity).

Note that set $\mathcal{U}$ is not exactly known to the BS even after the uplink phase. Each of the $K$ active users decodes the feedback using its individual decoder

$$\bar{g}_k : \bar{\mathcal{Y}}^{n_{\mathrm{dl}}} \to \{0, 1\}, \tag{H.7}$$

where $\bar{g}_k(\bar{\mathbf{y}}_k) = 1$ if $k \in \widetilde{\mathcal{U}}$ and otherwise $\bar{g}_k(\bar{\mathbf{y}}_k) = 0$.

For the channel in the downlink we assume that the received signal at the $k$-th active user is given as

$$\bar{\mathbf{y}}_k = \left( \sum_{m=1}^{M_A} \sqrt{g_k} \bar{h}_{k,m} \bar{\mathbf{x}}_m \right) + \bar{\mathbf{z}}_k, \tag{H.8}$$

where $\bar{h}_{k,m} \sim \mathcal{CN}(0, 1)$ and $\bar{\mathbf{z}}_k \sim \mathcal{CN}(\mathbf{0}, N_{0,dl}\mathbf{I}_{n_{\mathrm{dl}}})$ are the fading coefficient and the additive noise, which are both independent of $\bar{\mathbf{x}}$. We assume that the path loss $g_k$ follows a log-distance path loss model such that $g_k = d_k^{-\alpha}$, where $d_k$ is the distance between the BS and user $k$ and $\alpha$ is the path loss exponent. Furthermore, we assume that the users are uniformly distributed in a $[d, D]$ annulus around the BS such that their distances are distributed as $f(d_k) = (2d_k)/(D^2 - d^2)$.

We define three events of errors in the downlink for a user $k$. The first is when a user fails to decode the feedback, which happens if the rate supported by the channel is less than the transmission rate $R_{dl}$. Neglecting finite blocklength effects, which are negligible for Rayleigh fading with the blocklengths that we will consider [7], this event is defined as $\bar{E}_k^{\mathrm{out}} = \{\log_2(1 + \bar{P}|\bar{h}_k|^2) < R_{dl}\}$. The other two error events are false positive acknowledgment and false negative acknowledgment, defined as $\bar{E}_k^{\mathrm{FP}} = \{k \notin \widetilde{\mathcal{U}} \cap \bar{g}_k(\bar{\mathbf{y}}_k) = 1\}$ and $\bar{E}_k^{\mathrm{FN}} = \{k \in \widetilde{\mathcal{U}} \cap \bar{g}_k(\bar{\mathbf{y}}_k) = 0\}$, respectively. Combining all error types, we obtain the event that the user correctly decodes the feedback message as $\bar{E}_k = \bar{E}_k^{\mathrm{out}} \cup \bar{E}_k^{\mathrm{FP}} \cup \bar{E}_k^{\mathrm{FN}}$.

The overall scheme and its steps is depicted in Fig.H.1.

# 3 Characterization of uplink and authentication error events

In this section we will analyze step-by-step the possible outcomes of the uplink transmissions of the packets followed by the decoding and authentication at the BS.

## 3.1 Uplink transmissions and decoding

We will investigate two distinct scenarios. In the first one, the BS is assumed to have a single antenna, $M_A = 1$, and each uplink phase is preceded by the

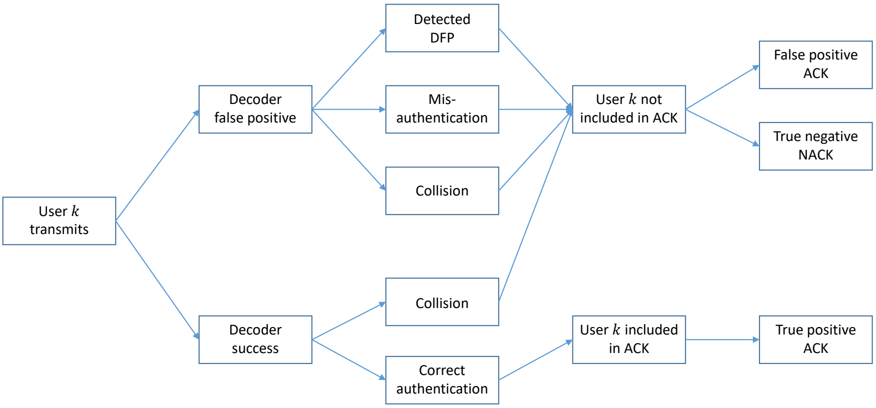


**Fig. H.2:** The possible outcomes of the combined uplink, authentication and downlink phases.

downlink beacon that allows users to estimate and invert their channels, i.e. $s_k = (h_{k,1}\sqrt{g_k})^{-1}$. As a result, the BS observes a simple AWGN channel. The error probability and, consequently, decoder false positive probability in such case can be obtained from random coding bound proposed in [3, eq. (3)-(10)].

In the second scenario, we assume that the BS is equipped with $M_A > 1$ antennas, but that there is no downlink beacon that the users can use to estimate and invert their channels. Thus, the users can only adapt their power to the path loss $g_k$, i.e., $s_k = 1/\sqrt{g_k}$. As a consequence, the individual signals are subject to the small-scale (Rayleigh) fading at the receiver. Furthermore, users have to rely on transmitting the pilot symbols along with the data so that the channel estimation can be performed by the BS. For this scenario, we will assume that the scheme presented in [8] is used to carry out the uplink transmissions and decoding. It can be shortly summarized as follows. First, each transmitting user chooses one of the $2^J$ non-orthogonal pilot sequences based on the first $J$ data bits (out of $B$) it wants to transmit. These are transmitted over the first $n_p$ channel uses and allow the BS to estimate the channels, as well as recover the implicitly encoded $J$ data bits. In the second step the remaining $B - J$ bits are transmitted over the $n_{ul} - n_p$ channel uses. The achievable error probability of such scheme (and consequently, false positive) can be obtained through normal approximation, i.e.,

$$p_{fp} = Q\left(\sqrt{\frac{n_{ul} - n_p}{2V}}\left(\log_2(1 + \mathrm{SINR}) - \frac{B - J}{n_{ul} - n_p}\right)\right) \tag{H.9}$$

where $Q(\cdot)$ is the Q-function, $V = \frac{\mathrm{SINR}}{2}\frac{\mathrm{SINR}+1}{(\mathrm{SINR}+1)^2}\log_2^2 e$ is the channel dispersion, and SINR is the effective SINR obtained after maximum ratio combining

(MRC) of signals and can be approximated as

$$\mathrm{SINR} = \frac{M_A(1-\sigma_{CE}^2)P}{N_{0,ul} + \sigma_{CE}^2 P + (K-1)P} \tag{H.10}$$

where $\sigma_{CE}^2 \geq N_{0,ul}/(N_{0,ul} + n_p P)$ is the lower bound on the MSE of the channel estimation assuming orthogonal pilots.

## 3.2 Identification and authentication

In order to enable identification and authentication capabilities, some additional bits need to be appended to the message and sent along with its information bits. The traditional way of achieving this is to add an address field, enabling identification, and message authentication code (MAC), ensuring integrity of the message by protecting it from accidental errors, as well as tampering and impersonation attempts. However, as demonstrated in [5], by constructing MAC in a specific way, namely by making it a function of the sender's identity, it is possible to omit the address field and still provide the identification feature with high reliability. This is especially desirable in the massive access scenarios, such as the one considered here, where the messages are typically very short and any type of overhead (such as address) accounts for large portion of the packet.

Throughout this paper we will rely on the following scheme. Each user $i$ possesses a unique, secret key $k_i$ known only by that user and the BS[1]. Whenever, a user has data to send $\mathbf{d}_i$, it will generate the MAC of length $L$ $\mathbf{m}_i \in \{0,1\}^L$ based on that data and its secret key as $\mathbf{m}_i = h(\mathbf{d}_i, \mathbf{k}_i)$. The function $h(\cdot)$ is common and known to all the communicating parties. Furthermore, it should be computationally hard to invert and produce sequences that have low collision probability (such functions are known as hashing functions). In this work, to simplify the analysis, we will assume an ideal hashing function whose outputs are uniformly distributed, i.e. the probability that a tuple $(\mathbf{d}_i, \mathbf{k}_i)$ produces an arbitrary MAC is $p = 1/2^L$. Once the MAC $\mathbf{m}_i$ is computed, the user concatenates it with the data and together they constitute the message $W_i = [\mathbf{d}_i, \mathbf{m}_i]$ as described in Section 2.

At the BS, upon decoding the messages $\widehat{\mathcal{W}}$, the receiver proceeds with the identification and authentication procedure. For each message $\widehat{W}_i = [\widehat{\mathbf{d}}_i, \widehat{\mathbf{m}}_i]$ on the list, the receiver does the following. First, it extracts the part corresponding to the data bits $\widehat{\mathbf{d}}_i$. Then, it recomputes the MACs using keys of

[1] We do not go into details of how to distribute such secret keys, however this is a very well-studied problem that can be addressed with, e.g. public-key cryptography. The distribution of keys can also be incorporated into the initial registration phase that is performed whenever the user attaches to the BS. This procedure, however, is relatively infrequent and the keys themselves do not need to be updated, hence its impact on the overall scheme is minimal.

different users stored in its database as $\mathbf{m}'_{i,j} = h(\widehat{\mathbf{d}}_i, \mathbf{k}_j)$ for $j = 1, 2, \ldots, N$. Each MAC produced this way is compared with the one in the decoded message and as long as there is only a single match (no collisions), i.e. $\widehat{\mathbf{m}}_i = \mathbf{m}'_{i,j}$ for some $j$, the receiver declares that the message is valid and comes from user $j$. This procedure is performed for each message produced by the decoder and can easily be parallelized to decrease the total computation time.

It is important to note that, unlike the address, MAC might not be a unique identifier which can lead to mis-identification and MAC collisions. More formally, for a given $[\widehat{\mathbf{d}}_i, \widehat{\mathbf{m}}_i]$ tuple, there might exist one or more matching key $\mathbf{k}_l \neq \mathbf{k}_j$ s.t. $h(\widehat{\mathbf{d}}_i, \mathbf{k}_l) = h(\widehat{\mathbf{d}}_i, \mathbf{k}_j) = \widehat{\mathbf{m}}_i$. Let us analyze the possible outcomes, which are also summarized in Fig.H.2.

First, let us assume that the decoded message $\widehat{W}_i$ is genuine, i.e. a true positive $\widehat{W}_i \in \mathcal{W}$. In that case, there is at least one matching key - the one that corresponds to the actual sender. If there are no others, receiver authenticates the message and correctly concludes the identity of the sender. If there are more matching keys, the receiver is not able to tell who the sender is and will not authenticate the message nor acknowledge it later on. As such, we have:

$$p_{s\_id} = (1-p)^{N-1} \tag{H.11}$$

$$p_{coll\_tp} = 1 - p_{s\_id} \tag{H.12}$$

On the other hand, if the decoded message is a false positive $\widehat{W}_i \notin \mathcal{W}$, then there are the following three possibilities. In case the receiver checks all the keys and none of them produces a matching MAC (most likely outcome), then it is able to detect that the message is, in fact, a decoder false positive and discard it. The probability of this happening is

$$p_{d\_fp} = (1-p)^N \tag{H.13}$$

Similarly, the message will not be authenticated if there is more than one matching key, which occurs with probability

$$p_{coll\_fp} = \sum_{i=2}^{N} \binom{N}{i} p^i (1-p)^{N-i} = 1 - (1-p)^N - Np(1-p)^{N-1} \tag{H.14}$$

In the worst case scenario, there is a single matching key which will cause the receiver to incorrectly accept the message and lead to mis-authentication. This event has probability

$$p_{mis\_id\_fp} = Np(1-p)^{N-1} \tag{H.15}$$

# 4 Characterization of downlink error events

Following the uplink and authentication phases, the BS multicasts a packet that acknowledges the user identities $\widetilde{\mathcal{U}}$, whose messages were decoded and authenticated. Because of its multicast nature, the BS cannot simply invert the channel. Furthermore, since the identities of senders are only estimates at this point, it has to transmit with enough power to ensure that all the devices in its range can decode the downlink packet. For simplicity, we will assume a very simple scheme where the BS transmits with the power $P_{dl} = \beta D^{\alpha}$ divided equally between its $M_A$ antennas and that some form of full-diversity space-time block coding is applied, such that the expected received power at a user at the edge of the network is equal to $\beta$. In general, the power received by user $k$ is $\beta D^{\alpha} d_k^{-\alpha}$ and the SNR follows a gamma distribution $f_{gam}(x; M_A, \frac{D^{\alpha} d_k^{-\alpha}}{M_A N_{0,dl}})$. Note that in the special case of $M_A = 1$, the received SNR is exponentially distributed with mean $\beta D^{\alpha} d_k^{-\alpha}$. By marginalizing over the distribution of $d_k$ we obtain the outage probability as a function of the downlink transmission rate $R_{dl}$:

$$\Pr(\bar{E}_k^{\text{out}} \mid R_{dl}) = \int_d^D F_{gam}\left(R_{dl}; M_A, \frac{\beta D^{\alpha} d_k^{-\alpha}}{M_A N_{0,dl}}\right) \frac{2d_k}{D^2 - d^2} \mathrm{d}d_k, \tag{H.16}$$

where $F_{gam}(\cdot)$ is the cumulative distribution function of the gamma distribution.

Because multiple users need to be acknowledged and the transmitting users are unknown to the BS, the acknowledgment message needs to address the users that should be acknowledged. We will consider two methods of encoding the message acknowledgments that are transmitted by the BS in the downlink. The first is to simply concatenate the identities of the decoded users, while the other is a more efficient encoding based on Bloom filers, which has the cost of introducing false positive acknowledgments. We will assume that the feedback message encodes the number of decoded users $K'$, which requires a field of at most $\lceil \log_2(N) \rceil$ bits.

We start by analyzing the concatenation based scheme. Since the identities of each of the $K'$ active users can be encoded using $\lceil \log_2(N) \rceil$ bits, the total number of bits required for the acknowledgment message is simply

$$B_{\text{concat}} = \lceil \log_2(N) \rceil + K' \lceil \log_2(N) \rceil, \tag{H.17}$$

where the first term comes from encoding the number of decoded users $K'$. Clearly, provided that the message is correctly decoded, this encoding scheme introduces no false negatives or no false positives, so the only error event is outage, i.e. $\bar{E}_k = \bar{E}_k^{\text{out}}$. Provided that the number of symbols in the downlink is $n_{\text{dl}}$, the required rate is

$$R_{dl,\text{concat}} = \frac{\lceil \log_2(N) \rceil + K' \lceil \log_2(N) \rceil}{n_{\text{dl}}}, \tag{H.18}$$

which plugged into Eq. (H.16) allows to determine the outage probability.

We now turn our attention to the Bloom filter encoding [6]. To encode the set of decoded user identities $\widetilde{\mathcal{U}}$, the Bloom filter uses $\ell$ independent hash functions $l_1(k), \ldots, l_\ell(k)$ that map from the set of user identities $\{0, 1, \ldots, N - 1\}$ to $\{0, 1, \ldots, B - 1\}$ representing the bits in the message. We will assume that the hash functions are ideal, i.e., they uniformly map each element to a random element in $\{0, 1, \ldots, B - 1\}$. The feedback message is constructed by setting the bits at indices $\cup_{k \in \widetilde{\mathcal{U}}} \cup_{l=1,\ldots,\ell} l_l(k)$ equal to '1' and the remaining bits to '0'. By checking whether a user with identity $n$ is in the message, it simply checks whether all bit indices resulting from the $\ell$ hash functions applied to its own ID are equal to '1' in the received message (we refer to these bits as *test bits*).

Because the outputs of the hash functions may overlap between different users (i.e. are subject to collisions), there is a risk of false positives, but no false negatives. A false positive happens if all test bits of a user that was transmitting but was not decoded/authenticated are equal to '1' in the acknowledgment message. The false positive probability depends on $\ell$, $K'$, and $B$. It has been shown that the false positive probability is minimized when approximately $\ell = \ln(2)(B/K')$ [6], in which case the false positive probability is

$$\Pr(\bar{E}_k^{\text{FP}}) = (1/2)^{\ln(2)(B/K')}. \tag{H.19}$$

Note that in practice $\ell$ needs to be an integer, which introduces a small penalty in the performance. Neglecting this and rearranging, we obtain that the number of bits required to achieve a false positive probability of at most $\epsilon_{\text{FP}}$ is

$$B_{\text{bloom}} = \lceil \log_2(N) \rceil + K' \frac{\log_2(1/\epsilon_{\text{FP}})}{\ln(2)}. \tag{H.20}$$

Repeating the calculation of the rate as done for concatenation, we obtain the rate

$$R_{dl,\text{bloom}} = \frac{\lceil \log_2(N) \rceil + K' \frac{\log_2(1/\epsilon_{\text{FP}})}{\ln(2)}}{n_{\text{dl}}}. \tag{H.21}$$

Since the false positive event is only relevant when the packet is decoded and there are no false negatives, we obtain the total error probability for the Bloom filter encoded acknowledgments as

$$\Pr(\bar{E}_k) = \Pr(\bar{E}_k^{\text{out}}) + (1 - \Pr(\bar{E}_k^{\text{out}})) \Pr(\bar{E}_k^{\text{FP}}). \tag{H.22}$$

# 5 Results

As could be seen from the previous sections, due to the multiple phases, each with its own set of possible outcomes, it is not immediately obvious what

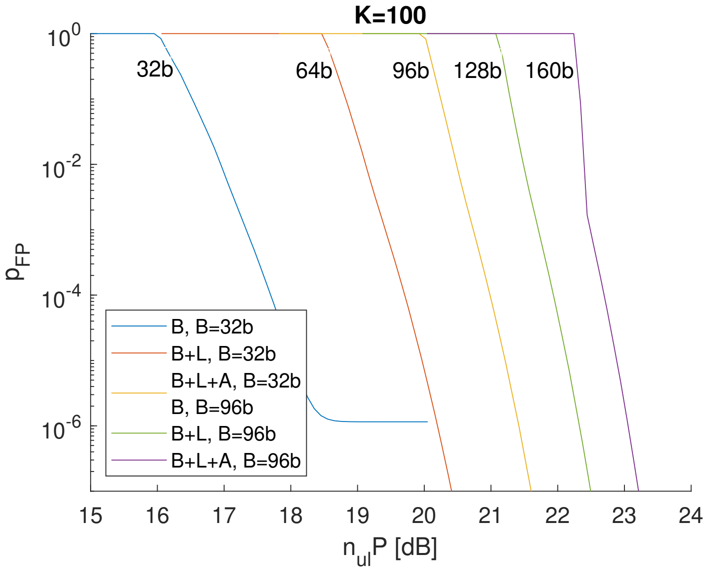


**Fig. H.3:** The probability of decoder false positive (physical layer error) for the AWGN case. The number of channel uses is $n_{ul} = 2^{15}$.

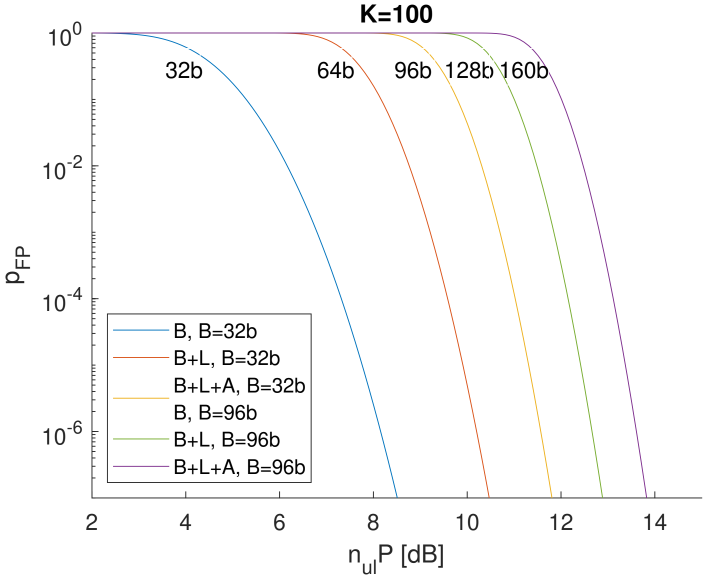
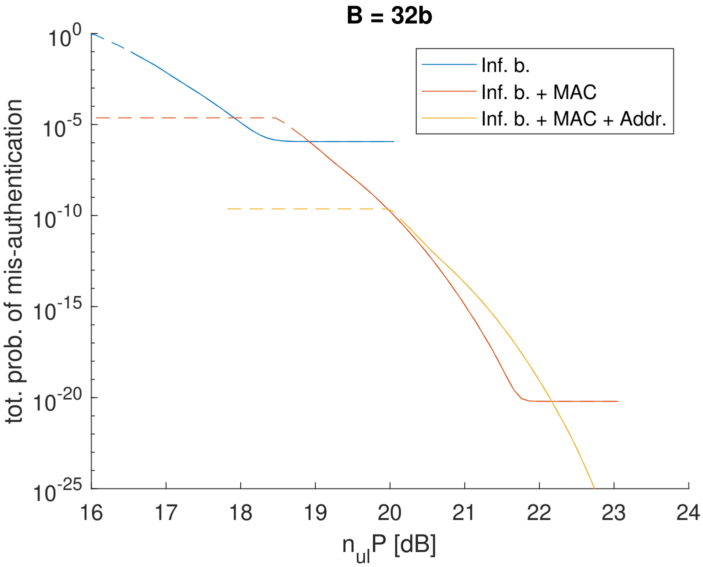


**Fig. H.4:** The probability of decoder false positive for the uplink SIMO scenario with fading. The number of antennas is $M_A = 32$. The number of channel uses is $n_{ul} = 4800$ out of which $n_p = 960$ serve as pilots.

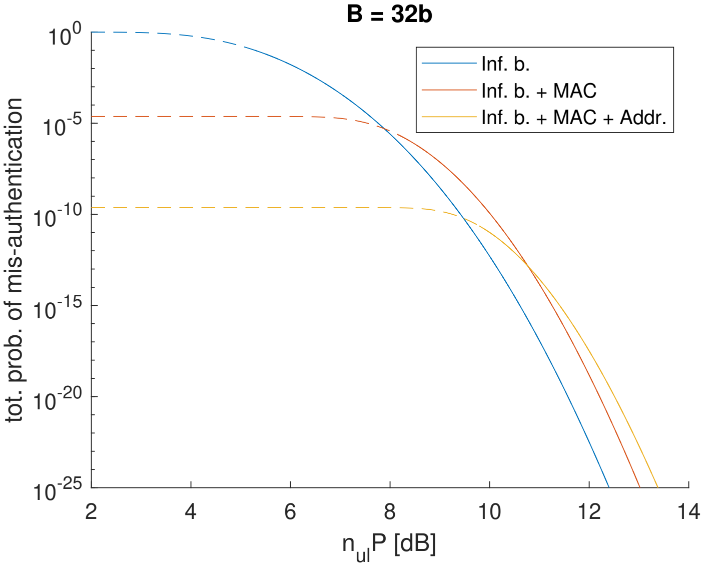
exactly constitutes an "error" or a "success". As such, we will be investigating the performance in terms of different types of events. Throughout this section we will be comparing the following three methods in the uplink: a) the "pure" URA where the packet consists solely of $B$ data bits, b) proposed method where the packet contain additionally $L$ MAC bits, c) traditional technique where in addition to data and MAC the packet contains also the explicit address for a total of $B + L + A$ bits. Unless otherwise stated we set $L = 32$ b, $A = 32$ b and the total population of users is $N = 10^5$. The number of antennas in the fading case is $M_A = 32$. We also set $d = 20$ m, $D = 150$ m, $\alpha = 3$.

We start by looking into the probability of mis-authentication, i.e. the probability that the BS accepts the incorrectly decoded packet, which is the most severe type of error. This phenomenon is dependent only the the decoder and authentication performance, hence it can be analyzed independently from the ACK procedure. Clearly, in the pure URA scheme, every packet produced by the decoder has to be accepted since there are no other means of verifying it. As such, the probability of mis-authentication is equal to the probability of decoder false positive $p_{mis\_auth} = p_{fp}$. In the proposed scheme, mis-authentication can only occur if the packet is a decoder false positive, and there is exactly one key that accidentally produces a matching MAC, hence $p_{mis\_auth} = p_{fp} p_{mis\_id\_fp}$. Similar is true for the traditional scheme with full address, however the authenticator module checks only a single key - the one that corresponds to the address explicitly given in the packet, so $p_{mis\_auth} = p_{fp} p$. In Fig. H.3 and H.4 we depict the decoder false positive probability as a function of the total transmitted energy $n_{ul} P$ for different packet sizes and fixed number of channel uses $n_{ul}$. Different curves represent the degradation in performance due to addition of auxiliary bits, e.g. if $B = 96$ b, then the third curve from the left corresponds to the pure URA scheme, fourth (128 b) corresponds to data bits and MAC while the rightmost to the traditional scheme with data, MAC and address. Based on those results and the preceding discussion, in Fig. H.5 we show the total probability of mis-authentication that depends on both physical layer errors (false positive probability) and mis-identification events. Although the pure URA scheme tends to outperform the others it is important to remember that it does not allow to identify the transmitters, which might disqualify it in many applications. On the other hand, the proposed scheme tends to perform better than the traditional one. This might seem surprising, considering that the lack of the address entails much higher probability of mis-identification. However, the reduction of the overhead leads to lower effective rate which in turn increases the probability of successful decoding.

In Fig.H.6 we compare the performance of the acknowledgment procedures based on the concatenation and Bloom filter. We do that, by examining the total probability of success which involves simultaneous: successful up-



(a) AWGN



(b) Fading

**Fig. H.5:** Total probability of mis-authentication. The dashed lines denote the range where $p_{FP} > 10\%$, i.e. despite the low mis-authentication probability significant portion of the packets is lost.

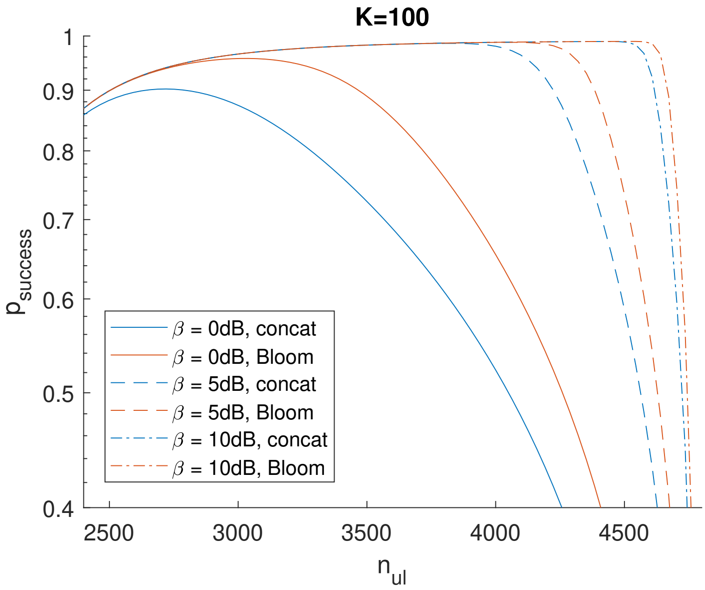


**Fig. H.6:** The comparison between acknowledgments based on concatenation and Bloom filter. The probability of false positive ACK is set to $\epsilon_{\mathrm{FP}} = 10^{-3}$ and $N = 10^6$.

link transmission, authentication, and reception of the downlink acknowledgment. The results are based on the proposed scheme ($B = 32$ b and $L = 32$ b) and are generated as follows. First, we fix the total power in the uplink $nP$ such that the decoder false positive probability is $\approx 1\%$ for $n_{ul} = n$. Then, keeping this power fixed we vary the number of channel uses in the uplink (x-axis) and dedicate the rest for the downlink acknowledgment. Based on the eq. (H.18) and (H.21) we are able to determine the rate of downlink transmission, which in turn determines the probability of its successful decoding (through eq.(H.16)). The results are obtained for three different values of $\beta$, i.e. the mean power at the cell edge. We can see that implementing the acknowledgments based on the Bloom filter can offer sizable gains compared to the simple concatenation. For each curve, there exists and optimal operating point, i.e. the division between uplink and downlink channel uses, which offers highest total probability of success.

## 6 Conclusions

In this work we have demonstrated a scheme that is suitable for supporting a massive, randomly activated population of devices. The proposed solution combines the URA-based uplink, identification and authentication with reduced overhead, and efficient joint acknowledgments that are broadcasted in the downlink. Compared to the "pure" URA, this scheme re-introduces the capability to identify the users, which in many applications is a necessity. We also show how to further extend the baseline procedure and improve its reliability by integrating it with jointly encoded acknowledgments. The work put forward here presents several interesting directions for future investigations including joint optimization of uplink and downlink, and introduction of retransmission techniques. The latter seems particularly interesting since in addition to further increasing the reliability, retransmissions would allow to implement mechanisms capable of detecting and rectifying false positives.

# References

[1] N. Abramson, "THE ALOHA SYSTEM: another alternative for computer communications," in *Proceedings of the November 17-19, 1970, fall joint computer conference*. ACM, pp. 281-285, 1970.

[2] X. Chen, T.-Y. Chen, and D. Guo, "Capacity of Gaussian Many-Access Channels," *IEEE Transactions on Information Theory*, vol. 63, no. 6, pp. 3516–3539, 2017.

[3] Y. Polyanskiy, "A perspective on massive random-access," in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527.

[4] G. Durisi, T. Koch, and P. Popovski, "Toward massive, ultrareliable, and low-latency wireless communication with short packets," *Proceedings of the IEEE*, vol. 104, no. 9, pp. 1711–1726, Sept 2016.

[5] R. Kotaba, A. E. Kalør, P. Popovski, I. Leyva-Mayorga, B. Soret, M. Guillaud, and L. G. Ordóñez, "How to Identify and Authenticate Users in Massive Unsourced Random Access," *IEEE Communications Letters*, vol. 25, no. 12, pp. 3795–3799, 2021.

[6] A. Broder and M. Mitzenmacher, "Network applications of bloom filters: A survey," *Internet mathematics*, vol. 1, no. 4, pp. 485–509, 2004.

[7] W. Yang, G. Durisi, T. Koch, and Y. Polyanskiy, "Quasi-static SIMO fading channels at finite blocklength," in *2013 IEEE International Symposium on Information Theory*. IEEE, 2013, pp. 1531–1535.

[8] A. Fengler, P. Jung, and G. Caire, "Pilot-Based Unsourced Random Access with a Massive MIMO Receiver in the Quasi-Static Fading Regime," in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2021, pp. 356–360.

# Paper I

## Common Message Acknowledgments: Massive ARQ Protocols for Wireless Access

Anders E. Kalør, Radosław Kotaba and Petar Popovski

The paper has been submitted to
IEEE Transactions on Communications, *arXiv:2201.03907*

*The layout has been revised.*

# Abstract

*Massive random access plays a central role in supporting the Internet of Things (IoT), where a subset of a large population of users simultaneously transmit small packets to a central base station. While there has been much research on the design of protocols for massive access in the uplink, the problem of providing message acknowledgments back to the users has been somewhat neglected. Reliable communication needs to rely on two-way communication for acknowledgement and retransmission. Nevertheless, because of the many possible subsets of active users, providing acknowledgments requires a significant amount of bits. Motivated by this, we define the problem of massive ARQ (Automatic Retransmission reQuest) protocol and introduce efficient methods for joint encoding of multiple acknowledgements in the downlink. The key idea towards reducing the number of bits used for massive acknowledgement is to allow for a small fraction of false positive acknowledgments. We analyze the implications of this approach and the impact of acknowledgment errors in scenarios with massive random access. Finally, we show that these savings can lead to a significant increase in the reliability when retransmissions are allowed since it allows the acknowledgment message to be transmitted more reliably using a much lower rate.*

***Keywords***— Automatic repeat request, feedback, internet of things, massive random access

# 1 Introduction

A fundamental challenge in supporting the Internet of Things (IoT) is to enable grant-free, or uncoordinated, transmissions from a very large number of users [1]. Furthermore, as the user activation is often triggered by physical phenomena, such as an event that generates sensory data, the traffic patterns are sporadic. Thus, at any instant, the resulting subset of active user that have something to transmit is random. This has initiated a large amount of research devoted to the design of random access schemes that can decode messages from a small random subset of users, often based on techniques derived from ALOHA [2, 3] or compressed sensing [4, 5].

However, despite the great interest in transmission schemes for massive access, the problem of efficiently providing packet reception acknowledgments to a large number of users has been somewhat neglected. Yet, a message acknowledgment is often an useful signal for the application layer, and is necessary in order to implement retransmission schemes, which can greatly increase the transmission reliability. Furthermore, several transmission schemes directly rely on such feedback in order to achieve high performance, e.g., by using rateless codes [2, 6]. Although these schemes require only a single bit of common feedback, it can be beneficial in practice to provide early feedback as soon any individual user is decoded in order to minimize the interference from imperfect SIC. In essence, our work treats the problem of massive ARQ (Automatic Retransmission reQuest) and thus expands the problem space of the area of massive wireless access.

Compared to grant-based access scenarios, where the BS can send an acknowledgment to a user using a single bit (ACK/NACK), acknowledging a set of users decoded from a grant-free access scenario requires the BS to encode the user identities or some other information that can be used to identify the users that it wants to acknowledge. Encoding the user identifiers requires a significant number of bits when the number of users is large. A naïve attempt to encode acknowledgments to $K$ users out of a total of $N$ users could be to simply concatenate the identifiers of the $K$ users and transmit an acknowledgment packet of $K \log_2(N)$ bits. However, this approach has two significant drawbacks. First, it requires a variable-length packet, which may not be desired in many protocols that rely on time-division multiplexing. Second, as we will show, it is possible to significantly reduce the number of bits required to encode the acknowledgments by applying source coding techniques to jointly encode the acknowledgments for all $K$ users. A similar idea was exploited in [7] to design feedback for collision-free scheduling of $K$ out of $N$ users succeeding a massive random access scenario. However, they assumed that the uplink was error-free, which makes the use of acknowledgments obsolete in the first place.

In order to achieve substantial reductions in the acknowledgment message length, our key proposal is to allow for a small but non-negligible fraction of *false positive* acknowledgments, i.e., that a transmitting user erroneously determines that its message is among the acknowledged messages. Such errors are atypical in existing systems, which are often designed to suppress false positives using error detection mechanisms such as cyclic redundancy checks (CRCs), or by encoding the feedback message such that false positives are very rare at the cost of a larger *false negative* probability [8]. The reason for this is that false positive acknowledgments remain undetected after a transmission round and thus can be hard to resolve and lead to unreliable communication. This is in contrast to false negative errors, which may for instance occur if there are errors in the CRC but the message is intact, and for which the cost is merely an unnecessary retransmission. In this sense, a false positive acknowledgment can be "fatal" as it leads to the situation where the user believes that its message was successfully received by the BS when it in fact was lost.

The practical consequences of false positive acknowledgments depend on the application. Applications that require high reliability are likely to be severely impacted by even a small fraction of false positive acknowledgments. On the other hand, in exemplary IoT applications, such as sensing or monitoring, false positive acknowledgments may result only in missed sensor readings because failed measurement transmissions will not be retransmitted, which is unlikely to have big consequences. However, if such events cannot be tolerated, false positive acknowledgments can be detected and subsequently resolved using mechanisms at higher layers such as packet numbering at the cost of a detection latency.

The impact of false positives and false negatives in feedback has been studied thoroughly for automatic repeat requests (ARQ) and hybrid automatic repeat requests (HARQ) in the single-user setting, where only a single-bit acknowledgment message is needed. The general conclusion from these studies is that the probability of false positive acknowledgments needs to be significantly smaller than the uplink error probability, since they, contrary to the uplink and a false negatives, cannot be repaired by a retransmission [9, 10]. The same result holds in the finite blocklength regime, where the downlink message should be designed to achieve low false positive probability, but the false negative probability should be held constant and relatively large independently of the total reliability requirement [8]. Although the reliability of the feedback is generally less important when the maximum number of transmissions is small since the uplink reliability plays a more significant role in determining the total reliability, these results hold even with as few as two transmission rounds [10]. Nevertheless, because of the large feedback message required in massive access regime and the fact that the feedback acknowledges multiple users, these results cannot be directly transferred to the scenario that we consider.

The paper has three main contributions. *First,* it is the core idea of allowing false positives. We show that by allowing a small fraction of false positive acknowledgments, the number of bits required for the feedback message can be significantly reduced, while the introduction of false negative acknowledgment does not yield comparable savings. Furthermore, we present various practical methods for efficient encoding of acknowledgments with false positives. *Second,* we study how the distribution of the number of active users impacts the feedback message, and derive closed-form bounds on the false positive probability based on the first and second moments of the distribution. *Third,* we quantify the impact of false positive acknowledgments on the overall reliability by studying transmission schemes with multiple transmission rounds. In this context, we show that the message length reduction that results from introducing false positives allows the feedback to be transmitted with a much lower rate, which in turn results in a significant increase in the overall reliability.

We note that, in both grant-free and grant-based settings and irrespectively of the feedback encoding, feedback can be designed either in an adaptive or non-adaptive manner. Adaptive feedback schemes are intrinsically non-trivial due to the half-duplex structure of most wireless systems, which requires the feedback instants to be either fully pre-planned or controlled by the transmitting user. In this paper, the focus is on the feedback message and assume that the feedback moments are known.

The remainder of the paper is organized as follows. Appendix 2 introduces the overall system model. Information-theoretic bounds for a fixed number of decoded users are presented in Appendix 3, and Appendix 4 introduces and analyzes a number of practical encoding schemes for this setting. Appendix 5 analyzes the case in which the number of decoded users is random, and the case with multiple transmission rounds is analyzed in Appendix 6. Finally, numerical results are presented in Appendix 7 and the paper is concluded in Appendix 8.

## 2 System Model

We consider a typical massive access scenario comprising a single base station (BS) that serves a massive set of potentially active users $[N] = \{1, 2, \ldots, N\}$ (typically $N$ is in the order of thousands). As is often the case in practical systems, we assume that each of the $N$ users has a unique identifier known to both the users and the BS. If the BS requires an initial handshake procedure for users to join the network, then $N$ corresponds to the number of users associated with the BS, and $N$ will be in the order of thousands (for instance in NB-IoT, the Cell Radio Network Temporary Identifier (C-RNTI) can identify up to $N = 65523$ users [11]). On the other hand, if no such procedure exists, then each user can have a globally unique identifier, such as a MAC address, and $N$ will be in the order of $2^{32}$ to $2^{64}$.

We assume a general frame structure in which the air interface is divided into a number of recurring random access opportunities in which a random subset $\mathcal{A} \subseteq [N]$ of users are active and transmit their messages in the uplink. The uplink transmission is followed by a downlink feedback message, multicasted by the BS, that provides acknowledgments to the users that the BS decoded in the uplink. Users that receive an acknowledgment have completed their transmission, while users that do not receive an acknowledgment are allowed to retransmit up to $L - 1$ times. We assume that each uplink message contains the transmitter's identifier such that the BS is able to determine the identity of the sender upon decoding of the packet[1]. In general, the number of active users is random and typically will be much smaller than $N$. The set of active users (including its cardinality) is unknown to the BS, which tries to recover it from the received signals. We denote the set of recovered users by $\mathcal{S} = \{s_1, s_2, \ldots, s_K\}$, where $s_k \in [N]$ and assume that, conditioned on $K$, $\mathcal{S}$ is drawn uniformly from the set of all $K$-element subsets of $[N]$, denoted $[N]_K = \{\mathcal{K} \subseteq [N] \mid |\mathcal{K}| = K\}$. Due to decoding errors, $\mathcal{S}$ may be different

[1] We make this assumption for clarity of presentation, but the analysis holds even if there is no identity (e.g., as in unsourced random access [12]) by treating the messages as temporary identities. In that case $N$ corresponds to the number of distinct messages.

from the actual set of active users $\mathcal{A}$. We denote by $\epsilon_{\mathrm{ul},n}$ the probability that a transmitting user $n$ is not decoded. This probability typically depends on the random access mechanisms as well as the value of $K$, the signal-to-noise ratios (SNRs) of the transmitting users, etc.

To make the transmitters aware of potential errors and to ensure reliable transmission, the BS transmits a common $B$-bit feedback message after the random access opportunity that allows the users to determine whether their own identifier is a member of $\mathcal{S}$. The message is transmitted through a packet erasure channel so that the packet is received by user $n$ with probability $1-\epsilon_{\mathrm{dl},n}$. The erasure probability depends on the SNR of the individual users, the channel, and the transmission rate of the feedback[2]. Formally, such a feedback scheme is defined by an encoder, the downlink channel, and a set of decoders, one for each user. We define the encoder as

$$f : [N]_K \to \{0,1\}^B, \tag{I.1}$$

and the erasure channel as

$$\Pr(Y_n = X \mid X \in \{0,1\}^B) = 1 - \epsilon_{\mathrm{dl},n}, \tag{I.2}$$

$$\Pr(Y_n = \mathrm{e} \mid X \in \{0,1\}^B) = \epsilon_{\mathrm{dl},n}, \tag{I.3}$$

where $X$ is the packet transmitted by the BS, $Y_n$ is the packet received by user $n$, and e denotes an erasure. Finally the individual decoders are defined as

$$g_n : \{0,1\}^B \cup \mathrm{e} \to \{0,1\}, \qquad n = 1, \dots, N \tag{I.4}$$

which output 1 if user $n$ is believed to be a member of $\mathcal{S}$ and 0 otherwise (throughout the paper we will assume that the decoder outputs 0 if it observes an erasure). Both the encoder and the decoders may depend on $K$ (which is random), but this dependency can be circumvented by encoding $K$ separately in the feedback message at an average cost of approximately $H(K)$ bits where $H(\cdot)$ is the entropy function[3]. As we will see, this overhead is minimal when compared to the number of bits required to encode the acknowledgments in most settings of practical interest. We will refer to $B$ as the message length of a scheme.

To characterize the performance of a feedback scheme, we define the false positive (FP) probability, denoted $\epsilon_{\mathrm{fp}}$, as the probability that a user whose

[2] In practice, an erasure channel represents the case where the decoder can detect if a packet is decoded incorrectly, e.g., through an error-detecting code. The size of such a code with negligible false positive probability is small compared to the size of the feedback message, and thus we will ignore the overhead it introduces.

[3] A pragmatic alternative when the activation distribution is unknown would be to assume that at most $K'$ users can be decoded simultaneously and then dedicate a fixed number of $\log_2(K')$ bits to describe $K$.

uplink message was not decoded $n \in \mathcal{A} \setminus \mathcal{S}$ erroneously concludes that it belongs to $\mathcal{S}$

$$\epsilon_{\mathrm{fp}} = \mathbb{E}\left[\Pr\left(g_n\left(f\left(\mathcal{S}\right)\right) = 1 \mid n \in \mathcal{A} \setminus \mathcal{S}\right) \mid K\right], \tag{I.5}$$

where the expectation is taken over $n$ and the distribution $p(\mathcal{S}|K)$ (but not the channel, which we will treat independently). Similarly, we define the false negative (FN) probability $\epsilon_{\mathrm{fn}}$ as the probability that a decoded user $n \in \mathcal{S}$ incorrectly concludes that it does not belong to the set

$$\epsilon_{\mathrm{fn}} = \mathbb{E}\left[\Pr\left(g_n\left(f\left(\mathcal{S}\right)\right) = 0 \mid n \in \mathcal{S}\right) \mid K\right]. \tag{I.6}$$

Note that these definitions ignore the channel, and thus allow us to treat $\epsilon_{\mathrm{fp}}$ and $\epsilon_{\mathrm{fn}}$ independently of the event of an erasure. Note also that both $\epsilon_{\mathrm{tp}}$ and $\epsilon_{\mathrm{fp}}$ are conditioned on $K$. We discuss the case when $K$ is random further in Appendix 5.

Using these definitions, we denote by $B^*$ the minimum message length $B$ required for a scheme with $K$ active users out of $N$ that achieves false positive and false negative probabilities at most $\epsilon_{\mathrm{fp}}$ and $\epsilon_{\mathrm{fn}}$, respectively.

# 3 Information Theoretic Bounds

We first consider the source coding part of the problem, namely the functions $f$ and $g_n$ defined previously, while for clarity ignoring the erasure channel between the BS and the users. Specifically, in this section we derive information theoretic bounds on the minimum message length $B$ required for the feedback message. To start with, we treat $K$ as constant, and thus neglect the bits required to encode $K$ in the message, which would be the same for all schemes.

## 3.1 Error-Free Coding

We first consider error-free schemes, i.e., schemes that have $\epsilon_{\mathrm{fp}} = \epsilon_{\mathrm{fn}} = 0$. A naïve construction of the feedback message is to concatenate the $K$ identifiers in $\mathcal{S}$ to produce a message of $B = K \log_2(N)$ bits. However, such a construction is sub-optimal because there are only $\binom{N}{K}$ subsets of $K$ users, and $\log_2 \binom{N}{K}$ bits are sufficient to distinguish each subset. This leads to the feedback message length

$$B^*_{\text{error-free}} = \left\lceil \log_2 \binom{N}{K} \right\rceil \tag{I.7}$$

$$\geq \left\lceil K \log_2(N/K) \right\rceil, \tag{I.8}$$

where the inequality follows from $\binom{N}{K} \geq (N/K)^K$.

A message length of $B^*_{\text{error-free}}$ can be achieved using e.g., enumerative source coding [13]. However, the decoding is impractical for large sets because each user needs to check each of the $\binom{N-1}{K-1}$ subsets that it can belong to.

## 3.2 Encoding with Bounded Errors

The required feedback message length for the error-free encoding scales with the logarithm of $N$, which can be significant when $N$ is in the order of $2^{32}$ or $2^{64}$. One way to reduce the impact of $N$ is to allow for non-zero false positive and false negative probabilities. To do so, a feedback message must acknowledge at most $K + \epsilon_{\text{fp}} N$ users, out of which at least $(1 - \epsilon_{\text{fn}})K$ must be in $\mathcal{S}$. For $\epsilon_{\text{fp}} < 0.5$ (which is typically the region of interest), it can be shown using combinatorial arguments that [14] (see Appendix A for details)

$$B^*_{\text{fp,fn}} \geq \log_2 \binom{N}{K} - \log_2 \left( K \binom{\lfloor \epsilon_{\text{fp}} N \rfloor + K}{\lceil (1-\epsilon_{\text{fn}})K \rceil} \binom{N}{\lfloor \epsilon_{\text{fn}} K \rfloor} \right) \tag{I.9}$$

$$\begin{aligned} &\geq K \log_2 \left( \frac{1}{\epsilon_{\text{fp}} + \frac{K}{N}} \right) - K \log_2 \left( \frac{e}{1 - \epsilon_{\text{fn}}} \right) \\ &\quad - \epsilon_{\text{fn}} K \log_2 \left( \frac{1 - \epsilon_{\text{fn}}}{\epsilon_{\text{fn}} \left( \epsilon_{\text{fp}} + \frac{K}{N} \right)} \right) - \log_2(K), \end{aligned} \tag{I.10}$$

where (I.10) follows from the observation that rounding cannot decrease the message length and the inequality $\left(\frac{N}{K}\right)^K \leq \binom{N}{K} \leq (eN/K)^K$. Note that if $K$ is held constant, the bound is independent of $N$ as $N \to \infty$.

The introduction of false positives has the potential to offer significantly greater gains than false negatives. In particular, when $\epsilon_{\text{fn}}$ is small as is typically desired, the required message length is only negligibly smaller than the one required if no false negatives were allowed. The reason for this is that the set of potential false negatives, $\mathcal{S}$, is much smaller than the set of potential false positives $[N] \setminus \mathcal{S}$. When $\epsilon_{\text{fn}} = 0$, the bound can be tightened further as [15, 16]

$$B^*_{\text{fp}} \geq K \log_2 \left( 1/\epsilon_{\text{fp}} \right) - \frac{\log_2(e)(1 - \epsilon_{\text{fp}})K^2}{\epsilon_{\text{fp}} N + (1 - \epsilon_{\text{fp}})K}, \tag{I.11}$$

where the last term vanishes as $N \to \infty$.

A (non-constructive) achievability bound for the case with $\epsilon_{\text{fn}} = 0$ and $K \leq N\epsilon_{\text{fn}}$ was provided in [15]. The overall idea is to sequentially generate all $\lfloor N\epsilon_{\text{fp}} \rfloor$-element subsets of $[N]$, and then transmit the index of the

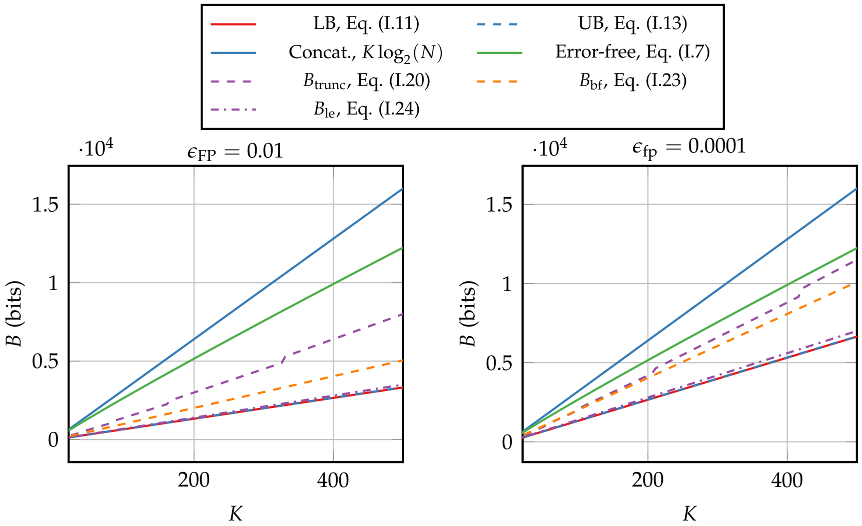


**Fig. I.1:** Message length, $B$, required to provide acknowledgment feedback for $N = 2^{32}$ and $\epsilon_{\text{fn}} = 0$ with $\epsilon_{\text{fp}} = 0.01$ and $\epsilon_{\text{fp}} = 0.0001$.

first subset that includes all $K$ elements in $\mathcal{S}$. Using a set-cover theorem by Erdős and Spencer [17, Theorem 13.4], they show that the number of $\lfloor N\epsilon_{\text{fp}} \rfloor$-element subsets required to cover all $K$-element subsets of $[N]$, denoted $M(N, \lfloor N\epsilon_{\text{fp}} \rfloor, K)$, is upper bounded by

$$M(N, \lfloor N\epsilon_{\text{fp}} \rfloor, K) \leq \left(1 + \ln \binom{\lfloor N\epsilon_{\text{fp}} \rfloor}{K}\right) \frac{\binom{N}{K}}{\binom{\lfloor N\epsilon_{\text{fp}} \rfloor}{K}}. \tag{I.12}$$

Taking the logarithm and bounding the binomial coefficients gives the following upper bound on the required feedback message length

$$\begin{aligned} B^*_{\text{fp}} &\leq \log_2 \binom{N}{K} - \log_2 \binom{\lfloor N\epsilon_{\text{fp}} \rfloor}{K} + \log_2 \left(1 + \ln \binom{\lfloor N\epsilon_{\text{fp}} \rfloor}{K}\right) && \text{(I.13)} \\ &\leq K \log_2 \left(e/\epsilon_{\text{fp}}\right) + \log_2 \left(1 + K \ln \left(\frac{N\epsilon_{\text{fp}}}{K}\right)\right). && \text{(I.14)} \end{aligned}$$

Note that this also serves as an upper bound for the case with $\epsilon_{\text{fn}} > 0$. By comparing (I.14) to the lower bound in (I.11), it can be seen that the bounds are tight within an additive term $O(\log N)$ as $N \to \infty$, i.e., for sufficiently large $N$,

$$B^*_{\text{fp}} = K \log_2(1/\epsilon_{\text{fp}}) \pm O(\log N), \tag{I.15}$$

which is lower than the error-free scheme in Eq. (I.8) when $\epsilon_{\text{fp}} \geq K/N$. To illustrate the potential gain of introducing a small fraction of false positives, suppose $N = 2^{32}$ and $K = 100$. Encoding the acknowledgment in an error-free manner requires approximately $B = \log_2 \binom{2^{32}}{100} \approx 2675$ bits, while only $B = 100 \log_2(100) \approx 664$ bits are required if we can tolerate $\epsilon_{\text{fp}} = 0.01$, and $B = 100 \log_2(10000) \approx 1329$ bits for $\epsilon_{\text{fp}} = 0.0001$. The required feedback message lengths for these cases are shown in Fig. I.1 and compared to a number of practically realizable schemes presented next. As expected, the upper (UB) and lower (LB) bounds are very tight (within 14 bits in the considered range).

# 4 Practical Schemes

In this section, we present a number of practical designs of $f$ and $g_n$, and compare them to the bounds derived in the previous section. Motivated by the fact that false negatives provide little reduction in the feedback message length, we will restrict ourselves to schemes with $\epsilon_{\text{fn}} = 0$. Furthermore, we will again assume that the number of decoded users $K$ is fixed and defer the discussion of random activations to Appendix 5.

## 4.1 Identifier Truncation

We start by considering a simple truncation scheme in which the feedback message is constructed by first truncating the identifiers of each of the $K$ decoded users to $b < \log_2(N)$ bits (say, the $b$ least significant bits), and then concatenating them to construct a feedback message of $Kb$ bits. To check whether a user with identifier $n$ is among the $K$ decoded users, one simply checks whether the $b$ least significant bits of $n$ is contained in the message. Clearly, while this cannot cause false negatives, it can lead to false positives if a user that is not decoded in the uplink shares the same $b$ least significant bits with a decoded user. Assuming that the identifiers are uniformly distributed, the false positive probability is

$$\epsilon_{\text{fp}} = 1 - \left(1 - \frac{1}{2^b}\right)^K . \tag{I.16}$$

By rearranging and ceiling to ensure $b$ is integer we obtain $b = \left\lceil -\log_2 \left(1 - (1 - \epsilon_{\text{fp}})^{1/K}\right) \right\rceil$. The feedback message length is then bounded by

$$
\begin{aligned}
B_{\text{trunc}} &= K\left\lceil -\log_2\left(1-(1-\epsilon_{\text{fp}})^{\frac{1}{K}}\right)\right\rceil && \text{(I.17)}\\
&\geq K\left\lceil -\log_2\left(1-e^{-\frac{\epsilon_{\text{fp}}}{K(1-\epsilon_{\text{fp}})}}\right)\right\rceil && \text{(I.18)}\\
&\geq K\left\lceil -\log_2\left(\frac{\epsilon_{\text{fp}}}{K(1-\epsilon_{\text{fp}})}\right)\right\rceil && \text{(I.19)}\\
&= K\left\lceil \log_2\left(1/\epsilon_{\text{fp}}\right)+\log_2\left(K(1-\epsilon_{\text{fp}})\right)\right\rceil, && \text{(I.20)}
\end{aligned}
$$

where the first inequality follows from $1-x \geq e^{-\frac{x}{1-x}}$ for $0 \leq x < 1$ and that $-\log_2(1-x)$ is monotonically increasing for $x < 1$, and the second inequality is due to $1-e^{-x} \leq x$ for $x \geq 0$ and that $-\log_2(x)$ is monotonically decreasing. The last term in Eq. (I.20) is strictly positive when $K > \frac{1}{1-\epsilon_{\text{fp}}}$, which is the case for the values of $K$ and $\epsilon_{\text{fp}}$ that we are interested in. Thus, the scheme requires approximately $K\log_2(K(1-\epsilon_{\text{fp}}))$ bits more than the lower bound in Eq. (I.15).

## 4.2 Universal Hashing

The downside of the identifier truncation scheme presented earlier is that it requires the identifiers to have high entropy, which may be difficult to guarantee in practice. To circumvent this, we can *hash* the identifier instead of using the identifier directly. To illustrate, we consider a scheme based on universal hashing, which can be implemented efficiently in practice. Formally, an $(n, v)$-family of universal hash functions is a family of functions $h : [n] \to [v]$ such that for a hash function $h$ chosen uniformly at random and for any two distinct values $x, y \in [n]$, $\Pr(h(x) = h(y)) \leq 1/v$. The event that $h(x) = h(y)$ is typically referred to as a *collision*. Using this assumption, we can concatenate the hash of each of the $K$ users to construct a message of $K2^v$ bits. The probability that the hash of an arbitrary user that is not among the $K$ decoded users collides with any of the decoded users is

$$
\epsilon_{\text{fp}} = 1-\left(1-\frac{1}{v}\right)^K, \tag{I.21}
$$

which is exactly the same as in the previous section, but does not require that the user identities are uniformly distributed, and thus is a practically appealing alternative to identifier truncation. However, the required number of bits is still quite far from the lower bound.

## 4.3 Bloom Filter

A Bloom filter [18] uses $T$ independent universal hash functions $h_i : [N] \to [B]$ for $i = 1, \dots, T$, and is constructed by setting the message bits at positions $\{h_i(s_k) \mid s_k \in \mathcal{S}, i = 1, \dots, T\}$ equal to '1' and the remaining bits equal to '0'. In order to decode the message and check whether an identifier $n$ belongs to the set, the decoder simply checks if the bits at positions $\{h_i(n) \mid i = 1, \dots, T\}$ are equal to '1'. Clearly, the decoder can only observe false positives and not false negatives.

It can be shown that the minimum false positive probability is obtained when the probability that a given bit is '1' is exactly $1/2$, and that $T$ should be chosen as $T = (B/K)\ln(2)$ to achieve this [19] (in practice, one needs to round to the nearest integer). The resulting false positive probability is non-trivial, but can be approximated as [19]

$$\epsilon_{\text{fp}} \approx 2^{-\lceil (B/K)\ln(2)+0.5 \rceil}. \tag{I.22}$$

By assuming equality in the approximation we obtain

$$B_{\text{bf}} = K \log_2(e) \log_2(1/\epsilon_{\text{fp}}), \tag{I.23}$$

revealing that Bloom filter is approximately within a factor $\log_2(e) \approx 1.44$ of the asymptotic lower bound in Eq. (I.15). Nevertheless, it is better than the previous two schemes approximately when $K > \frac{\epsilon_{\text{fp}}^{1-\log_2(e)}}{1-\epsilon_{\text{fp}}}$.

## 4.4 Linear Equations

An alternative family of constructions is based on solving a set of linear equations in a Galois field, first proposed in [16, 20]. To simplify the analysis, we will assume that we have access to a fully random hash function. An $(n, b)$-family of fully random hash functions is a family of functions $h : [n] \to [b]$ such that for each value $x \in [n]$, it outputs a value chosen uniformly at random from $[b]$. While such hash functions have desirable properties, they are not practical as they require an exponential number of bits to store. Nevertheless, in many practical problems the fully random hash function can be replaced by a simpler hash function with a negligible penalty, especially when the input is randomized [21].

Returning to the encoding scheme, suppose we have a fully random hash function $h_1 : [N] \to \text{GF}(2^{\lceil \log_2(1/\epsilon_{\text{fp}}) \rceil})^K$, i.e., mapping from $[N]$ to $K$-element vectors in $\text{GF}\left(2^{\lceil \log_2(1/\epsilon_{\text{fp}}) \rceil}\right)$, and a universal hash function $h_2 : [N] \to 2^{\lceil \log_2(1/\epsilon_{\text{fp}}) \rceil}$. Then, we can construct the equation $h_1(s_k) \cdot z = h_2(s_k)$ in $\text{GF}(2^{\lceil \log_2(1/\epsilon_{\text{fp}}) \rceil})$, where $\cdot$ is the inner product. By constructing an equation for each $s_k \in \mathcal{S}$, we obtain the set of $K$ equations with $K$ variables $H_1 z = h_2$, where $H_1 \in \mathrm{GF}(2^{\lceil \log_2(1/\epsilon_{\mathrm{fp}})\rceil})^{K\times K}$ is the matrix of rows $h_1(s_1),\ldots,h_1(s_K)$ and $h_2 \in \mathrm{GF}(2^{\lceil \log_2(1/\epsilon_{\mathrm{fp}})\rceil})^{K}$ is the vector with elements $h_2(s_1),\ldots,h_2(s_K)$. In order for this system to have a solution, we require $H_1$ to be full rank. It can be shown that this happens with probability at least $1-\frac{1}{2^{\lceil \log_2(1/\epsilon_{\mathrm{fp}})\rceil}-1}$ [20], which is large for the values of $\epsilon_{\mathrm{fp}}$ that we consider (e.g., greater than 0.99 for $\epsilon_{\mathrm{fp}} = 0.01$ and greater than 0.9999 for $\epsilon_{\mathrm{fp}} = 0.0001$). By repeating the procedure with new hash functions $h_1', h_1'', \ldots$, the probability of generating a matrix with full rank can be made arbitrarily large at the cost of a message length penalty required to store the number of trials. In practice, this penalty is negligible compared to the total size of the message. For instance, with $\epsilon_{\mathrm{fp}} = 0.01$ and up to 16 trials, requiring only four additional bits, the failure probability is in the order of $10^{-34}$.

Provided that the resulting matrix $H_1$ has full rank, we can obtain the solution $z$ to the set of equations. A decoder can then check whether an identifier $n$ is contained in the set by simply checking if $h_1(s_k)\cdot z = h_2(s_k)$. Thus, neglecting the potential overhead caused by repeating the hashing procedure, only the vector $z$ needs to be communicated, which contains $K$ entries of $\lceil \log_2(1/\epsilon_{\mathrm{fp}})\rceil$ bits each. Combining these observations, we obtain the feedback message length

$$B_{\mathrm{le}} = K\lceil \log_2(1/\epsilon_{\mathrm{fp}})\rceil, \tag{I.24}$$

which, disregarding the rounding, matches the asymptotic information theoretic bound in Eq. (I.15). As we will see in Appendix 7, the practical performance matches closely with the bound.

It is worth noting that finding $z$ uses Gaussian elimination, which requires $O(K^3)$ operations. This makes the method infeasible for large $K$. However, the operation can be performed fast as long as $K$ is at most in the order of hundreds, which is the main interest in this paper. When $K$ is larger, the construction can be improved by introducing sparsity in $H_1$ at the cost of a small overhead, see e.g., [16, 20, 22].

# 5 Random User Activity

So far, we have assumed that $K$ is fixed and optimized the feedback for a specific value of $K$. In practice, the number of active devices is random and unknown to both the BS and the devices, and thus the number of messages produced by the random access decoder at the BS, $K$, is in general also random. Furthermore, $K$ may even be correlated over time and depend on the feedback itself. However, to simplify the analysis, we will here assume that $K$ is independent across frames and drawn from the distribution $p(K)$.

The optimal designs of the feedback schemes depend on $K$ and in many of the schemes the recipient must know $K$ to be able to decode the message.

Hence, it is reasonable to include the value of $K$ in the feedback message, which incurs only a very small overhead. To illustrate, suppose the random access mechanism is designed to support at most $K' = 1024$ simultaneously active users, then a 10-bit overhead is required to encode $K$. On the other hand, if the desired false positive probability is $\epsilon_{\text{fp}} = 0.001$, then approximately $\log_2(0.001) \approx 10$ bits are required per user in the feedback message, so the overhead introduced by encoding $K$ merely corresponds to encoding one additional user. When more than $K'$ users are active, we accept that the error probability can be arbitrarily high. In that case, we may decide to pick a random subset comprising $K'$ of the $K > K'$ decoded users at the cost of $K - K'$ false negatives.

If we allow the feedback message to have a variable length, then we can achieve the desired $\epsilon_{\text{fp}}$ (and $\epsilon_{\text{fn}}$) as long as $K \leq K'$ without significant overprovisioning when $K < K'$ by adjusting the message length to $K$. However, the random user activity has a more significant impact on the performance when the length of the feedback message needs to remain fixed for any value of $K$, e.g., due to protocol constraints, as the error probabilities depends on the instantaneous value of $K$. It seems reasonable in this case to optimize message length either based on the average false positive/negative probabilities or by the probability that the these probabilities exceed some thresholds $\tilde{\epsilon}_{\text{fp}}$ and $\tilde{\epsilon}_{\text{fn}}$. Assuming for clarity that $\epsilon_{\text{fn}} = 0$, we can formalize the first case by defining the message length selection rule

$$B = \inf\left\{B' \geq 0 : \mathbb{E}_{K\sim p(K)}[\epsilon_{\text{fp}}(K, B')] \leq \tilde{\epsilon}_{\text{fp}}\right\}, \tag{I.25}$$

where $\epsilon_{\text{fp}}(K, B')$ is the false positive probability achieved with $K$ users and a message length of $B'$ bits, and $\tilde{\epsilon}_{\text{fp}}$ is the specified false positive probability target. Similarly, for the second case we have

$$B = \inf\left\{B' \geq 0 : \Pr(\epsilon_{\text{fp}}(K, B') > \tilde{\epsilon}_{\text{fp}}) \leq \delta\right\}, \tag{I.26}$$

where $\delta$ specifies the maximum allowed probability that the false positive probability exceeds $\tilde{\epsilon}_{\text{fp}}$.

Computing these feedback message lengths requires complete knowledge of the distribution of $K$, which is often not available. Instead, we proceed by deriving bounds based on the first moments of $p(K)$ using the asymptotic expression for the feedback message length given in Eq. (I.15) for $\epsilon_{\text{fn}} = 0$, which is accurate for large $N$. We first present an upper bound on the expected false positive $\bar{\epsilon}_{\text{fp}} = \mathbb{E}_{K\sim p(K)}[2^{-B/K}]$.

**Proposition 1.** *Let the number of decoded users $K$ be random with mean $\bar{K} = \mathbb{E}[K]$ and variance* $\mathrm{Var}[K]$. *Then for a given message length $B$ the expected false positive* *probability $\bar{\epsilon}_{\text{fp}}$ is upper bounded as*

$$\bar{\epsilon}_{\text{fp}} < 2^{-B/\bar{K}} + \frac{2.66\text{Var}[K]}{B^2}. \tag{I.27}$$

*Proof.* By rearranging the expression in Eq. (I.15) we obtain $\epsilon_{\text{fp}}(K, B) \approx 2^{-B/K}$. To bound $\bar{\epsilon}_{\text{fp}}$, we consider the first-order Taylor expansion of $2^{-B/K}$ around $\bar{K} = \mathbb{E}[K]$ given as

$$2^{-B/K} =$$
$$2^{-B/\bar{K}} + \frac{2^{-B/\bar{K}}}{\bar{K}^2}(K - \bar{K}) + \frac{2^{-B/Z} B \ln(2)\,(B\ln(2) - 2Z)}{Z^4}\frac{(K-\bar{K})^2}{2}, \tag{I.28}$$

for some $Z$ between $\bar{K}$ and $K$. By analyzing its derivatives it can be shown that the term $\frac{2^{-B/Z} B \ln(2)(B\ln(2)-2Z)}{Z^4}$ is bounded and attains its maximum at $Z = \frac{\ln(2)(3-\sqrt{3})}{6}B$. From this we obtain the bound

$$\frac{2^{-B/Z}\,(B\ln(2) - 2Z)}{Z^4} \leq \frac{e^{-\frac{6}{3-\sqrt{3}}}\left(\frac{6}{3-\sqrt{3}} - 2\right)}{\left(\frac{3-\sqrt{3}}{6}\right)^3 \ln^2(2) B^2} \tag{I.29}$$
$$= \frac{\zeta}{B^2}, \tag{I.30}$$

where $\zeta = e^{-\frac{6}{3-\sqrt{3}}}\left(\frac{6}{3-\sqrt{3}} - 2\right)\left(\frac{6}{3-\sqrt{3}}\right)^3 \ln^{-2}(2)$. By inserting into Eq. (I.28), taking expectation and rearranging we obtain

$$\bar{\epsilon}_{\text{fp}} \leq 2^{-B/\bar{K}} + \frac{\zeta\text{Var}[K]}{2B^2}. \tag{I.31}$$

The proof is completed by noting that $\zeta/2 < 2.66$. □

The result in Proposition 1 can be used to select the feedback message length according to the rule in Eq. (I.25). We now derive a similar general bound on the probability that $\epsilon_{\text{fp}}$ exceeds $\tilde{\epsilon}_{\text{fp}}$ that can be used for the alternative feedback message length selection rule in Eq. (I.26).

**Proposition 2.** *Let the number of decoded users $K$ be random with mean $\bar{K} = \mathbb{E}[K]$ and variance* Var$[K]$*. For a given message length $B$ and $\tilde{\epsilon}_{\text{fp}} \geq 2^{-B/\bar{K}}$, the probability that the false positive probability $\epsilon_{\text{fp}}$ exceeds $\tilde{\epsilon}_{\text{fp}}$ can be bounded as*

$$\Pr\left(\epsilon_{\text{fp}} > \tilde{\epsilon}_{\text{fp}}\right) \leq \frac{\text{Var}[K]}{\left(\frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})} - \bar{K}\right)^2} \tag{I.32}$$

*Proof.* Note first that

$$\Pr\left(\epsilon_{\text{fp}} > \tilde{\epsilon}_{\text{fp}}\right) = \Pr\left(K > \frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})}\right) \tag{I.33}$$

$$\leq \Pr\left(K \geq \frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})}\right). \tag{I.34}$$

Applying Chebyshev's inequality yields

$$\Pr\left(K \geq \frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})}\right) \leq \Pr\left(|K-\bar{K}| \geq \frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})} - \bar{K}\right) \tag{I.35}$$

$$\leq \frac{\operatorname{Var}[K]}{\left(\frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})} - \bar{K}\right)^2}. \tag{I.36}$$

□

Because the bound in Proposition 2 does not assume much about the distribution of $K$, it is in general not very tight. If we further assume that the users activate independently (but not necessarily identically distributed), we can tighten the bound as follows.

**Proposition 3.** *Let the number of decoded users $K = \sum_{i=1}^{N} k_i$ where $k_i \in \{0,1\}$ are independently Bernoulli random variables with $\Pr(k_i = 1) = p_i$, and let $\bar{K} = \mathbb{E}[K] = \sum_{i=1}^{N} p_i$. For a given feedback message length $B$ and $\tilde{\epsilon}_{\text{fp}} \geq 2^{-B/\bar{K}}$, the probability that the false positive probability $\epsilon_{\text{fp}}$ exceeds $\tilde{\epsilon}_{\text{fp}}$ can be bounded as*

$$\Pr\left(\epsilon_{\text{fp}} > \tilde{\epsilon}_{\text{fp}}\right) < \Pr\left(\frac{e^{\left(\eta_{\tilde{\epsilon}_{\text{fp}}}-1\right)}}{\left(\eta_{\tilde{\epsilon}_{\text{fp}}}\right)^{\eta_{\tilde{\epsilon}_{\text{fp}}}}}\right)^{\bar{K}}, \tag{I.37}$$

*where $\eta_{\tilde{\epsilon}_{\text{fp}}} = B/\left(\bar{K}\log_2(1/\tilde{\epsilon}_{\text{fp}})\right)$.*

*Proof.* As in Proposition 2 we have

$$\Pr\left(\epsilon_{\text{fp}} > \tilde{\epsilon}_{\text{fp}}\right) \leq \Pr\left(K \geq \frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})}\right). \tag{I.38}$$

Defining $\eta_{\tilde{\epsilon}_{\text{fp}}} = B/\left(\bar{K}\log_2(1/\tilde{\epsilon}_{\text{fp}})\right)$ and applying the Chernoff bound for Poisson trials (see e.g., Theorem 4.4 in [23]), we obtain

$$\Pr\left(K \geq \frac{B}{\log_2(1/\tilde{\epsilon}_{\text{fp}})}\right) = \Pr\left(K \geq \eta_{\tilde{\epsilon}_{\text{fp}}}\bar{K}\right) \tag{I.39}$$

$$< \Pr\left(\frac{e^{\left(\eta_{\tilde{\epsilon}_{\text{fp}}}-1\right)}}{\left(\eta_{\tilde{\epsilon}_{\text{fp}}}\right)^{\eta_{\tilde{\epsilon}_{\text{fp}}}}}\right)^{\bar{K}}, \tag{I.40}$$

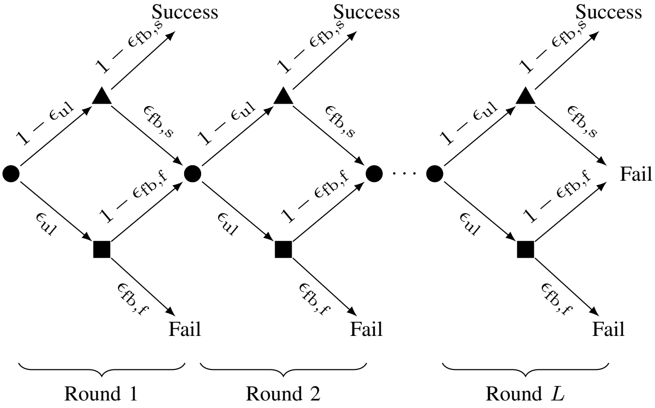

Success
Success
Success
$1-\epsilon_{\mathrm{fb,s}}$
$1-\epsilon_{\mathrm{ul}}$
$\epsilon_{\mathrm{fb,s}}$
$1-\epsilon_{\mathrm{fb,f}}$
$\epsilon_{\mathrm{ul}}$
$\epsilon_{\mathrm{fb,f}}$
Fail
Fail
Fail
Fail
Round 1
Round 2
Round $L$

**Fig. I.2:** Events in the case with $L$ transmission rounds. The circles represent the start of a round, triangles represent the case when a packet is successfully decoded by the BS, and the squares are when the packets are not decoded by the BS. A success occurs when a packet is both decoded by the BS and the user decodes the acknowledgment.

which completes the proof. □

We remark that although these bounds are based on the asymptotic bound from Eq. (I.15), similar bounds can be obtained for the practical schemes by first bounding the rounding error. For instance, for the scheme based on linear equations, we have $B_{\mathrm{le}} = K\lceil\log_2(1/\epsilon_{\mathrm{fp}})\rceil \leq K(\log_2(1/\epsilon_{\mathrm{fp}})+1)$, which is straightforward to bound using the same methodology as in Propositions 1 to 3.

# 6 Random Access with Feedback and Retransmissions

In this section, we analyze the impact of feedback in a scenario with $L$ transmission rounds, each comprising an uplink and a downlink phase. We first analyze the problem with packet erasure channels in both uplink and downlink, and then extend the analysis to a richer channel model that allows us to characterize the trade-off between false positives/negatives in the feedback message and the transmission rate. We assume that the transmission rounds are independent, and that the channel erasure probabilities are the same in all rounds.

## 6.1 Packet Erasure Channels

We consider the transmission scenario from the perspective of a single user and assume an uplink erasure probability $\epsilon_{\mathrm{ul}}$, downlink erasure probability $\epsilon_{\mathrm{dl}}$, false positive probability $\epsilon_{\mathrm{fp}}$, and false negative probability $\epsilon_{\mathrm{fn}}$, which are the same in all transmission rounds. Under these conditions, the transmission process is illustrated in Fig. I.2, where $\epsilon_{\mathrm{fb,s}} = 1 - (1 - \epsilon_{\mathrm{dl}})(1 - \epsilon_{\mathrm{fn}})$ and $\epsilon_{\mathrm{fb,f}} = (1 - \epsilon_{\mathrm{dl}})\epsilon_{\mathrm{fp}}$ are the probabilities that the user makes a wrong decision based on the feedback, conditioned on success or failure in the uplink, respectively. We assume that the user succeeds only if it received an acknowledgment for the packet transmitted in the same round, i.e., a false positive acknowledgment is a failure even if the uplink was successful in a previous round but the downlink in that round was unsuccessful. Furthermore, the user retransmits if it is unable to decode the feedback. The failure probability is then given as

$$\Pr(\mathrm{fail}) = 1 - \sum_{l=1}^{L} \ell^{l-1}(1 - \epsilon_{\mathrm{ul}})(1 - \epsilon_{\mathrm{fb,s}}), \tag{I.41}$$

where $\ell = \epsilon_{\mathrm{ul}}(1 - \epsilon_{\mathrm{fb,f}}) + (1 - \epsilon_{\mathrm{ul}})\epsilon_{\mathrm{fb,s}}$ is the probability that the user proceeds from one transmission round to the next. To gain some insight into the behavior, suppose first that $L = 1$, in which case the expression reduces to

$$\Pr(\mathrm{fail}) = 1 - (1 - \epsilon_{\mathrm{ul}})(1 - \epsilon_{\mathrm{dl}})(1 - \epsilon_{\mathrm{fn}}), \tag{I.42}$$

suggesting that $\epsilon_{\mathrm{ul}}$, $\epsilon_{\mathrm{dl}}$ and $\epsilon_{\mathrm{fn}}$ have equal importance in minimizing the failure probability. Furthermore, because false positives can only occur when the uplink fails, in which case the entire transmission fails since $L = 1$, the failure probability is independent of the false positive probability $\epsilon_{\mathrm{fp}}$. Similarly, suppose now that we allow an infinite number of retransmissions. By taking the limit $L \to \infty$ in Eq. (I.41) we obtain

$$\Pr(\mathrm{fail}) = 1 - \frac{(1 - \epsilon_{\mathrm{ul}})(1 - \epsilon_{\mathrm{fn}})}{1 - \epsilon_{\mathrm{fn}} - \epsilon_{\mathrm{ul}}(1 - \epsilon_{\mathrm{fn}} - \epsilon_{\mathrm{fp}})}. \tag{I.43}$$

Note that this expression depends on the uplink erasure probability and the false positive/negative probabilities, but not on the downlink erasure probability $\epsilon_{\mathrm{dl}}$, which is because an erasure in the downlink always will result in a retransmission. When $\epsilon_{\mathrm{fp}} = 0$ the expression reduces to $\Pr(\mathrm{fail}) = 0$ indicating that even with false negatives (but no false positives), an arbitrarily high reliability can be achieved by increasing the transmission rounds. On the other hand, when $\epsilon_{\mathrm{fn}} = 0$ the expression in Eq. (I.43) reduces to

$$\Pr(\mathrm{fail}) = 1 - \frac{1 - \epsilon_{\mathrm{ul}}}{1 - \epsilon_{\mathrm{ul}}(1 - \epsilon_{\mathrm{fp}})}, \tag{I.44}$$

suggesting that when $\epsilon_{\text{fp}} > 0$ retransmissions cannot fully compensate for an unreliable uplink channel. The intuition behind this is that an unreliable uplink increases the probability of receiving a false positive, which in turn increases the failure probability.

Note that the analysis above holds even when the number of users $K$ is random if the length of the feedback message and the transmission rate (channel coding rate) of the feedback message are adapted based on the instantaneous $K$ to match $\epsilon_{\text{dl}}$, $\epsilon_{\text{fp}}$, and $\epsilon_{\text{fn}}$. However, if the length of the feedback message remains fixed, $\epsilon_{\text{fp}}$ and $\epsilon_{\text{fn}}$ depends on the instantaneous $K$. In this case, a reasonable strategy is to use Eq. (I.41) to determine an appropriate $\epsilon_{\text{fp}}$, and then apply either Proposition 1 or Proposition 2 to select the feedback message length such that the target false positive probability is satisfied with the desired probability.

## 6.2 Source/Channel Coding Trade-off

In practice, the erasure probability of the downlink transmission, $\epsilon_{\text{dl}}$, is a function of the transmission rate and depends on the SNR at the receiver. Furthermore, for a given number of symbols transmitted over the channel, the rate depends on the length of the feedback message $B$, which directly impacts the false positive/negative probabilities. Consequently, there is an inherent trade-off between $\epsilon_{\text{dl}}$, $\epsilon_{\text{fp}}$ and $\epsilon_{\text{fn}}$, which determine the overall reliability of the system. To illustrate the trade-off, suppose we can construct a feedback message with $\epsilon_{\text{fn}} = 0$ and false positive probability $\epsilon_{\text{fp}}$ using $K \log_2(1/\epsilon_{\text{fp}})$ bits, and that we aim to transmit it over a quasi-static fading channel with additive noise and instantaneous SNR given by $\gamma$. For a given number of symbols $c$, the transmission rate is given as $K \log_2(1/\epsilon_{\text{fp}})/c$ and the probability of decoding error is thus

$$\begin{aligned}\epsilon_{\text{dl}} &= \Pr\left(\log_2(1+\gamma) < \frac{K \log_2(1/\epsilon_{\text{fp}})}{c}\right) && \text{(I.45)}\\ &= \Pr\left(\gamma < \epsilon_{\text{fp}}^{-K/c} - 1\right), && \text{(I.46)}\end{aligned}$$

illustrating, as expected, that decreasing $\epsilon_{\text{fp}}$ causes $\epsilon_{\text{dl}}$ to increase since a higher transmission rate is required.

If the number of symbols for the feedback message and the number of retransmissions $L$ are fixed, the transmission rate and the false positive/negative probabilities need to be jointly optimized to minimize the failure probability in Eq. (I.41). This is in general a non-convex optimization problem that requires numerical evaluation of Eq. (I.41) over a range of $\epsilon_{\text{fp}}$. The asymptotic expression in Eq. (I.43) suggests that when $L$ is large, we should aim at minimizing $\epsilon_{\text{fp}}$ and $\epsilon_{\text{fn}}$ since $\epsilon_{\text{dl}}$ has no impact on the failure probability. In

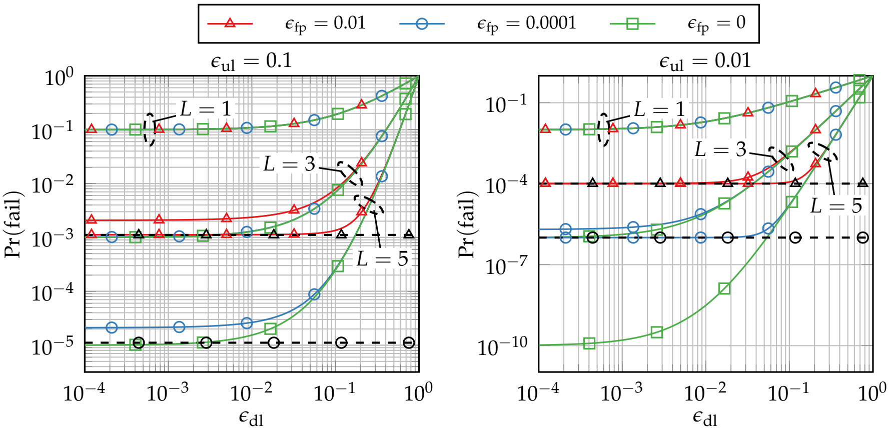


**Fig. I.3:** The probability that a transmission fails with $L$ rounds obtained using Eq. (I.41). The dashed lines show the asymptotic results for $L \to \infty$.

particular, in this case the introduction of false positives will lead to a worse performance compared to identifier concatenation, as there is no gain in reducing the length of the acknowledgment packet. However, when $L$ is small, the downlink erasure probability has an increasing impact since a successful downlink transmission is required to succeed. In particular, to minimize the failure probability for $L = 1$ the downlink probability and the false negative probability should be equal, while $\epsilon_{fp}$ has no impact, as can be seen in Eq. (I.42).

# 7 Numerical Results

In this section, we evaluate the feedback schemes in a typical massive random access setting. We first present results that illustrate the impact of false positives in the setting with multiple transmission rounds and fixed $K$ over a simple erasure channel. We then investigate the case with random $K$, and finally we exemplify the trade-off between allocating channel symbols for the uplink and the feedback under a random access channel in the uplink and a Rayleigh fading channel in the downlink. Except for the cases where it is explicitly mentioned, we will assume that $\epsilon_{fn} = 0$.

## 7.1 Fixed $K$ and $L$ Transmission Rounds

When $K$ is fixed, the false positive probability $\epsilon_{fp}$ is constant and can be picked arbitrarily by choosing an appropriate feedback message length $B$.

The probability that a transmission fails in a setting with $L$ retransmissions is shown in Fig. I.3 along with the asymptotic results for $L \to \infty$, obtained using Eqs. (I.41) and (I.44), respectively. Although the downlink erasure probability $\epsilon_{\mathrm{dl}}$ has no impact as $L \to \infty$, it has a significant impact when $L$ is small. In particular, for finite $L$ the failure probability is at least $(\epsilon_{\mathrm{dl}})^L$, as a successful downlink transmission is required in order for the user to succeed. Similarly, we can observe an error floor as $\epsilon_{\mathrm{dl}}$ approaches zero caused by both the false positive probability and the uplink erasure probability. When $\epsilon_{\mathrm{fp}}$ is small the floor is approximately at $(\epsilon_{\mathrm{ul}})^L$. On the other hand, when $\epsilon_{\mathrm{ul}}$ is small, the error floor is dominated by $\epsilon_{\mathrm{fp}}$.

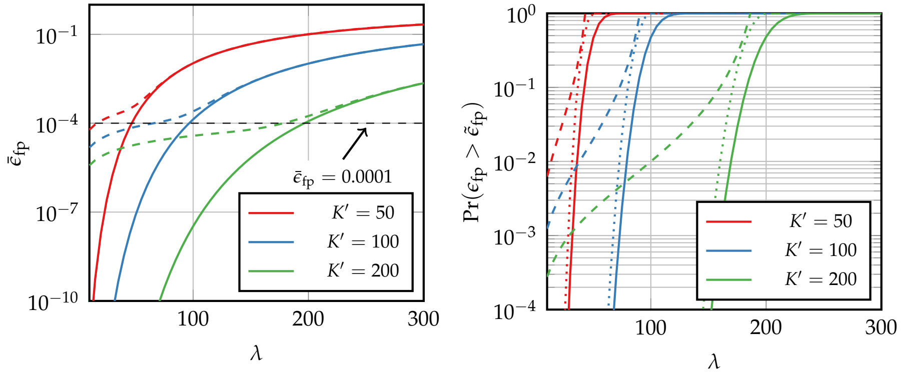
For low $\epsilon_{\mathrm{dl}}$, the failure probabilities are rather close to the asymptotic failure probabilities despite $L$ being as low as 3 or 5 (where the solid and dashed lines coincide). In this regime, the failure probability is limited only by $\epsilon_{\mathrm{fp}}$ and $\epsilon_{\mathrm{ul}}$, and increasing the downlink reliability or the number of transmission rounds will not lead to a reduced failure probability.

## 7.2 Random $K$ and $L$ Transmission Rounds

We now study the case when $K$ is random, and start by assessing the accuracy of the bounds derived in Propositions 1 to 3 and investigating how the false positive probability depends on the distribution of $K$. The impact of the distribution of $K$ is illustrated in Fig. I.4, where $K$ follows a Poisson distribution with mean $\lambda$. Fig. I.4(a) shows the expected false positive probability $\bar{\epsilon}_{\mathrm{fp}}$, computed numerically, and the bound from Proposition 1 when the feedback message length $B$ is optimized to provide a false positive probability $\tilde{\epsilon}_{\mathrm{fp}} = 0.0001$ when $K = K'$. As can be seen, the expected false positive probability is larger than the target false positive probability of $\tilde{\epsilon}_{\mathrm{fp}} = 0.0001$ when $\lambda = K'$, suggesting that optimizing based on only the expected number of active users is insufficient. However, the bound, which also takes into account the variance of $K$, is accurate when $\lambda$ is close to and greater than $K'$, and can be used to pick a feedback message length that satisfies the target false positive probability when $\lambda = K'$ at the cost of only a minor message length penalty.

The probability that $\epsilon_{\mathrm{fp}}$ exceeds $\tilde{\epsilon}_{\mathrm{fp}} = 0.0001$ is shown in Fig. I.4(b) along with the bounds from Propositions 2 and 3. While the bound from Proposition 2 is reasonable when $\lambda$ is close to $K'$, it is generally quite weak due to the strong concentration of the Poisson distribution around its mean. However, by assuming that the users activate independently as in Proposition 3 the bound can be significantly tightened especially for low $\lambda$.

We now turn our attention to the case with $L$ transmission rounds, and assume that the length of the feedback message $B$ is selected using Proposition 1 to satisfy a given false positive requirement $\tilde{\epsilon}_{\mathrm{fp}}$ on average. The failure probability when $K$ is Poisson distributed is shown in Fig. I.5 for $L = 5$ and



(a)

(b)

**Fig. I.4:** Illustrations of the bounds for (a) the expected false positive probability $\bar{\epsilon}_{\mathrm{fp}}$, and (b) the probability that $\epsilon_{\mathrm{fp}}$ exceeds $\tilde{\epsilon}_{\mathrm{fp}} = 0.0001$ when $K$ is Poisson distributed with mean $\lambda$. The feedback message length $B$ is optimized to guarantee a false positive probability of $\tilde{\epsilon}_{\mathrm{fp}} = 0.0001$ when $K = K'$. In (a) the dashed line shows Proposition 1; in (b) the dashed and dotted lines show Proposition 2 and Proposition 3, respectively.

$\epsilon_{\mathrm{dl}} = 0.01$. Because the message length is selected using the bound from Proposition 1, the failure probability for random $K$ is lower than the one with deterministic $K$, indicated by the dashed lines. The gap between the failure probability for deterministic $K$ and random $K$ decreases as $\lambda$ increases, which is due to the bound becoming more tight in this regime. This confirms that the bound can be used as a useful tool to select the message length.

## 7.3 Source/Channel Coding Trade-off

We finish the section by studying the trade-off between the number of bits used to encode the acknowledgments and the transmission rate. We assume that $K$ is Poisson distributed with arrival rate $\lambda$ and the uplink reliability is $\epsilon_{\mathrm{ul}} = 0.1$. For the downlink, we pick $\epsilon_{\mathrm{dl}}$ using Eq. (I.46) for the case in which the BS has 64 antennas, there are $L = 5$ transmission rounds, and $c = 2048$ channel symbols are available for the feedback, such that the transmission rate is $B/2048$ bits/symbol. Assuming a quasi-static flat-fading Rayleigh channel with average SNR $\overline{\mathrm{SNR}}$, the instantaneous SNR at the user, $\gamma$, is Gamma distributed with shape and scale parameters equal to 64 and $\overline{\mathrm{SNR}}/64$, respectively. We consider four encoding methods, namely identifier concatenation, the error-free (EF) method from Eq. (I.7), the scheme based on linear equations (LE) presented in Appendix 4.4, and the asymptotically optimal scheme from Eq. (I.15). For each value of $\lambda$ and each encoding scheme, we optimize $B$ so that $\Pr(\mathrm{fail})$ is minimized when averaged over the instanta-

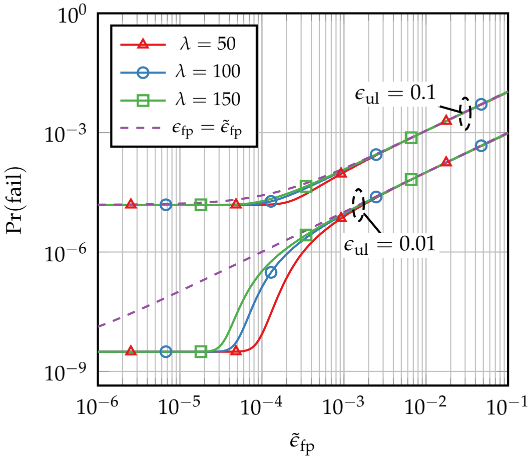


**Fig. I.5:** Failure probability vs. target average false positive probability $\tilde{\epsilon}_{\text{fp}}$ for Poisson arrivals with mean $\lambda$, for $L = 5$ and $\epsilon_{\text{dl}} = 0.01$. The message lengths are selected using the bound in Proposition 1, and the dashed lines show the failure probability when $K$ is deterministic.

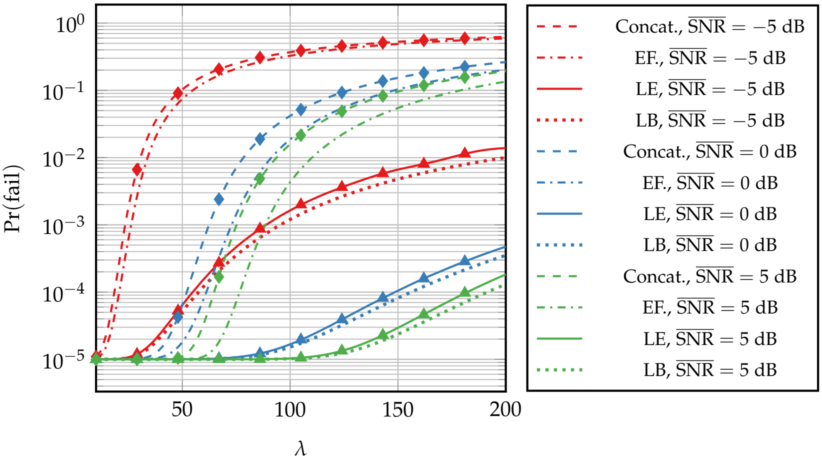


**Fig. I.6:** Failure probability for the concatenation based encoding compared to the error-free (EF) bound (Eq. (I.7)), the linear equations (LE) scheme (Eq. (I.24)), and the lower bound (LB, Eq. (I.15)) when the acknowledgment is transmitted over a Rayleigh fading channel with $c = 2048$ symbols and 64 transmitter antennas. $K$ is Poisson distributed with mean $\lambda$, $\epsilon_{\text{ul}} = 0.1$ and $L = 5$. Diamond and triangle markers are obtained by simulation of the concatenation and LE schemes, respectively.

neous arrivals given $\lambda$. Thus, the transmission rate remains fixed for a given $\lambda$. In the concatenation and error-free schemes, we assume that each identifier requires 32 bits, and when the length of the feedback message is less than $32K$ (if the instantaneous $K$ is large compared to the message length), we encode a random subset comprising $\lceil B/32 \rceil$ identifiers, which results in a false negative probability of $\epsilon_{\text{fn}} = 1 - \lceil B/32 \rceil / K$ (but no false positives). Therefore, while these representations are error-free when the number of bits $B$ is adapted to $K$, they are *not* error free here where $K$ is random and the number of bits is optimized to minimize the failure probability.

The results are shown in Fig. I.6 for $\overline{\text{SNR}} \in \{-5, 0, 5\}$ dB. The figure shows that, despite introducing false positives, the failure probability can be substantially decreased when the scheme based on linear equations is used compared to both the straightforward concatenation scheme and the bound given by the EF scheme. This is because admitting false positives allows the message length to be significantly reduced (and thus, the transmission rate), which in turn leads to much higher reliability of the downlink feedback. Furthermore, as expected the scheme based on linear equations performs close to the asymptotically optimal bound, with a gap caused only by the rounding in Eq. (I.24). Finally, we see that simulations, indicated by the markers, agree with the theoretical analysis, manifesting that the gains can be attained in practice. We note that these results have been obtained under the assumption that the BS does not have channel state information (CSI) of the decoded users, although many massive random access schemes obtain this as part of the decoding procedure [4, 5]. When CSI is available, the BS can increase the SNR at the devices that were successful in the uplink, while the unsucessful users will experience a lower SNR. This effectively suppresses the false positive probability, leading to an even smaller failure probability.

Fig. I.7 shows the failure probability vs. feedback message length $B$ for $\lambda = 100$, $\overline{\text{SNR}} = -5$ dB, and various number of transmission rounds $L$. As also suggested by the analysis in Appendix 6.2, for the asymptotic lower bound and the scheme based on linear equations, the feedback message length $B$ that minimizes the failure probability increases as the number of transmission rounds increases, causing $\epsilon_{\text{dl}}$ to decrease. On the other hand, for the concatenation scheme, the feedback message length that minimizes the failure probability is the same independently of $L$. This is because $\epsilon_{\text{fn}}$, which increases with $B$, and $\epsilon_{\text{dl}}$, which decreases with $B$, both lead to the same event, namely a retransmission. This point has a high false negative probability of $\epsilon_{\text{fn}} \approx 0.78$, while the outage probability is low ($\epsilon_{\text{dl}} \approx 0.05$). On the other hand, the schemes that allows for false positives has $\epsilon_{\text{fp}}$ in the range 0.005 to 0.021, while $\epsilon_{\text{dl}}$ is ranges from approximately 0.01 to 0.30. This illustrates, in line with existing literature, the fact that the false positive probability should generally be kept smaller then the false negative probability. Despite this, the resulting failure probability is significantly smaller when

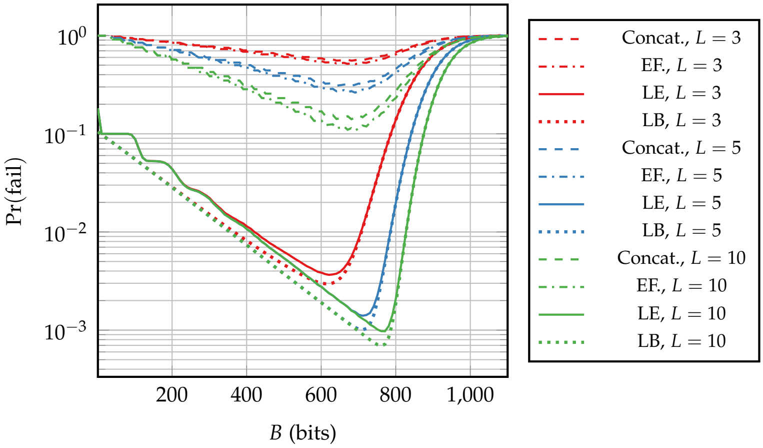


**Fig. I.7:** Failure probability for various acknowledgment message lengths $B$ using the concatenation scheme, the error-free (EF) bound, the linear equations (LE) scheme, and the asymptotic lower bound (LB) with false positives over a Rayleigh fading channel with $c = 2048$ symbols, 64 transmitter antennas. $K$ is Poisson distributed with mean $\lambda = 100$, $\epsilon_{\text{ul}} = 0.1$ and $\overline{\text{SNR}} = -5$ dB.

false positives are allowed, compared to the case where they are not.

# 8 Conclusion

In this work, we have studied the use of message acknowledgments in a massive random access setting. We have shown that because of the large number of users that are active at any given time, encoding the feedback message requires a significant number of bits. To reduce this amount. we propose to allow for a small fraction of false positive acknowledgments, which results in a significant reduction in the length of the acknowledgment message. We have presented and analyzed a number of practical schemes of various complexity that can be used to realize these reductions, and shown that their performance is close to the information-theoretic optimum. With the basis of these schemes, we have studied their performance when the number of decoded users is random, and derived bounds on the false positive probability in this setting. Finally, we have studied how the schemes perform in a scenario with retransmissions, and shown, through numerical results, the extent to which reducing the feedback message length can improve the overall reliability of the random access scenarios.

# A Appendix

## Derivation of Eq. (I.9)

For completeness, we derive here the lower bound in Eq. (I.9) presented as Proposition 4 in [14].

Suppose we construct a feedback message that acknowledges a set of users $\mathcal{W} \subset [N]$. We are interested in finding the number of sets $\mathcal{S}$ of size $K$ that such a message can acknowledge while the requirements in terms of false positives and false negatives are satisfied. Clearly, in order to meet the false positive requirement, we must have $|\mathcal{W}| \leq K + \lfloor \epsilon_{\text{fp}} N \rfloor$.

Consider first the sets $\mathcal{S}$ for which the message $\mathcal{W}$ has exactly $i$ false negatives and thus $K-i$ true positives. In order for $\mathcal{W}$ to be a valid message for such a set, at least $K-i$ users of $\mathcal{S}$ must belong to $\mathcal{W}$, while the remaining $i$ users can be any of the $N - |\mathcal{W}|$ users that are not acknowledged by $\mathcal{W}$. For a given message $\mathcal{W}$, the number of such sets is $\binom{|\mathcal{W}|}{K-i}\binom{N-|\mathcal{W}|}{i} \leq \binom{K+\lfloor \epsilon_{\text{fp}} N \rfloor}{K-i}\binom{N}{i}$. Thus, the number of sets with up to $\lfloor \epsilon_{\text{fn}} K \rfloor$ false negatives is at most

$$\sum_{i=0}^{\lfloor \epsilon_{\text{fn}} K \rfloor} \binom{K + \lfloor \epsilon_{\text{fp}} N \rfloor}{K-i} \binom{N}{i} \leq K \binom{K + \lfloor \epsilon_{\text{fp}} N \rfloor}{K - \lfloor \epsilon_{\text{fn}} K \rfloor} \binom{N}{\lfloor \epsilon_{\text{fn}} K \rfloor}. \tag{I.47}$$

The total number of bits to represent all $\binom{N}{K}$ possible sets $\mathcal{S}$ is therefore at most

$$B^*_{\text{fp,fn}} \geq \log_2 \left( \frac{\binom{N}{K}}{K \binom{K+\lfloor \epsilon_{\text{fp}} N \rfloor}{K - \lfloor \epsilon_{\text{fn}} K \rfloor} \binom{N}{\lfloor \epsilon_{\text{fn}} K \rfloor}} \right) \tag{I.48}$$

$$= \log_2 \binom{N}{K} - \log_2 \left( K \binom{K + \lfloor \epsilon_{\text{fp}} N \rfloor}{\lceil (1-\epsilon_{\text{fn}}) K \rceil} \binom{N}{\lfloor \epsilon_{\text{fn}} K \rfloor} \right). \tag{I.49}$$

Note that this bound is valid only when $\epsilon_{\text{fp}} < 1/2$, as it otherwise might be beneficial to encode the users that should *not* be acknowledged instead of the users that should.

# References

[1] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, C. Stefanovic, P. Popovski, and A. Dekorsy, "Massive machine-type communications in 5G: physical and MAC-layer solutions," *IEEE Communications Magazine*, vol. 54, no. 9, pp. 59–65, 2016.

[2] C. Stefanovic and P. Popovski, "ALOHA Random Access that Operates as a Rateless Code," *IEEE Transactions on Communications*, vol. 61, no. 11, pp. 4653–4662, 2013.

[3] E. Paolini, G. Liva, and M. Chiani, “Coded Slotted ALOHA: A Graph-Based Method for Uncoordinated Multiple Access,” *IEEE Transactions on Information Theory*, vol. 61, no. 12, pp. 6815–6832, 2015.

[4] L. Liu, E. G. Larsson, W. Yu, P. Popovski, C. Stefanovic, and E. de Carvalho, “Sparse Signal Processing for Grant-Free Massive Connectivity: A Future Paradigm for Random Access Protocols in the Internet of Things,” *IEEE Signal Processing Magazine*, vol. 35, no. 5, pp. 88–99, 2018.

[5] A. Fengler, P. Jung, and G. Caire, “Pilot-Based Unsourced Random Access with a Massive MIMO Receiver in the Quasi-Static Fading Regime,” in *2021 IEEE 22nd International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, 2021, pp. 356–360.

[6] M. Shirvanimoghaddam, M. Dohler, and S. J. Johnson, “Massive Multiple Access Based on Superposition Raptor Codes for Cellular M2M Communications,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 1, pp. 307–319, 2017.

[7] J. Kang and W. Yu, “Minimum Feedback for Collision-Free Scheduling in Massive Random Access,” *IEEE Transactions on Information Theory*, vol. 67, no. 12, pp. 8094–8108, 2021.

[8] J. Östman, R. Devassy, G. Durisi, and E. G. Ström, “Short-Packet Transmission via Variable-Length Codes in the Presence of Noisy Stop Feedback,” *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 214–227, 2021.

[9] S. C. Draper and A. Sahai, “Variable-length channel coding with noisy feedback,” *European Transactions on Telecommunications*, vol. 19, no. 4, pp. 355–370, 2008.

[10] P. Wu and N. Jindal, “Coding versus ARQ in Fading Channels: How Reliable Should the PHY Be?” *IEEE Transactions on Communications*, vol. 59, no. 12, pp. 3363–3374, 2011.

[11] 3GPP, “Medium Access Control (MAC) protocol specification,” 3rd Generation Partnership Project (3GPP), TS 36.321, 2021, v16.6.0.

[12] Y. Polyanskiy, “A perspective on massive random-access,” in *2017 IEEE International Symposium on Information Theory (ISIT)*, 2017, pp. 2523–2527.

[13] T. Cover, “Enumerative source encoding,” *IEEE Transactions on Information Theory*, vol. 19, no. 1, pp. 73–77, 1973.

[14] R. Pagh and F. F. Rodler, “Lossy dictionaries,” in *Proceedings of the 9th Annual European Symposium on Algorithms*. Springer, 2001, pp. 300–311.

[15] L. Carter, R. Floyd, J. Gill, G. Markowsky, and M. Wegman, "Exact and approximate membership testers," in *Proceedings of the tenth annual ACM symposium on Theory of computing (STOC)*. ACM Press, 1978.

[16] M. Dietzfelbinger and R. Pagh, "Succinct data structures for retrieval and approximate membership," in *35th International Colloquium on Automata, Languages and Programming*. Springer, 2008, pp. 385–396.

[17] P. Erdős and J. Spencer, *Probabilistic methods in combinatorics*. Academic Press New York, 1974.

[18] B. H. Bloom, "Space/time trade-offs in hash coding with allowable errors," *Communications of the ACM*, vol. 13, no. 7, p. 422–426, Jul 1970.

[19] A. Broder and M. Mitzenmacher, "Network Applications of Bloom Filters: A Survey," *Internet Mathematics*, vol. 1, no. 4, p. 485–509, Jan 2004.

[20] E. Porat, "An optimal Bloom filter replacement based on matrix solving," in *International Computer Science Symposium in Russia*. Springer, 2009, pp. 263–273.

[21] K.-M. Chung, M. Mitzenmacher, and S. Vadhan, "When Simple Hash Functions Suffice," p. 567–585.

[22] P. C. Dillinger, L. Hübschle-Schneider, P. Sanders, and S. Walzer, "Fast succinct retrieval and approximate membership using ribbon," *CoRR*, vol. abs/2109.01892, 2021. [Online]. Available: https://arxiv.org/abs/2109.01892

[23] M. Mitzenmacher and E. Upfal, *Probability and computing: Randomization and probabilistic techniques in algorithms and data analysis*. Cambridge university press, 2017.

„Wobec kłamliwych jawnie snów, wobec zmarniałych w nicość cudów,
Potężne młoty legły w rząd, na znak spełnionych godnie trudów.

I była zgroza nagłych cisz. I była próżnia w całym niebie!
A ty z tej próżni czemu drwisz, kiedy ta próżnia nie drwi z ciebie?”

Bolesław Leśmian, *„Dziewczyna”*

ISSN (online): 2446-1628
ISBN (online): 978-87-7573-933-2

AALBORG UNIVERSITY PRESS