Technische Universität München
Fakultät für Elektrotechnik und Informationstechnik
Lehrstuhl für Kommunikationsnetze

Delay-Constrained Reliable Cellular Uplink Radio Resource Management for Industrial Internet of Things

Halit Murat Gürsu, M.Sc.

Vollständiger Abdruck der von der Fakultät für Elektrotechnik und Informationstechnik der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktor-Ingenieurs (Dr.-Ing.)

genehmigten Dissertation.

Vorsitzender:		Dr. sc. tech. Gerhard Kramer
Prüfer der Dissertation:	1.	Prof. Dr.-Ing. Wolfgang Kellerer
	2.	Assoc. Prof. Cedomir Stefanovic, PhD

Die Dissertation wurde am 02.10.2019 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 20.02.2020 angenommen.

Delay-Constrained Reliable Cellular Uplink Radio Resource Management for Industrial Internet of Things

Halit Murat Gürsu, M.Sc.

20.02.2020

Abstract

Communication is merely one of the pillars of productivity and there are many more such as automation. As part of automation, machines are replacing humans and the amount of machines are more common on both ends of the communication. Wireless communication enables us to communicate while we are on the move and leads to higher productivity. Radio resource management is the problem of coordination of wireless communication among multiple entities. The radio resource management for automation enabling machines defines the upcoming phenomenon of industrial internet of things that poses new challenges differing from human communications: (1) The communication patterns of machines are more correlated with one another such that dimensioning the system resources assuming users are independent is not valid anymore, (2) a huge amount of machines transmit small amount of data that makes the overhead of current communication significant such that new radio resource management techniques are required, (3) machines are shouldering safety-critical tasks requiring the underlying communication system between them to be highly reliable.

This thesis introduces a delay-constrained reliable radio resource management system for industrial internet of things. This is achieved in three main components dealing with the three challenges: (1) The unpredictable activity pattern is proposed to be smoothed with an admission control. (2) Efficient resource management algorithms are proposed to maximize the capacity for dynamic delay-constraint guarantees. (3) The system is prototyped using IEEE 802.15.4 sensors to show reliable operation solving many physical layer based reliability challenges.

The activity patterns for machines is predicted to have a Beta distribution by 3GPP. This distribution presents a bursty behavior and trafic shaping techniques can be used to smooth this burstiness. However, the shaping operation has to be adjusted to hyperparameters of the distribution which would vary drastically depending on the application. A solution is to use an estimator to predict the hyperparameters of the activity. This provides a self-adaptive solution to the heterogeneous application space for machines. This thesis shows that use of an estimator enables higher efficiency for radio resource management. This estimator is integrated to an admission control before the initial access as an adaptive technique that accepts and rejects

machines with respect to the system capacity. Even though this guarantees reliable serving for accepted users, we have to eventually provide service for the rejected machines. This is possible through an increase in the system capacity.

Currently, as the drivers of the cellular networks are multimedia communications, the radio resource management is based on high payload and low amount of users. Grants are used to avoid large scale interference. The introduction of the industrial internet of things reverses this picture to low payload and high number of users. Sending a grant to send a payload that is smaller than the grant request is inefficient. Thus, grant free algorithms are evaluated as a solution. The delay-constraints have been neglected by the grant-free algorithms so far. The only solution to achieve these constraints is to add more capacity in terms of channels. This thesis proposes grant-free algorithms that can make use of a throughput-optimal selection for the number of channels to increase the efficiency of the system. The efficiency can be improved with *successive interference cancellation* that is a complex decoding technique enabled by increasing edge computing capabilities. This thesis proposes successive interference cancellation based new delay-constraint aware algorithms to provide the best radio resource management efficiency.

This thesis prototypes the proposal with the commercial off-the-shelf boards supporting IEEE 802.15.4. Thus, the modulation, the maximum power and the channelization are not investigated and reliability improvements are provided with techniques such as decoding, channel estimation and frequency planning. The techniques we focus on deal with the challenges: *external interference*, *channel variations* and *hardware related problems*. External interference is caused by the co-existence of multiple technologies on the same band which is inevitable due to the scarcity of the frequency spectrum. We propose to use a frequency planning, *whitelisting* to reduce the effect of *interference* in a controlled environment. Stochastic variations on the hardware and the wireless channel cannot be dealt with through planning. We provide a practical model for the SINR that incorporates the soldering imperfections in the system on chips that causes a constant phase noise and demonstrate that successive interference cancellation based grant free algorithms are ready to be deployed in industrial environments.

Acknowledgment

Hello wandering soul. Firstly, I thank you for taking time for reading my thesis and this acknowledgement. Knowledge is meaningless in one´s brain unless it is shared by someone. As with any communication, with passive form there is a two-ways communication it takes one to spell out an information and it takes one to find out that piece of information hidden in a thesis.

The journey that lead to this thesis is with many caveats and I cannot do enough even if I thank to all of them separately. I will try nevertheless.

My story starts with my mother, Nuran Gürsu, being the biggest believer in me. I am glad that I am one of the people in life that had a strong believer from the start. Without her I wouldn´t have the courage to accomplish and motivate myself for any of the steps that lead to this thesis. I also have to mention that my father, Halil Gürsu, not appreciating(!) any of the achievements I obtained up to this thesis, is again one of the main factors motivating me to work for this target to finally show him that I did it.

I cannot thank my wife, Hilal Gürsu, enough for taking the leap of faith with me, deciding to come to Germany together. This decision was to give it all to acquire a PhD from one of the best universities in the world. It was not an easy decision, it was not an easy process. She has been there for me when I needed the support or about to turn away from my focus. I hope everyone is blessed with a companion like her in their lives. I hope I can be as much to her as she has been for me. I also cannot thank my in-laws enough for raising her with such love that to this day she radiates back to people around her.

I have to specially mention my brother, Ali Emre Gürsu, which I turn for consulting even though I am the elder brother. He is the one I call to have a hard reset in my mind in my most stressful moments. I am really lucky to have him.

Looking to the details of the story, I want to specially thank Prof. Yasemin Kahya for going over my cover letter with me together with a really tight deadline and probably influenced highly my chances of getting in Technical University of Munich. Likewise, I also want to thank Dr. Helmut Gräb that conducted my interview and tutored me about how to undertake my studies at TUM.

I would like to extend my thanks to my friends, that made me love Munich during my graduate studies, Emrah Karabacak, Raul Figuera, Serhan Gül, Vianney Martinez, Malek El-Khatibi, Mustafa Cemil Coskun.

I would like to thank Prof. Wolfgang Kellerer for his supervision and for the working environment he provided at the chair of Communication Networks thanks to which I flourished. I could not have completed my studies if it hadn´t been his understanding and kindness. I would like to thank Carmen Mas Machuca also for telling me during my first conference visit that I should use "being easy to talk to", that lead me to having many collaborators during my thesis.

I would like to thank to my chair mates, Amaury van Bemten, Arled Papa, Samuele Zoppi, Onur Ayan, Amir Varasteh, Nemanja Djeric, Raphael Durner, Ermin Sakic, Mikhail Vilgelm, Markus Klügel, Andreas Blenk, Christian Sieber, Alberto Martinez Alba, Johannes Zerwas, Patrick Kalmback, Jochen Guck and others that I may have forgotten. You probably made LKN a home for me, and it will be one of the most memoriable time of my life.

Also I would like to extend my thanks to friends who accompanied me in many occasions to keep my heads up during hard times, Mahmut Kafkas, Ebru Yildirimli, Emrecan Tasgin, Efe Yigitbasi, Cagatay Cobanoglu, Fatih Göksu, Umur Karabulut, Okan Köpüklü, Halil Yoldas, Orhan Öcal, Yalcin Ozhabes, Lars Teiwes, Johannes Stadler, Jan Jansen, Onur Günlü, Ramazan Alishov, Ralf Bär.

Contents

Chapter 1

Introduction

Mobile wireless communication networks have been designed to create a ubiquitous communication system for humans to interact with each other anywhere and anytime. Thus, voice has been the main driver until the end of 3G era where data communications started to catch up in terms of volume. Nowadays, voice communication is also treated as data communications and mobile networks have converged to the single goal of serving data packets. Some data packets, like voice packets require certain metrics to be fulfilled. This defines quality of service that enables data service to be used seamlessly by the end users. Such requirements are then enforced to be sustained on average as none of the requirements pose any life-critical threats. This is about to change with introduction of automation to wireless communication.

Digital wired communication has been in use for safety critical tasks such as aviation and industrial automation already for some decades. So why are safety critical applications so much more difficult to provide with wireless communication? Intuitively, the limitation comes from the single difference of wired to wireless communication that is the communication medium. Advantages of wiring is (1) isolation from external electromagnetic signals, (2) high link capacity and (3) dedicated link per user. Conversely, none of these are given for wireless communication, where (1) the medium is susceptible to electromagnetic interference and other diversions to electromagnetic signals due to physical objects; (2) the total link capacity depends on the total bandwidth allocated to the wireless technology and (3) the link capacity per user depends on the bandwidth allocated to each user as due to the broadcast nature of wireless, all users have to share the total bandwidth. Each of these problems require special attention on their own.

This chapter is organized as follows. Sec. 1.1 shortly introduces the challenges tackled in the thesis. Sec. 1.2 summarizes the contributions with respect to the challenges. Sec. 1.3 shares an overview about the organization of the thesis.

1.1 Challenges

The isolation of the medium is not possible, thus the status of the wireless channel has to be tracked and managed continuously. Tracking and reversing the wireless effects on a signal does not provide the same performance as isolating electromagnetic signals as in wires. This is called channel equalization and has to be adapted with respect to the underlying events concerning the mobility of the environment or of the transmitter and the receiver device. As long as the characteristics of the wireless communication environment are known, these effects can be mostly reversed. One unknown effect is the electromagnetic interference caused by sporadic activity of wireless transmitters. The interference problem is solved to an extent with separation of bandwidth allocated to each wireless technology as illustrated in Fig. 1.1. A timeline for commercial wireless technologies is given vertically while the horizontal axis demonstrates the wireless spectrum.

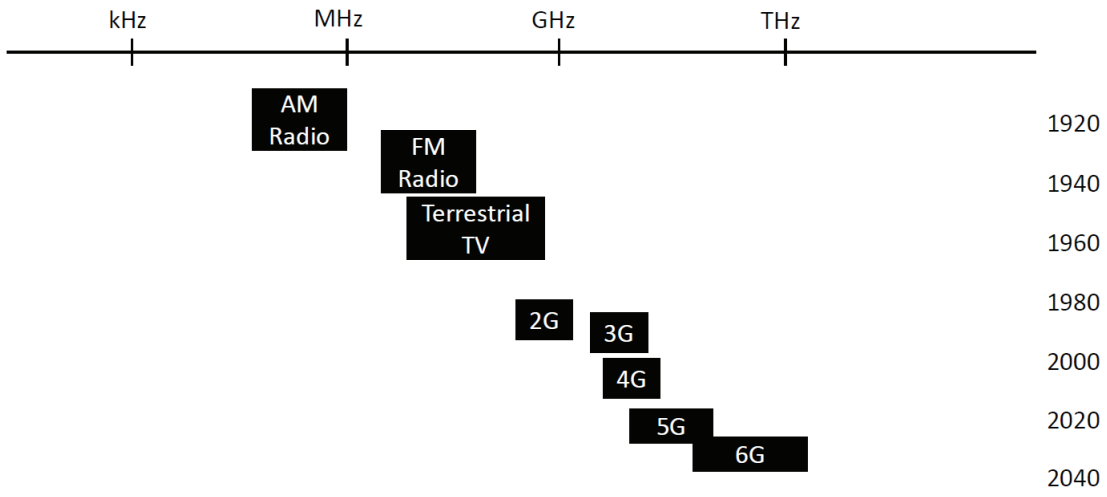


Figure 1.1: The spectrum allocation evolution through time for wireless technologies.

The distribution of frequency spectrum among multiple wireless technologies is not always done in an orthogonal fashion due to scarcity of the bandwidth. Especially for lower frequency range, as it is relatively easy to design transceivers, there are many technologies that use the same band. The overlap is guided via certain rules, such as geographical and power limitations to guarantee fair sharing of the spectrum. When it comes to guarantee a certain performance, these rules have to be tested. However, strict guarantees are not possible as long as full control of one technology over another is assumed. This can only be assumed for non-public areas, which is a valid assumption for certain use-cases like industrial automation, as factories should be the initial deployment location for industrial internet of things networks.

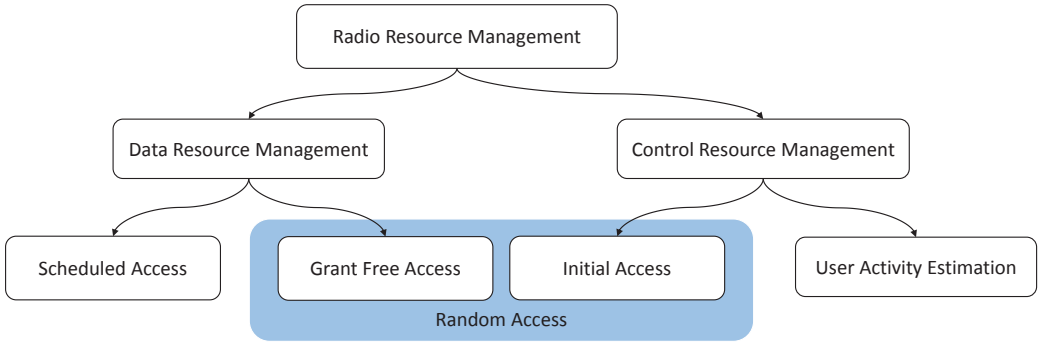


Figure 1.2: The spectrum allocation evolution through time for wireless technologies.

Virtual slices of the frequency spectrum is called a radio resource. The allocation of radio resources for different tasks is called the radio resource management (RRM). Different parts of RRM is illustrated in Fig. 1.2. There are resources dedicated for data communications and control communications. Data resource management, is mainly separated into two parts as scheduled access and grant free access, depending on if the users are allocated dedicated resources or not, respectively. Control communications encompasses many different aspects but we limit its scope to only initial access and user activity estimation in this thesis. The reason we involve these aspects is that there is an interplay between these control resources and data resources such that a holistic optimization has to be taken into account. In the state of the art, both initial access algorithms and grant free algorithms are considered as random access algorithms.

With the scheduled access, the scheduling resources are allocated to users with respect to the instantaneous requirements of the users. Obtaining this information is possible through the initial access but it requires its own bandwidth. It is clear that allocating too much bandwidth to initial access such that the scheduled access can be done efficiently does not make sense. The optimal trade-off has to be found. The initial access bandwidth is dependent on the activity of the users. If users rarely send requests on initial access, it is not a problem. But if they send requests in synchronization, which is the case for some users belonging to industrial internet of things applications, then the sharing of the band is not possible anymore. The shared spectrum with synchronized activity defines a new challenge for wireless communication. It

is clear that the resources allocated for initial access has to be related to the user activity. So the amount of resources to estimate the user activity is another radio resource management problem. The connection of initial access and scheduled access can be skipped if the users transmit their payloads in a grant-free fashion on shared resources. In both cases, we have to deal with the user activity estimation problem.

Consequently, the bottlenecks of access protocols with respect to new radio performance metrics, related to Industrial IoT, have to be evaluated. State-of-the-art protocols that focus on average access delay metric cannot be used anymore due to criticality of the current use cases. This is currently an unsolved issue for most of the use cases of the Industrial IoT. The behavior of these protocols has to be analyzed with higher precision.

The main metric used to evaluate access protocols is fulfillment of delay constraints, i.e., if a request is allocated a resource with a limited time. Assuming the user activity estimation is available, the tight delay constraints imposed by Industrial IoT cannot be supported by any random access protocol. Thus, random access protocols have to be adapted for such constraints. The challenges of using the delay constraint metric is two-fold: (1) the mapping from application requirements to radio performance metrics and (2) developing access protocols that fulfills the radio performance metrics.

The mapping of application requirements to radio performance metrics is application specific and involves multiple system level assumptions. The interaction between different sub-systems have to be clearly stated and inter-dependencies have to be defined. Only a detailed analysis would result in sub-system performance metrics for a real world performance fulfillment.

The full characterization of the access protocols is not possible without any assumptions for the activity of the users. This requires modeling of the user traffic. Currently, multiple models are accepted as the ground truth as 3GPP, the standardization body of mobile networks, has considered them for specific scenarios. However, connection of these models to real world applications is still missing, and as long as industry is reluctant to open their data to public, we will observe a divergence between the models and the real activity. One way to overcome this problem is to introduce a solution that learns the traffic on the fly.

1.2 Contribution

This section summarizes the contributions of this thesis. An illustration of the contributions is given with Fig. 1.3.

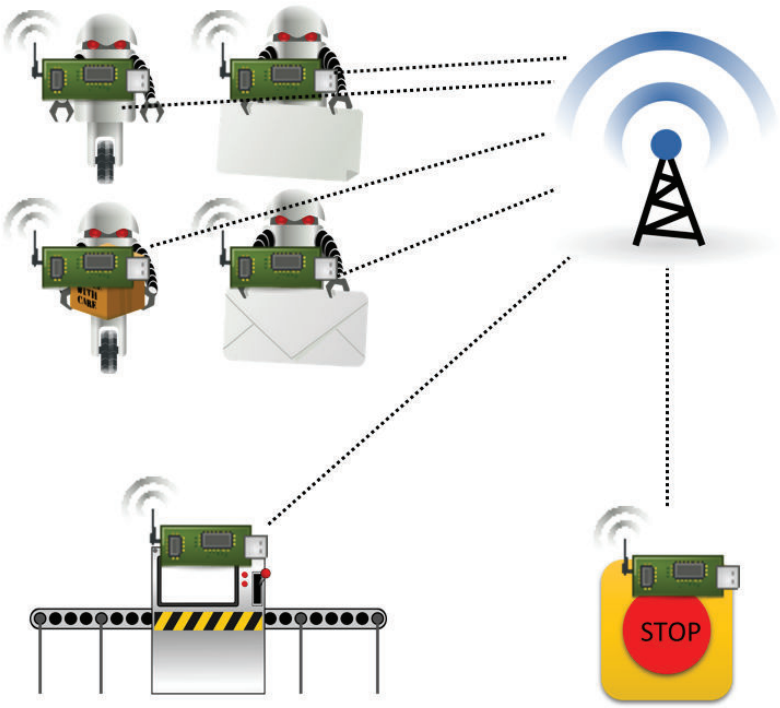


Figure 1.3: Industrial IoT involves multiple applications such as mobile robots, control of production lines and safety-critical kill switches all connected to the same base station.

The thesis emphasizes on four main aspects of the aforementioned problems. Reliability requirements for wired communication are defined for safety critical communications. As we are interested in migrating these use cases from wired communication to wireless,

1. the mapping of the system level requirements to wireless communication metrics has been achieved and latency-reliability metrics are underlined. Following, the access protocols have to be evaluated with respect to latency-reliability metrics. The evaluation of access protocols resulted in underlining user activity and algorithm re-activity as main bottlenecks.

2. The access protocols are adapted with respect to user activity through use of a novel admission channel before the initial access.

3. Following, further the feedback scheme used in access algorithms is improved in order to provide strict reliability and latency guarantees.

4. Finally, a real world implementation of the proposed protocols on commercial of the shelf sensors is demonstrated in order to validate the contributions through measurements and to show that the proposed solution is ready for wide-scale deployment for industrial internet of thing.

The break down of safety-critical communication requirements sheds light to multiple aspects of the reliability definition based on each application in our work in [Gür+15] ,[Gür+17a]. One major aspect is that the reliability is defined per flight phase and would require different evaluation considering its effect on the wireless communication. For instance human mobility would be higher on in-flight phase while shielding vibration would be dominant for landing and take-off phases. This clearly has a limited effect on wired communication while wireless communication is affected by the surroundings.

Consequently, we evaluate the performance of the state-of-the-art access algorithms for strict latency-reliability requirements versus the assumed user activity models in state of the art with our works [GAK17a] , [Vil+17].In [Gür+17b], we propose improvements to these algorithms using combination of conventional techniques. However, the validity of such models are unclear for industrial use-cases. Thus, we investigate a solution in our work [Gür+19a] that would be reactive to any traffic. This solution adapts a admission control policy, that only admits certain users to the system, at the same time guaranteeing the fulfillment of their requirements. The admission is decided through a simple estimation algorithm that enables reactivity against traffic. The algorithm uses a closed formula derived from the famous Coupon Collector's Problem. The low complex solution enables real-time reactivity to traffic. The extension of this problem to successive interference is investigated in our work [Gür+18]. The efficiency of the access algorithms using the estimation techniques are analytically and simulatively investigated by our work [GKS19].

The next major step to guarantee performance is investigating capabilities of the access algorithms and providing improvements. To the best of our knowledge we provide the first analysis fit for delay-constrained reliable access in [GAK17b]. This analysis uses the tree structure that can be tracked in a step-wise fashion to provide resources for a high reliability requiring application. This is further extended in another work of ours [GGK19] to deterministic guarantees in a rather simplified fashion enforcing limitations to the access algorithm including the addresses of users enabled by successive interference cancellation. The novelty is that the analysis provides bounds for maximum access delays. Compared to the other solutions without feedback, this algorithm enables a low latency region with almost no added complexity.

The bottom of the communication stack, the physical layer, is investigated for reliability assumptions in terms of external interference and limitations of assumptions based on suc-

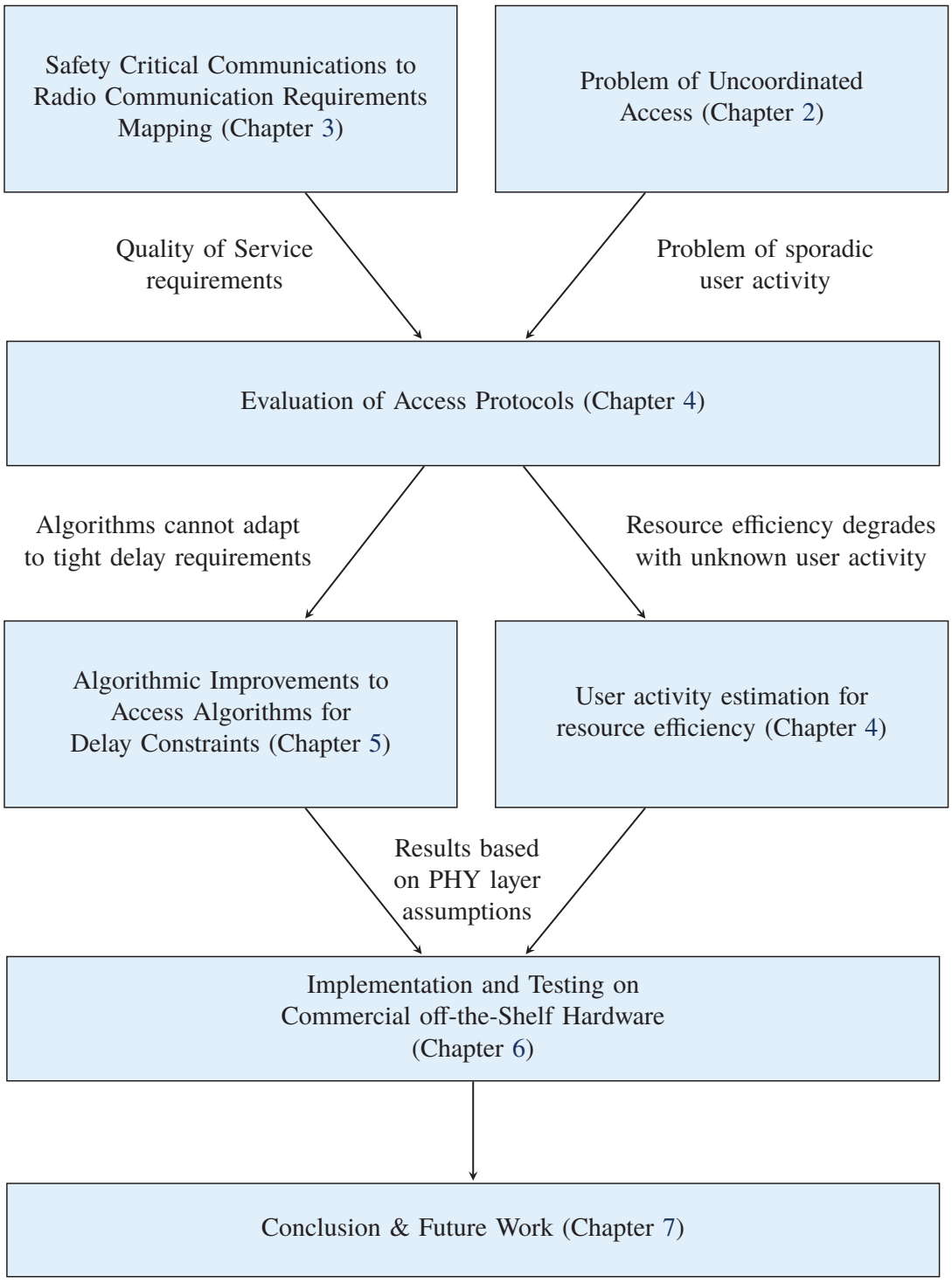


Figure 1.4: Thesis structure

cessive interference cancellation. The external interference is characterized with commercial off-the-shelf sensors in an aircraft like scenario where multiple technologies co-exist sharing

the same band [Gür+16]. We have evaluated the effects of cross-technology interference and the results showed crucial disturbance exists even though this can be overcome with complex internal or external interference cancellation techniques. For the internal case, the successive interference cancellation capabilities of commercial of the shelf sensors is evaluated to provide limitations to re-design such algorithms or to use this as a way to evaluate if it would make sense to upgrade the algorithms for the sensors in a factory.

1.3 Organization

The remainder of the thesis is organized as below.

Chapter 2 introduces the main challenge the thesis is dealing with through a human communications analogy. Next, background information for wireless communication is introduced and theoretical foundations of the challenge is elaborated. Challenges stemming from the background define the research questions investigated by the thesis.

Chapter 3 analyzes the wireless communication implications of the application layer requirements. The application under scope is safety-critical systems for enabling industrial internet of things. A safety critical application is considered as an example and broken down to components. The outcome translates into wireless performance constraints, user activity and wireless channel capacity are coined as the main challenges.

Chapter 4 starts with evaluating state-of-the-art algorithms for delay constraints and sporadic user activity. Results point out that with limited number of resources the algorithms cannot fulfill the delay requirements. Thus, the algorithms has to use resources more efficiently. One way to achieve this is to react to user activity achieving higher resource efficiency. This hypothesis is mathematically proven, motivating implementation of user activity estimation. Through introduction of an user activity estimation based admission control before the access algorithms, the throughput improvement of estimation algorithms is simulatively shown.

Chapter 5 presents two new access algorithms that are analytically characterized for delay-constrained reliable access. One algorithm is for successive interference cancellation the other one is without. The successive interference cancellation due to its high-complexity receiver and the required low noise floor is not usable in every scenario. For this reason, the advantage of feedback is used to introduce a multichannel algorithm that can be paralleled. The high reliability guarantee for delay is enhanced to deterministic delay guarantees through successive interference cancellation.

Chapter 6 describes a real world implementation of the suggested Successive Interference Cancellation based algorithm on commercial-of-the-shelf sensors. Following, a new analytical

model is introduced to include the hardware effects for the resource efficiency calculation. The results shed light on assumptions that the analytical work is based on.

Chapter 7 summarizes the work and points out directions for future work.

Chapter 2

Problem of Uncoordinated Uplink Radio Resource Management

This thesis builds up on a fundamental problem in wireless communications, that is the coordination problem for uplink radio resource management, namely the "Problem of Uncoordinated Uplink Radio Resource Management". In this chapter this problem is re-stated to familiarize the reader for the following chapters. Sec. 2.1 introduces a real world analogy to the problem. Sec. 2.2 explains the basics of wireless communications and provides a problem definition related to these fundamentals. Sec. 2.2.2 explains the limitations and challenges of focusing the problem on cellular networks with real world applications.

2.1 Introduction

This section introduces a real world analogy to the "Problem of the Uncoordinated Uplink Radio Resource Management".

It is typical that multiple stations communicate to a central station. However, it may be that none of the stations are aware of each other. This is critical if the stations communicate at the same time. If the central station receives simultaneous signals from multiple stations, it can decode neither of them. So the stations try to minimize such occurrences. This defines the problem of uncoordinated uplink radio resource management.

The uncoordinated uplink radio resource management is a specific problem. To comprehend in depth, an analogy helps to explain: In our daily lives, we do not realize how important the eye contact and non-verbal exchanges are for the coordination of communication. For instance, while communicating a group of people without visual contact have problems in taking-turns and signaling readiness [QHR15]. They have to interact with tactile senses or auditory signals. The problem of taking-turns is less of a problem in a classroom setting,

where one person from the group is assigned as the head. The head grants right to speak for all group members. However, such coordination is possible if there is a protocol to assign the head and information exchange between the members, i.e., if the head knows names of all the members of the group, then he can call each member by their name. For instance, in a setting where new members are joining to the classroom everyday, such coordination cannot be established. In a group of people without visual contact, how would the members, unknown to one another, would communicate to the head of the group? This is the same question posed by the uncoordinated uplink communication problem.

One social auditory signal that depicts the intent of communication is throat clearing noise. Such noise can be used to reserve the next communication turn [PB09]. If two members do not hear while reserving the turn for communicating, then they will start speaking simultaneously. If we assume a general polite community, at the moment that they realize the overlap, they will stop and wait the other to re-initiate communication again. As they can hear each other, this problem will self-organize in quite short amount of time. However, if there were hundreds of members trying to coordinate in this way, one can imagine that it would not work. This example mimics the current problem with WiFi communication. A technique called carrier sense is used to listen to the wireless channel before transmitting any message. With low number of users this technique works and WiFi functions properly, but it becomes inefficient (if not impossible) with increasing number of users [Bia00].

In an unknown members setting, we can still involve the head of the group that can give commands to resolve the conflict of taking turns. The conflict is defined as more than one member attempting to speak at the same time. For instance, before the session starts the head declares that when a conflict occurs, all the members involved in a conflict has to pick "*heads* or *tails*". The intent is that after a conflict occurs the head can then coordinate by saying, "The member(s) that have selected heads, please speak now". If another conflict happens again this process can be repeated until only one member is speaking, so the speaker is successfully elected. A similar scheme used to resolve uplink coordination problem in wireless communication is the "Tree Resolution Algorithm" that will be analyzed in depth in this thesis. So our interest lies in solving the coordination problem in an efficient manner. We aim to define tangible metrics to evaluate the performance of the coordination algorithm. In the following, we define the uncoordinated uplink problem formally.

2.2 Basics of Wireless Communication

This section introduces fundamentals of wireless communication required to define the "Problem of the Uncoordinated Uplink Radio Resource Management".

Wireless communication is based on emitting electromagnetic waves from a transmitter, that propagates over space and can be detected by a receiver. Light is a good example here for visualization, and itself is also an electromagnetic wave. The oscillation frequency of the wave depicts which part of the wave carries the information for us, i.e., the color of the light. This is called the *carrier* in wireless communication, as it carries the information. In order to transfer information from the transmitter to the receiver with the electromagnetic wave, we also have to modify specific characteristics of it. Sending always the same wave does not have any information as pointed out by Shannon [Sha48].

The characteristics of an electromagnetic wave can be defined by the power and time of arrival. We can modify its amplitude and time of arrival to send information to the receiver. This operation of modifying the waveform for transmitting information is called *modulation*. Modulation is the interface between analog and the digital world. Through modulation techniques, the waveform can be seen as a sequence of bits, 1s and 0s, propagating through the wireless medium.

Unfortunately, once the waveform is received, the conversion from the analog to the digital signal cannot always be done reliably. There are many physical limitations for wireless communication. For instance, assuming an isotropic transmission, the waveform starts as a point in space and it dissipates in a spherical way into the space. As can be imagined, all the energy focused on the single point, also known as the transmit energy P_t, is distributed all over the surface of the sphere with area S, given as

$$S = 4\pi d^2 \tag{2.1}$$

where d is the radius of the sphere. So the energy of the waveform, dubbed as energy per surface P_s, at a point with distance d from the transmitter, is

$$P_s = \frac{P_t}{S}. \tag{2.2}$$

Assuming an isotropic aperture with room temperature and below 1 THz frequency [Bro03][1], we have the aperture gain A_e, set by the wavelength of the waveform λ, as

$$A_e = \frac{\lambda^2}{4\pi}. \tag{2.3}$$

The received energy by the antenna can be calculated by multiplying the aperture gain with the energy per surface,

$$P_r = \frac{\lambda^2}{4\pi} \cdot \frac{P_t}{4\pi d^2} = P_t \cdot \left(\frac{\lambda}{4\pi d}\right)^2. \tag{2.4}$$

[1]The aperture gain is expected to increase with decreasing temperature and increasing frequency.

This formula summarizes the physical limitations of propagating a waveform in vacuum. As the wavelength λ is relatively small compared to the distance in wireless communication, we can say that the energy is inversely proportional to the squared distance. Once the energy reaches below the thermal noise energy, it cannot be even detected by the receiver. The noise power with Boltzmann constant k_B, temperature in Kelvin T and the bandwidth of the waveform Δf is $N = k_B \cdot T \cdot \Delta f$ [TV05].

The problem of having a strong signal versus the noise power is dealt with setting the right transmission power. This is measured with a metric called Signal to Noise ratio (SNR) that measures the received signal power relative to the thermal noise at the receiver. It is a unitless metric as it is a ratio. Shannon in 1948, with his famous formula [Sha48], has related the SNR to a term called capacity. The capacity is defined as the maximum bits that can be transferred per second on a link. A link is a virtual connection, a pipe, where the waveform travels and connects the transmitter and the receiver. The formula provided us a simple tool to evaluate the reliability of the wireless communication, or in other words received waveform to bit sequence conversion. The formula takes also the bandwidth of the signal of interest Δf in to account,

$$C = \Delta f \log_2(1 + \text{SNR}). \tag{2.5}$$

He declares that any bit rate R greater than C cannot be reliably transferred with this SNR and this bandwidth Δf. The noise power is the measurement precision of the receiver. Assuming a noise power of 3 mW and a signal power of 10 mW, the receiver can easily distinguish a signal of 10 mW and 5 mW, but the same is not true for 10 and 9 mW. Basically the division, is extracting how many discrete power levels the receiver can distinguish. Actually, since the detection is on the voltage level and SNR is on the power level the number of distinguishable levels cannot be extracted in this way. But it provides an intuition for with a certain SNR the number of binary options can be represented through $\log_2$ SNR. He also shows that the capacity C can be reached with a vanishing error probability for infinite latency if an error correction technique is used, that is called coding, and is not further investigated in this thesis.

2.2.1 Multiple Access

However, the noise energy is not the only energy masking/limiting the received waveform. Two waveforms from two transmitters, arriving at the receiving antenna at the same instance can also mask each other. This effect is called the interference effect. Waveforms establish a link, link 1 and link 2 with capacities C_1 and C_2, respectively. The received power of waveforms from each link is P_{r1} and P_{r2} respectively. From each link point of view, the other

link is limiting the differentiation of voltage levels that can be written as

$$C_1 = \Delta f \log_2(1 + \frac{P_{r1}}{P_{r2} + N}), \tag{2.6}$$

and

$$C_2 = \Delta f \log_2(1 + \frac{P_{r2}}{P_{r1} + N}). \tag{2.7}$$

This creates an interplay between the transmission power of both waveforms. An interesting problem can be defined to optimize the transmission power for each of the waveforms in order to maximize the total system capacity. However, doing such an optimization requires that transmitters are coordinated in terms of transmission power and in terms of activity. Activity is a binary metric and refers to transmission of a waveform i.e. the activity for a transmitter is 1 if a waveform is transmitted and 0 otherwise. Through the activity metric it is known which waveforms will overlap.

Referring back to the initial problem definition, we are dealing with an uncoordinated scenario where up to n_{tot} waveforms may overlap. The capacity of a link i in this case is

$$C_i = \Delta f \log_2 \left(1 + \frac{P_{ri}}{\sum_{j=1}^{n_{tot}} \delta_j P_{rj} + N}\right) \text{ where } i \neq j, \tag{2.8}$$

and $\delta_j \in \{0, 1\}$ denotes the activity of link j at that instance.

Access protocols are composed of multiple rules that the transmitters have to obey in a wireless network. These rules can enforce decisions on transmission power, activity, bandwidth and more.

An example for a fair coordination of transmitters can be achieved with an access protocol that sets an equal minimum power providing the same rates, assuming all the users are active.

In this case, even though this protocol does not maximize the capacity, it results in a guaranteed capacity for each waveform, independent from the activity. On the other hand, as in Eq. (2.8), the number of links is a denominator, such a scheme would result in close to zero rate with increasing number of links.

Referring back to our analogy, an adaptation for the proposed protocol for low rate communication would be that every member is pronouncing a single syllable for a long time, before each member moves on to the next syllable. The capacity depends on the bandwidth but is quite low with for example $n_{tot} > 10$. So a coordination has to be established to have sizable increase in capacity for each link.

The coordination of multiple users is easily achieved with separation in terms of time, frequency, space and power. For instance, we define virtual time and bandwidth slices. We can call each bandwidth-time slice a resource. As the carrier frequencies are at the range of

radio frequencies, more specifically, we can call them, radio resources. Thus, the coordination problem is re-defined as the *radio resource management* (RRM) problem, that in broad part covers the problem investigated in this thesis.

2.2.2 Cellular Networks

The RRM problem can be defined from many perspectives but the problem definition that is accepted in Cellular Networks is allocating resources to links [Zan97]. In cellular networks, each cell is assumed to be independent from each other, such that the signals from one cell do not reach the other cell. Thus, the resources are allocated per cell. Conceptually, a cell is the same as the classroom in the initial example where a group of people are communicating with each other with the coordination of a group head. That head of the group in cellular context is the Base Station (BS). The BS coordinates the use of resources in a cell and responsible for the RRM. The members of the group in the cellular context is called the User Equipment (UE), that will be interchangeably referred to as user, device or sensor throughout the thesis. Throughout the thesis, a resource is treated as the atomic unit. Similarly, a packet requires a single resource to be transmitted. The problem of RRM is mapping the packets of users to resources. The efficiency of RRM is measured with the average number packets successfully transmitted per resource. This metric is called efficiency and throughput throughout the thesis.

2.2.2.1 Quality of Service

The rate allocation is not the sole goal of RRM. Many parameters that can be used to allocate radio resources is defined as Quality of Service (QoS). At its most simple version, a QoS can be just a rate requirement that can be controlled with Eq. (2.5). However, the rate, depending on the averaging period may lack the time perspective. If a QoS with a certain guarantee R, e.g. 90 percent of the time, is required, the Eq. (2.5) with an average SNR over infinite latency cannot be used any more. QoS can also encompass an average rate requirement for a certain period, dubbed as latency requirements L, e.g., an amount of bits that have to served before $L = 2$ seconds.

The latency is effected by many aspects of how the radio resource management is done. For example, one time taking aspect is the time required for coordination signals between BS and UE and another aspect is the re-transmission rules in case of a radio failure. Thus, latency is a parameter harder to optimize compared to rate. In this thesis we focus, though, specifically on the latency aspect of RRM.

The RRM is an optimization problem that considers QoS requirements from multiple UEs as a constraint input. If radio resources can be distributed in a way that fulfills the

requirements of all UEs, then the problem is solved. Algorithms addressing this problem are called *scheduling* algorithms. In cellular networks, the link that defines the communication from the UE to BS is called the Uplink (UL) while the communication from BS to UE is called the Downlink (DL). The scheduling algorithms are covered as UL and DL scheduling algorithms, for coordinated uplink.

In a cellular setting the DL and UL are asymmetric in terms of participants to the communication. While it is one-to-many in DL, it is many-to-one in UL. The members talking to the head of the group and the head of the group talking to the members is totally different in terms of coordination. If a resource is set for DL, any interested UE can listen to the BS and receive the incoming signal interference free.

The DL RRM has specific challenges such as, maximizing the DL capacity and minimizing delay [Mar05]. On top of this, as the base-station is a single device compared to many UEs, it has usually more complex transmitters. Multiple transmitters on one hand enables the BS to reach the capacity but on the other hand it complicates the processing of the received signal. We do not explore DL scheduling in the scope of this thesis.

The scaling requirement for hardware of UEs enforce it to be rather cheap compared to the BS. So the number of antennas in a UE is limited. This results in complex signal processing algorithms being not useful for UEs as a part of the UL RRM. Rather the coordination is a typical RRM problem for the UL resources [Saq+12].

There are two main approaches to the coordination problem in the UL, this can be summarized as allocating resources on a contention basis, or contention free. Contention basis relies on a distributed algorithm to coordinate the use of UL resources. With the contention free method, the BS obtains activity information of the UEs. Consequently, the BS allocates UL resources with respect to user activity. The activity information can only be obtained through another protocol called initial access protocol.

In the most current version of cellular networks that is the Long Term Evolution-Advanced (LTE-A), the UL activity information is obtained on a contention basis through a channel called Random Access Channel (RACH). The algorithm deployed in RACH is called Multichannel Slotted ALOHA MC-SA. This is an example for a scenario where the contention based access is used only for control signals, i.e. requests. In contrast, WiFi uses contention based access for all signals, and the scheme is called Carrier Sense Multiple Access (CSMA). Contention based protocols that directly work without exchanging activity information is called grant-free protocols. Their mathematical performance analysis is the same only with different assumption on packet sizes.

In Equation (2.8)

$$C_i = \Delta f \log_2 \left(1 + \frac{P_{ri}}{\sum_{j=1}^{n_{tot}} \delta_j P_{rj} + N} \right) \text{ where } i \neq j,$$

any access protocol tries to optimize the parameter δ_j or P_{ri} in a distributed manner for all the users with respect towards a specific goal. The goal can be to only have a single user active at a resource, i.e.,$\exists\, i \in \{1, \cdots, n_{tot}\}, \delta_i = 1 \forall\, j \neq= i, \delta_j = 0$. In this way the received power P_{ri} needs not to be optimized but only activity δ_j can be coordinated. The main idea is that even a small amount of interference can violate the allocated rate and the interference should be totally avoided. Such a model is called the *collision channel model* and is introduced in [MM85]. However, practical tolerance to a certain level of interference is possible. The model that involves conservation of the rate even subject to interference is called the *capture channel model* [GS87]. In this case, if one of the received signal powers is stronger compared to others, the interference on the strong signal does not deteriorate the signal and it can be decoded.

Considering signal processing based improvements, a complex receiver can decode the signal of interest from the received signals once it is successfully decoded. This is called the Successive Interference Cancellation (SIC) channel [Str+94]. This model allows the re-use of *capture channel model*. The power of the decoded signal is removed from the received signal power and the Capture Channel model is re-used if the power difference is sufficient for another capture. Capture channel tolerates different activities of UEs at maximum, while the collision channel tolerates the least. Recent work [NP12] has shown that SIC-Channel model enables achieving a capacity in uncoordinated uplink problem matching that of coordinated uplink. Meanwhile, the best practical algorithm for *collision channel model* has achieved less than half of the SIC Channel capacity [YG05].

The algorithmic challenges are similar for all channel models, as the main problem is to coordinate UEs in a distributed manner. So most of the results obtained for uncoordinated UL for Collision Channel model applies to SIC-Channel model also. However, the implementation of a receiver enabling a SIC channel model is quite complex and has limitations that needs investigation.

2.3 Summary

In this chapter we have introduced the fundamentals of wireless communication and radio resource management necessary for understanding the problem of uncoordinated uplink access. Specifically, we emphasized the challenges underneath the problem of uncoordinated

uplink access. This problem is motivated with an analogy to human communication without visual context. Next, we defined the wireless communication as distribution of discrete time-frequency resources in order to simplify the problem. Then, we have introduced possible mechanisms that the radio resource management can use to overcome uncoordinated uplink problem. Furthermore, we have defined how these techniques can be used for QoS for wireless communications.

In the next chapter, we introduce a real world scenario, intra-aircraft communications, that we use to define QoS requirements for radio resource management.

Chapter 3

From Flight Safety to Wireless Reliability for Intra-Aircraft Communications: A Motivation for Delay-Constrained Reliable Access

In this chapter we aim to explain practical aspects of the Quality of Service we introduced in Chap. 2. We consider a specific use-case of intra-aircraft communication. As aircraft industry is highly regulated, any system deployed in an aircraft has to follow a tight certification process. In this chapter we aim to bridge the Quality of Service constraints to the requirements of the certification process, embodying the requirements for radio resource management. Following, we investigate the commercial of the shelf chips for their suitability for such a certification process.

The structure of the chapter is as follows: In Section 3.1 the motivation for using wireless communication in aircraft is shared. In Section 3.2 a reliability assessment technique involving wireless communication for system failures in aircraft is introduced with a fault tree analysis. The analysis connects the system level parameters to communication parameters. In Section 3.3 performance parameters are summarized and in Section 3.4 wireless technologies considered for intra-aircraft communication are presented. In Section 3.5 the presented candidates are evaluated with the performance metrics based on the reliability analysis. This chapter is mainly based on our presented work in [Gür+15] and its extension as a book chapter [Gür+17a]. The results of this chapter, on one hand shows that commercial of the shelf systems can be used for restricted low number of users for intra-aircraft communication while on the other hand it motivates the investigation of scalability of the currently available wireless communication technologies.

3.1 Wireless for Intra-Aircraft Communication

The communication of systems inside the aircraft has always been a critical aspect of a safe flight. The sense of flying safely is not an intuitive feeling for human beings even though the accident reports prove that it is a lot safer than normal ground vehicle traffic [Aur08]. The deceived public opinion merged with safety concerns forces the regulators to impose strict safety regulations and certifications.

At the start of the epoch the control of aircraft was fully mechanical, so any input given by the pilot would be transferred to the concerning area with an increased force. Analog electronics converted any input to a mechanical input, e.g., Fly-By-Wire which was installed initially on a Concorde designed by Aerospatiale [TLS04]. Following, digital electronics enabled conversion of any kind of input to a digital message and triggering the action required anywhere in the system. This system introduced the first communication systems in the aircraft where the medium of communication is shared and different messages are interpreted at different applications. This increased the amount of functionalities embedded in a normal aircraft around 1980's e.g., A310. On the one hand, the removal of needed mechanical input opened areas such as the possibility of automated flights. On the other hand, the amount of communication infrastructure in the form of copper wires became a huge burden for the aircraft. For instance in A380, the copper wires had to be replaced with aluminum wires in order to make the communication infrastructure weigh less up to 50 percent.

Two facets of the wired communication can be summarized as the weight problem and the placement problem. The placement of wires is a constraint on deploying new systems since a new application may require new wire placement. Each wiring has to be planned and the planning time of the wiring infrastructure is costly. Introduction of a wireless communication system can easily solve this problem.

The replacement of a widely deployed communication system such as the wiring in the aircraft requires in-depth investigation. As it has been pointed out, (1) the reliability is the most important issue but many other requirements follow. The various characteristics of different applications provide a wide range of requirements, where one is (2) energy conservation as the equipment runs on batteries. While some applications require low power communications, (3) low delay is more critical for others and is affected by the (4) maximum payload. (5) Node density is an issue with the large amount of sensors that are to be introduced in the aircraft for passenger and application monitoring.

Considering these requirements we limit our study to 6 standards that are (1) the Wireless Sensor and Actuator Network for Factory Automation (WSAN-FA) [Sch+07], (2) the WirelessHART [Com+10] that is the wireless version of the HART standard, (3) ZigBee [Erg04],

the non-modified structure of the IEEE 802.15.4, (4) LTE [36313] due to its availability, (5) the ECMA-368 [All08] with the Ultra Wide Band (UWB) physical layer and lastly (6) the widely established Wi-Fi standardized as IEEE 802.11[Com+99], is chosen as a candidate due to the adaptability of the system for many applications.

Many surveys exists for wireless sensors [AE10], personal area networks [Cav+14] and sensor networks in aircraft [Har02] none of them covers all of the critical aspects for aircraft communications. In a use-case with similar requirements as intra-aircraft communications, the factory automation, there are also relevant studies. For instance, in [ISW12] the security aspect for is investigated while the reliability aspect is not covered in detail. In [GH+09] hardware limitations are investigated and delay constraint is neglected. The most relevant surveys for our investigation are [LSS07] [WMW05] but in their work, the factory automation related standards like WSAN-FA and WHART are left out. Most importantly none of these works has as a reliability perspective which is a must for the aircraft communications and with which we provide a detailed assessment of each candidate technology.

3.2 Reliability Assessment Framework

This section aims to introduce the mapping from certification requirements to quality of service requirements in terms of radio resource management. This is achieved in three steps: flight to application mapping, application to medium access mapping and medium access to transmission mapping. The three steps are introduced in the following parts of the section. The selection of these three steps among many is motivated by investigating the whole fault tree for a heat sensor application in the aircraft.

We assume that the communication inside the aircraft is set via a cellular topology. Delay constraints in multi-hop networks is investigated in terms of routing metrics in our work in [GK17a] and in terms of medium access control in our work in [Vil+16]. Multi-hop networks is not included in this thesis. The effect of the gateway node forwarding wireless packets to the wired domain in terms of delay constraints is investigated in our work in [Gür+19b], [GZK19] and [Zop+18]. This topic is not investigated in the scope of this thesis.

The fault tree analysis as in Fig. 3.1 helps us to breakdown the components of a failure. For the fault tree analysis, the addition $+$ represents the logical "OR" operation while the multiplication $\times$ represents the logical "AND" operation. As an example, we consider a Passenger Heat Sensor Application (PHSA). We aim to demonstrate the safety to QoS mapping methodology that can be later on used for any other application.

The first level of the tree is composed of the different systems failures forming the application. These are Power System Failure (PF), Sensor System Failure (SF), Control System

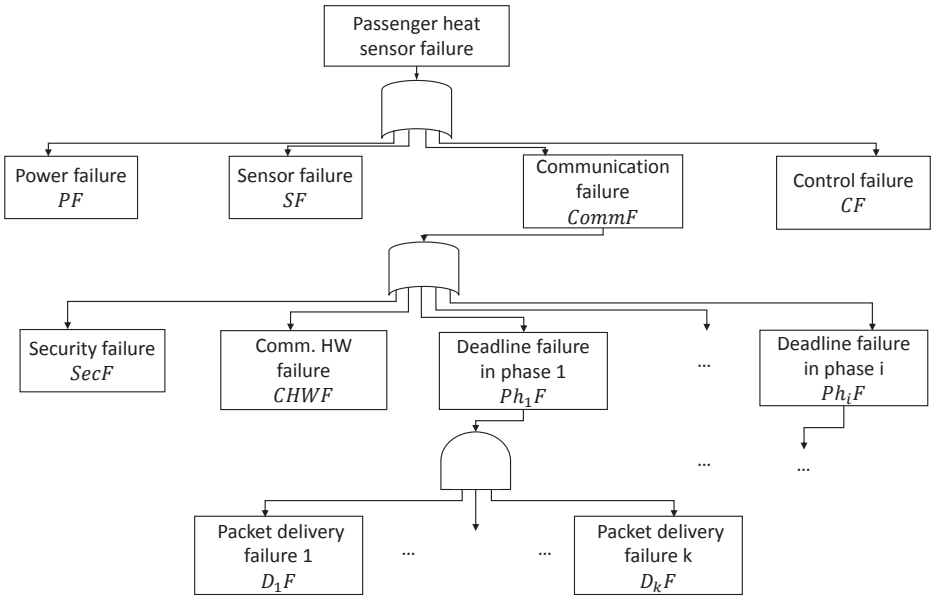


Figure 3.1: Exemplary application fault tree analysis: heat sensor application failure.

Failure (CF) and Communication System Failure ($CommF$). Each system failure probability represents the application failure as $PHSA$ in,

$$PHSA = PF + SF + CommF + CF. \tag{3.1}$$

For all use-cases we assume at least two order of magnitude smaller failure probability compared to the communication failure and we assume that their contribution can be ignored.

Furthermore, the communication failure ($CommF$) can be broken down further: (1) Communication Hardware Failure ($CHWF$), (2) the failure caused by a packet arriving later that the deadline T_W allowed by the application Ph_iF, as this failure is investigated differently for each flight phase i and (3) the Security Failure ($SecF$) where the communication security is breached. The multitude of delivered messages add another failure chance independently from each other. This results in a $CommF$ of

$$CommF = CHWF + SecF + \sum_{i=1}^{I} Ph_iF, \tag{3.2}$$

where I is the total number of flight phases as defined in [US 11] such as taxi, take-off, cruise and landing. Each phase can be broken down to combination of transmissions

$$Ph_iF = \prod_{k=1}^{Np} D_k^i F, \tag{3.3}$$

where the possible N_p transmissions from a device within a time-window T_W and k is the k^{th} transmission. The complete model of the fault tree is

$$PHSA = PF + SF + CHWF + SecF + CF + \sum_{i=1}^{I} \prod_{k=1}^{Np} D_k^i F. \tag{3.4}$$

The OR operation can be written with a sum, assuming the failure probabilities are close to 0 with two digits. In our analysis we only focus on the communication failure $CommF$. This simplifies the error rate of our application to

$$PHSA = \sum_{i=1}^{I} \prod_{k=1}^{Np} D_k^i F. \tag{3.5}$$

With a slight abuse of notation we move from the definition of failure to failure rates. We assume a deadline aware queue management protocol is used as introduced in our work [GK17b]. We replace the application failure $PHSA$ with application failure rate P_{PHSA} and the packet delivery failure $D_k^i F$ with packet delivery failure rate $P_{comm_k^i}$,

$$P_{PHSA} = \sum_{i=1}^{I} \prod_{k=1}^{Np} P_{comm_k^i}. \tag{3.6}$$

If the communication failure probability does not change within one time window we can write the AND operation without the product but as exponent. The OR operation can be simplified by calculating the success probability and converting that to the error probability,

$$P_{PHSA} = \sum_{i=1}^{I} {P_{comm_k^i}}^{Np} = \left(1 - \left(1 - {P_{comm^i}}^{N_p}\right)^I\right). \tag{3.7}$$

If the packet failure rate is constant between all of the flight phases I, we can drop the index I and treat the failures as equal. Given ${P_{comm}}^{N_p}$ is two orders of magnitude smaller than the inverse of the exponent, $1/I$, we can apply the linear approximation to simplify the failure probability,

$$P_{PHSA} \approx I \cdot {P_{comm}}^{Np}. \tag{3.8}$$

The fault tree analysis demonstrates a guideline to focus the wireless failure. In the following part we zoom in to investigate some parameters we introduced such as number of transmissions N_p. The same result is found without different flight phases in [BK06] for vehicular communication.

3.2.1 Medium Access Layer

For N users and transmitting for T_{SE} seconds, we define $N \cdot T_{SE}$ as the cycle time T_{cyc}. This summarizes the delay for the use of medium by N (all) users sequentially and fairly. T_{SE} can

be set with respect to the standard transmission rate, the packet size and the headers. Then we can extract the maximum number of transmissions before a deadline T_{TW} is violated, N_p using the previously defined cycle-time T_{cyc} as

$$N_p = \left\lfloor \frac{T_{TW}}{T_{cyc}} \right\rfloor . \tag{3.9}$$

The introduced model is used for investigating multiple technologies in the following parts. But initially, we detail the metrics and the parameters considered in the evaluation.

3.3 Metrics and Parameters

In this section we introduce the performance metrics and design parameters.

3.3.1 Design Parameters

The design parameters are the controllable parameters before a communication system is put in place.

3.3.1.1 Cycle Length and Packet Size

The cycle time is a design parameter of the layer 2 of the wireless technology. It varies among different technologies. The calculation of the cycle time is detailed for each technology.The multiplexing of the user on the medium is different on each standard and has varying overhead. For a fair allocation of resources a round-robin schedule is considered. Each user has a single transmission opportunity in a cycle.

3.3.1.2 Node Density

The node density is the number of users per cell, as we focus on a cellular topology. Hence, it is an advantage to have smaller cells for supporting denser communication.

3.3.2 Performance Metrics

The performance metrics are selected to measure the quality of service of the use-case.

3.3.2.1 Power Consumption

As some of the devices will be wireless sensors in an aircraft communication system, it is possible that they lack a powering cable and they operate on batteries. For such an application it is important that the communication uses only a small portion of the stored energy.

The power consumption depends on multitude of parameters. First, it depends on the carrier frequency. Higher frequencies requires more energy for more complex RF circuits. Second, synchronization requires periodic wake-up and extra signaling that drains additional power. Third, the complexity of the network protocol may result in signaling overhead, thus, extra power consumption.

3.3.2.2 Initialization Time

In case of a system restart or an emergency, the re-attach time of all of the wireless devices to the communication system should be bounded. Worst-case of this re-connection time is defined as initialization time. Two different ways of joining a network are: (1) Initially the network state is fixed and known to all users such that all of the users have dedicated slots. With a synchronization method, e.g. broadcast beacon, the user will align its message to its dedicated slot. Or (2) the network can use a resource request protocol. The (1) and (2) can be seen as a grant-based and connection based communication respectively.

3.3.2.3 Packet-pipe Reliability

We treat the wireless communication as a faulty pipe and investigate the required minimum reliability of this pipe on a packet basis.

3.4 Wireless Candidate Technologies

In this section we provide general information about technologies under consideration.

The selection of the candidate technologies are mostly based on the Commercial of the Shelf (CotS) systems. The reason for such a selection is to provide a practical solution for widely available chips.

3.4.1 WISA & WSAN-FA

Wireless Interface for Sensors and Actuators (WISA) is a factory automation standard built by ABB. It uses the PHY layer, IEEE 802.15.1, and deploys Frequency Hoping to avoid intra-technology interference. It is a proprietary protocol but the latest decision by ABB is to have it standardized as WSAN/FA with a set of improvements.

PHY specifications are not detailed here as our focus in on MAC layer. The standard deploys a time division multiplexing scheme. Each user has a dedicated downlink timeslot. Thus, the downlink delay is limited by the superframe length, 2048 ms. For separation of

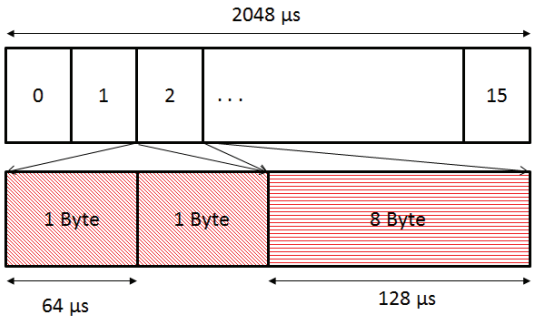


Figure 3.2: Superframe structure of WISA

uplink and donwlink a frequency division duplexing scheme is used. For redundancy, retransmission on 4 different frequencies is possible extending the cycle time to $8ms$. The payload considered for each time slot is 1 byte and the number of users is limited to 120 [Fro+14]. Alternatively, number of users can be limited to 60 for a payload size increase to 10 bytes [Val12]. It uses the CotS Bluetooth for PHY layer such that 1 Mbit/s data rate is expected.

As it is a sensor and actuator network protocol, it has more uplink channels than downlink which enables the sensors to report every cycle while the downlink channels are used for the control of the actuators.

In Fig. 3.2 the frame structure of WSAN/FA is depicted. The superframe of 2048 μs for a WSAN/FA supports two types of slot formats. First one is the small slot size of 64 μs which allows 120 users with 15 frequencies and 4 redundant transmissions. The second slot format has a twice bigger timeslot which allows 60 users due to increase in the slot size with a slot duration of 128 μs.

As the system is built for either 120 or 60 users and multiple slot assignments are not allowed, lower cycle time for lower number of users is not possible. On top of that, the advantage of low cycle time vanishes with increasing packet size. Here, the cycle-time includes all fragmented transmission of a single packet, as the full size payload does not fit in a single transmission opportunity. As the protocol is built for low payload, the overhead becomes significant with fragmentation of big packets.

3.4.2 ECMA-368

The European Computer Manufacturers Association standard ECMA-368, with a high carrier frequency, has small cells with 30 m range and, thanks to the wide bandwidth, a data-rate of 480 Mbit/s [Sav+13]. The carrier frequency of 3 to 10 GHz also comes with its disadvantages such as high fading and high power requirement due to complex physical layer chip structure.

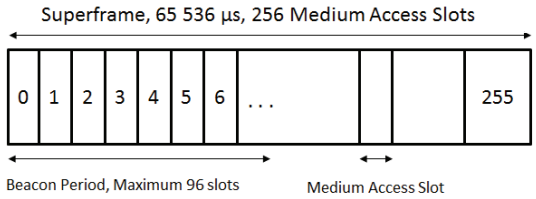


Figure 3.3: The superframe structure of ECMA

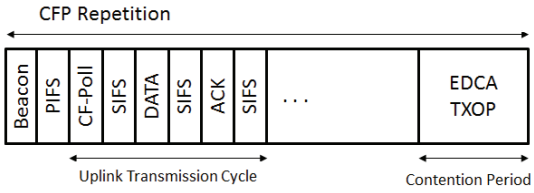


Figure 3.4: Reservation based access cycle in 802.11e

It has two resource management protocols. First, through a distributed reservation protocol the users access a guaranteed timeslot and secondly prioritized contention access can be used to access contention slots. There is no long-term dedicated timeslot, there are Medium Access Slot (MAS) which are allocated dynamically [Fan09]. There are 256 MAS in a ECMA superframe where first 96 MAS are filled with beacons for each attached device, The beacon holds information about the structure of the upcoming superframe. Following that, each user knows how many slots it is able to use in the upcoming superframe. After the end of a superframe a new superframe starts with new settings that will be conveyed to the system via the beacons. One advantage of ECMA is the variable data length. The minimum packet length with the most reliable coding rate is 1.6 Kbyte which is more than the required packet length for any sensor activity.

On the PHY layer it uses a multiband orthogonal frequency division multiplexing, which enables 110 subcarriers in a superframe. For interference mitigation: frequency domain spreading, time-domain spreading and forward error correction codes are used.

The superframe structure is what determines the layer 2 data rate and the cycle-time as can be seen from Fig. 3.3. The superframe structure creates a cycle time of $65ms$ with 256 MAS of $256\mu s$. This cycle time limit can be overcome with the allocation of multiple slots to a user in a superframe, but this will reduce the number of users in a timeslot. Due to allocated 96 beacons at the start of the superframe and the criteria that requires each of the users to possess a beacon, the number of users in a cell is limited to a maximum of 96 [Lei11].

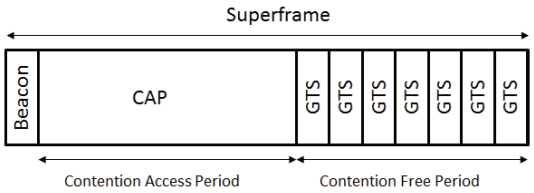


Figure 3.5: The superframe structure in IEEE 802.15.4

3.4.3 IEEE 802.11e

The PHY layer of IEEE 802.11 operates in the ISM frequency band. Interference is a problem due to the unregulated nature of the ISM band. Cell range is normally around 100 meters but can be increased and decreased with transmit power arrangements.

IEEE 802.11 uses contention based access schemes. However, in the 11e amendment, the option of contention-free access is added to the standard. There are three resource management protocols: (1) full contention access as legacy, (2) semi contention access with reservation (Enhanced Distributed Channel Access, EDCA) and (3) full contention-free access HCF Channel Access, HCCA [Vie+13]. The contention based access works as legacy. For (2) a request can be placed with the contention based scheme to reserve upcoming contention-free slots. It is possible that the contention-free slots are allocated to certain users without reservation. And (3) uses a polling technique to guarantee contention-free access. As we have focused just on contention-free access on this survey, for a fair comparison we will consider HCCA only. In Fig. 3.4 the HCCA frame structure is depicted. The figure reflects the amount of overhead required to guarantee the co-existence of 3 resource management protocols.

For transferring a packet of 40 bytes, the total time required is 1.6 ms per user with 11 Mb/s data rate. If we chose a scenario where no packet loss is considered and the service interval time is set to 100 ms we can calculate the cycle-time for each user. The standard has a parameter called *service interval time* that is the time between two consecutive beacons. With multiple users contention free period (CFP in Fig. 3.4) can be extended up to this parameter.

With 60 users the 100 ms timeslot length is reached. Before that the system benefits from allocating multiple Controlled Access Period (CAP)s to each user. Cycle time here is defined as the time before next packet sending opportunity.

3.4.4 IEEE 802.15.4

IEEE 802.15.4 is a standard originally designed for personal area network applications but in a short amount of time it found itself in the area of Internet of Things, factory automation and many more. The modification that enabled this compared to other standards was the

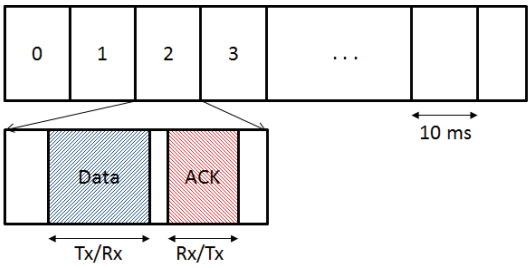


Figure 3.6: The timeslot structure in WirelessHART

lightweight protocol stack which enables low power communications. It also uses the ISM band as IEEE 802.11. Direct Sequence Spread Spectrum (DSSS) decreases the effect of any interference. Due to the low spectral efficiency the standard does not support a data rate higher than 250 kb/s but that is enough for sensor applications.

This standard, similarly to IEEE 802.11, is not designed for high reliability, and, due to that, it also lacks an optimized TDMA scheme. However, there is up to seven Guaranteed Time Slots (GTS)s. The packet size can be modified with setting the parameter "*macsuperframeorder*" (SO). As illustrated in Fig. 3.5 the superframe consists of Contention Access Phase CAP and Contention Free Period CFP. CFP can be enabled while CAP has a fixed minimum duration of $440 \cdot 2^{SO}$ bits. Additional $46 \cdot 2^{SO}$ bits is reserved for beacon and interframe spacing. This limits the capacity available for CFP to $474 \cdot 2^{SO}$ bits in a superframe [Che+09]. On the other hand, not the packet size but the maximum number of 7 users per superframe is the critical limitation. The CAP is neglected as the focus is on CFA. If the SO is fixed to 2. Cycle-times with 15 ms is possible for a packet length of 32 bytes.

3.4.5 WirelessHART

WirelessHART, WHART, is the protocol designed for process automation using the physical layer of IEEE 802.15.4. As a major modification to the IEEE 802.15.4 it has a TDMA based radio resource management. The improvement is based on the consideration of critical sensor networks. This use-case requires small and frequent chunks of data for uplink compared to downlink. So timeslots are arranged for parallel communication.

WHART supports up to 16 channels at a 10 ms timeslot [Dan+13]. In each timeslot an ACK response is expected. The frame structure is depicted in Fig. 3.6. The MAC payload in a WHART packet contains 133 bytes. The ACK packet is 26 bytes thus the total data exchange in a WHART timeslot is 159 bytes, with a data rate of 250 kbps [PC09]. The ACK can be disabled to trade-off delay versus reliability.

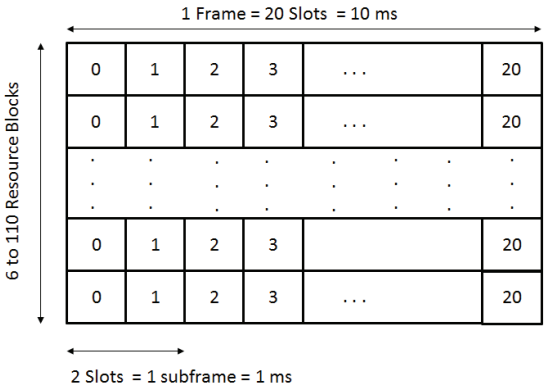


Figure 3.7: The frame structure of LTE [LA11]

16 users can use 16 orthogonal channels such that a cycle-time of 10 ms is sustained. However, as we show in Chapter 6, the parallel use of channels may cause interference. On a comparable level to previous protocols 60 users can be served within 40 ms.

3.4.6 LTE

3rd Generation Partnership Project Long Term Evolution (3GPP LTE) is the latest stable standard for mobile networks. Having access to only a limited band for wide use-cases, it proved to have a high spectral efficiency.

LTE uses multiple private bands which negates inter-technology interference. It deploys time and frequency resource blocks which allow flexible and efficient radio resource management. In Fig. 3.7 the frame structure of LTE is depicted. A frame is divided into 20 slots, 2 slots form a subframe. A resource block (RB) is formed of 12 carrier frequencies and 7 symbols and it is the minimum resource block of the system. A resource block can transmit up to 40 bytes of payload data with a large overhead [BK12]. Each subframe takes 1 ms. Therefore a minimum Transmission Time Interval will be 1 ms.

We will investigate the uplink resource management in LTE. In the TDD subsequent subframes are duplexed between uplink and downlink. This can be varied with different TDD configurations. We use in our analysis the TDD configuration 0 [BK12] where two of the subframes are downlink against six uplink and two special subframes in a total of 10 subframes. Special subframes can be control information or Random Access Channel for synchronization.

As presented in [DJ10], we try to calculate the worst cycle-time in an LTE frame of 10 ms for varying number of users. We use 25 resource blocks. Beyond 10 ms, the number of users and latency tend to increase linearly with increasing cycle-time. We see that due to the

Table 3.1: Compared Protocols

Technology	Cycle (ms)	Packet (Byte)	Power	Initialization	N. Dens.
WISA WSAN/FA	2	8	Low	-	60
WLAN	100	40	High	seconds	60
ECMA-368	22	1.6k	High	seconds	60
ZigBee	135	20	Very Low	seconds	60
WHART	40	133	Very Low	seconds	60
LTE	5	40	High	mseconds	60

number of RBs, 150 users are the maximum that can be supported. 75 users have a cycle-time of 5 ms. The achievable lowest cycle-time is 3 ms due to the TDD configuration. In FDD this would be as low as 1 ms.

3.4.7 Technology summary

The cycle-times are investigated fixing the number of users to 60. Tab. 3.1 shows the supported packet length and cycle-time of each technology. The ECMA has the highest packet size. The WSAN/FA has the lowest cycle-time of 2 ms. When both the packet length and the cycle-time is compared at the same time, LTE has the best trade-off. The problem with LTE, UWB and HCCA is that they do not support low power devices. The limitation of WSAN/FA with the a maximum of 120 users and of ECMA to 94 users, is a problem for scalability. Even though IEEE 802.11 does not have a protocol defined limit, the algorithm has a practical limit around 60 users. Other technologies do not have any standardized limits.

3.5 Evaluation

In this section we highlight two major problems for fulfilling the quality of service requirement as a result of the model we proposed. More specifically, we aim to extract communication reliability requirements of each technology given the application requirements. Thus, we use an exemplary application of cabin lighting which requires a 36 byte messages each 200 ms. This application has a safety requirement of 10^{-5} failures per flight. There are 60 sensors for this application that is the number of users in our evaluation.

Two main aspects of the quality of service problem is the communication reliability and the number of users. We can investigate one by fixing the other and vice-versa. Initially, we fix the number of users to 60 and investigate the required communication reliability.

We assume perfect layer 3 fragmentation without overhead, and if the packet size is bigger than supported by the technology, it is fragmented and sent in consequent fragments.

Table 3.2: Top-Down Reliability Comparison

Technology	WSAN/FA	ECMA-368	802.11e	802.15.4	WHART	LTE-A
T_{cyc} (ms)	8	22	100	270	40	5
N_p	25	9	2	0	5	40
P_{comm}	0.5916	0.2326	0.0014	0	0.0724	0.7203
Max. Users	1500	846	120	42	320	3000

Fragmentation is used to adjust cycle-times for the technologies in order to meet the required packet length. Using the values given in previous section some technologies cannot support the payload in one cycle such that the we need 4 and 2 times the cycle-time for WSAN and IEEE 802.15.4 to deliver the payload, respectively. This results in updated cycle times for both technologies of 8 ms and 270 ms, respectively. We see that IEEE 802.15.4 is not able to meet the delay constraint as one cycle-time is larger than the time window. The next step is the calculation of transmission attempts N_p through Eq. 3.9 which results in P_{app} with Eq. 3.8. As same application is considered the flight layer mapping is done only once. Using N_p values of each technology, this is then mapped to the required communication reliability P_{comm}.

Table 3.2 shows that LTE requires P_{comm} of less than 70%. As discussed in previous section, LTE turned out the be the best solution. It is followed by WSAN thanks to low cycle-time that allows for multiple retransmissions. Feasibility of these required reliability values is investigated in Chap. 6. The short conclusion is that time-diversity is a promising solution to achieve safety requirements of aircraft applications.

In a second step, we assume that the radio transmission has perfect reliability and no errors occur on the air. Through this we inspect the maximum number of users supported by each technology for a delay constraint of 200 ms. In this comparison we do not need the number of re-transmissions as the communication has perfect reliability. But still we need to scale the cycle-time with increasing number of users until it reaches the delay constraint. This gives the maximum number of users supported with perfect communication reliability in Tab. 3.2 on the last row. We did not consider signaling related or protocol related limitations. These are summarized in the previous sections. In short, the results show that scalability is also a limitation in most of the technologies even though a lot of bottlenecks are abstracted.

3.6 Summary

In this Chapter we provided a model to connect the certification requirements existing in aircrafts to Quality of Service (QoS) requirements that will be imposed on wireless technologies. We have used simplifying assumptions to compare the best-case scenario for the commercial

off-the-shelf technologies. We have demonstrated that even with a perfect physical layer assumption, some technologies are totally unable to fulfill the QoS requirements like the standardized contention free IEEE 802.15.4 and the others suffer from scalability.

Scalability is limited mostly due to scheduled resource allocation. To overcome this problem, in some technologies resources are allocated in a shared manner contrary to what we assumed in this Chapter, namely via the contention based access. However, to the best of our knowledge no contention based access protocol exists for strict delay constraints. We investigate and improve the access decisions of users to overcome this challenge in Chapter 5. However, an algorithmic solution is only feasible if the number of users accessing the system is in accord with the number of resources available in the system.

In order to provide effective guarantees, the user activity should be known by the access algorithm. In Chapter 4, we investigate the effect of user activity on the reliability of contention based access schemes. We demonstrate that user activity estimation is a solution to overcome the scalability limitation caused by the sporadic activity.

Chapter 4

Evaluation of User Activity for Delay-Constrained Reliable Access

In Chapter 3 we have broken down the system requirement to radio requirements. Following, we have identified that the two major sources of failure is the reliability of a single packet transmission and the multiplexing of the users on the radio resources. Reliability of a single packet transmission is investigated in Chapter 6. In Chapter 4 and 5 we focus on the latter problem, that is the multiplexing of the users to radio resources. In Chapter 5 we investigate an algorithmic approach for multiplexing users for strict stochastic and deterministic delay bounds. In this Chapter we start with evaluating state of the art protocols with respect to defined challenges and later on focus on the effect of user activity for this problem.

Reciting from Chapter 2, in a collaborative setting, the broadcast is an advantage for downlink communication such that every user can receive the message at the same time. However, for the uplink this means that users cannot communicate simultaneously or they will interfere with each other.

The BS, knowing the traffic profile of users, can broadcast the radio resource allocation for users to avoid interference. In other words it can set the activity parameter δ of each user as introduced in Chap. 2. This concept is called scheduling and can be formulated as an optimization problem constrained with user requirements and optimized with respect to efficient use of resources. However, the BS does not allocate resources in a proactive way, it reacts to the users. The reaction is triggered by requests of the users. Following, the BS distributes resources announcing it with a broadcast. In order to limit the overhead of distributing the scheduling decision, the BS does the scheduling with a certain frequency. Hence, the requests are accumulated and the decision can be made at the same time for many requests.

The collection of requests use a shared channel, named random access channel, in the current mobile networks. The users re-transmit until successful. However, the re-transmission behavior may result in continuous interfering of users with each other resulting in an unstable behavior. One of the reasons for such an unstable behavior is a burst of requests at the same times. The bursty behavior may be a result of machine communications due to correlations among users compared to the human communication that is independent in short time scale. Thus, the prior analysis of the stability of radio resources is not valid anymore and new solutions have to be developed regarding analysis of the new user activity.

The current solutions in random access algorithms encompass smoothing of the bursty distribution through shaping algorithms and expanding resource pool for re-stabilizing the access algorithm. Even when relaxed delay constraints are considered such as 10 minutes, these solutions fall short to provide a reliable access, or use abundant resources as we demonstrate in Sec. 4.2 of this chapter. The content in Sec. 4.2 is mostly from our work in [Gür+17b] and [GAK17a]. Sec. 4.2 demonstrates the problem of over-dimensioning as the only used solution. To solve this problem in Sec. 4.3 we undertake the problem of optimal resource provisioning by guaranteeing higher resource efficiency for delay constrained access, by theoretically estimating the activity of users. Sec. 4.3 is based on our work in [GKS19]. We show that estimating the activity of users improves resource efficiency when tight delay constraints are considered or with bursty arrivals. A practical user activity scheme is developed and analyzed in Sec. 4.4 where we propose the use of an admission control to guarantee delay-constrained reliable access in a scenario with limited resources such as an LTE cell, as based on our work in [Gür+19a].

4.1 Background and Related Work

In this section a short background on the radio resource management in cellular networks is given. Following, the related work that evaluates the effect of M2M activity, estimation performance and delay-constrained algorithm is included to emphasize the contributions.

4.1.1 System Model

We focus on a single cell with a homogeneous population of n_{tot} users that access the common base station (BS). The traffic is assumed to be uplink only. The users are randomly and sporadically activated, and their activity is modeled via a batch arrival of N_{a} users. In general, n_{tot} (or some upper bound on it) is assumed to be known, while N_{a} is a random variable. The time-frequency resources in the uplink are divided in a grid consisting of time-frequency slots (denoted simply as slots in further text), and without loss of generality, we assume that

the slot-bandwidth and slot-duration are of a unit size. We use the collision channel model i.e., channels have 3 distinct states, idle (0, no request), singleton (1, 1 request) or a collision (e, $>$1 requests) simplifying the interference effect. Unless physical layer enhancements are assumed the central entity cannot differentiate two or more users and treats them equally. We assume no capture capability. We also assume an instant feedback. Implementation of such feedback channels is discussed in previous work [GAK17a]. In order to distinguish, re-transmitting users and initially transmitting users, we will use the term backlogged user for the collided users and initial arrival for the first transmission of a user.

4.1.2 Machine-to-Machine (M2M) Traffic

In case of an emergency alarm, re-synchronization event, or system start e.g., in an aircraft, the n_{tot} users will try to simultaneously access the resources. The behavior of the users can be in general investigated in three different categories: Delta, Poisson and Beta arrival.

Delta An elementary M2M traffic model, the so-called Delta model, lets the n_{tot} users access in the same single slot, i.e., the synchronous arrival period (activation time period) is $T_A = 1$.

Poisson The arrival distribution of N_{a} can be modeled by a Poisson distribution if a set of independent users are considered. For the Poisson distribution, we have

$$\Pr[N_{\text{a}} = n_{\text{a}}|\lambda] = \frac{\lambda^{n_{\text{a}}}}{n_{\text{a}}!}e^{-\lambda} \tag{4.1}$$

where the mean number of arrived users is $\mathrm{E}[N_{\text{a}}] = \lambda$, assumed to be known.

Beta If the users are expected to react similarly to a timely event, then their traffic profile should be correlated. This correlation can be reflected with a Beta distribution. A sophisticated M2M traffic model can be based on the beta distribution [Kim+14; LAA14a; 11]. According to the beta distribution, the probability that a device is activated in time instant $t \in [0, T_A]$ is given by

$$p(t) = \frac{t^{\alpha-1}(T_A - t)^{\beta-1}}{T_A^{\alpha+\beta-1}\text{Beta}(\alpha, \beta)} \tag{4.2}$$

where $\text{Beta}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt$, α and β are shape parameters, and T_A is the activation time. We assume the 3GPP model [3rd00], where $\alpha = 3$, $\beta = 4$ and $T_A = 10$ s. Assuming that the activation time is discretized into L time units[1], the probability of n_{a}

[1] I.e., we assume that the arrivals are gated in batches of L time units.

Table 4.1: Summary of main model notations.

	User parameters
n_{tot}	Number of users (machines) = total number of user requests
T_A	Activation time period of the n_{tot} user requests
n_{a}	Active users at one slot
N_{a}	Random variable for active users at one slot
	System Parameters
M	Number of preambles
Γ	Number (const.) of back-off slots with index $\gamma = 1, 2, \ldots, \Gamma$
B_γ	Number (rand. var.) of users in given back-off slot γ
$\Pi_{\gamma,\pi}$	Number (rand. var.) of users on given preamble π in given back-off slot γ

arrivals in interval t_s, $t_s \in [0, \cdots, \frac{T_A}{\text{L}} - 1]$, can be approximated as,

$$\Pr[N_{\text{a}} = n_{\text{a}} | t_s] \approx \binom{n_{\text{tot}}}{n_{\text{a}}} P[t_s]^{n_{\text{a}}} (1 - P[t_s])^{n_{\text{tot}} - n_{\text{a}}} \tag{4.3}$$

where the total number of n_{tot} users is assumed known and $P[t_s]$ is given by

$$P[t_s] = \int_{t_s \text{L}}^{(t_s+1)\text{L}} p(t)\, dt. \tag{4.4}$$

The exact probabilities would require a partitioning based analysis. This approach is reminiscent of the one taken in [Lan+13a], where the discretization yields a time-modulated Poisson process, whereas in our case we deal with a binomial one, as indicated by Eq. (4.3).

4.1.2.1 Synchronous Arrival Detection

Access protocols for M2M traffic generally rely on some M2M traffic (synchronous arrival) detection mechanism to switch from regular (non-M2M) operation to the M2M random access operation mode and back to the regular mode when the M2M traffic subsides. For instance, the eNB can monitor an energy detector that detects a high number of MSG3 collisions, indicating the onset of M2M traffic [MSP14a]. For specific system settings, there may be outside triggers for switching to M2M traffic mode, e.g., alarms in smart meters [Che+12], or specific alarms or system start-up signals in an aircraft cabin. However, a system can also work dynamically adapting the operation on the fly for M2M and non-M2M through estimation techniques. This option is evaluated in the last section of this chapter.

4.1.3 Radio Resource Management in Cellular Networks

In LTE a hybrid radio resource management is utilized to allocate resources for users. This hybrid approach splits the request and the payload on different phases. When a user has a packet to transfer, first the BS is informed about the request using a Random Access Channel (RACH). RACH is a shared channel and user may fail due to lack of coordination if both users access the same resource at the same time. The failure is referred to as a collision in the rest of this chapter. Following the request reception, the BS allocates radio resources for the UL transmission with a downlink message. The payload can be transmitted on this resource without contention. The hybrid scheme is resource efficient for large payload and low request frequency. In case the payload decreases, the overhead of the request becomes significant. Or similarly, if the request frequency increases the contention based access becomes inefficient and takes too long. As these characteristics exactly represent the profile of M2M communication, the random access channel performance is evaluated in this chapter first in LTE scope with hybrid access scope followed with 5G adaptation that is the contention based access scope called the grant free access.

4.1.3.1 LTE Random Access Channel (RACH)

As specified by the 3rd Generation Partnership Project (3GPP), the random access scheme of the Long Term Evolution (LTE) wireless standard [16] is based on a variant of Slotted ALOHA with four message transfers.

A resource block (RB), the basic unit of the resource grid of LTE, is a defined group of sub-carriers within a slot. As any other LTE channel, the RACH works by using multiple assigned RBs. On top of that, any RACH message is sent by means of orthogonal preambles. These preambles are Zadoff-Chu sequences [Pop92], which are orthogonal to one another. Therefore, several preambles can be transmitted at the same time without causing decoding problems at the receiver.

The operation of the Random Access Procedure (RAP) of LTE is illustrated in Fig. 4.1. RAP is initiated by either the Radio Resource Control or PDCCH order on the user side, which is detailed in [16]. First step in the RAP is the preamble selection, which is controlled via the parameters, the Physical RACH (PRACH) Mask Index and *ra_preambleindex*. In case these parameters are not set, the user selects and transmits a random preamble. After the transmission, the user calculates on which resource block it will receive the response inferring it through the selected preamble and waits for it. Then, an active preamble is detected by the eNodeB, and it answers with a Random Access Response. This answer includes (1) the UL Grant that describes which preamble is allocated to which UL Shared Channel, (2) the target power for the MSG3, and (3) the timing advance to correct the synchronization of

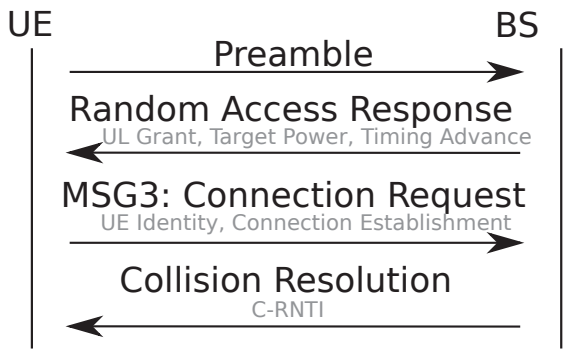


Figure 4.1: The messages exchanged for Random Access Procedure. The abstraction of first two messages and the periodic possibility to re-use RACH enables this process to be seen as a Slotted ALOHA.

the user. The user expects this message within a transmission time interval if MSG3 is not received then the first step is repeated. If successfully received, the user answers with its own identity and connection establishment details in MSG3. After that, the user waits for a *mac-ContentionResolutionTimer* until it receives the Collision Resolution (CR) from the eNodeB, which includes the Cell Radio Network Temporary Identifier (C-RNTI), mapping to a certain set of RBs. Successful reception of CR concludes the RAP. If the CR message is not received, the process starts over again from the first step. The reason of not receiving CR is usually two users interfering with each other causing a collision. The collision happens with the MSG3. Assuming successful reception of message 1 and message 2, as this is determined by the transmission power settings, we can focus on the message 3 and 4. Focusing on the these two messages we can abstract the RAP as a Slotted ALOHA (SA) system. If we consider multiple preamble selection possibilities, this creates a multi-channel SA (MC-SA).

In Fig. 4.2 the LTE RACH is depicted such that each column illustrates one subframe used by the RACH (called Random Access Opportunity or simply RAO), and each row represents a different preamble. As usual with SA, if a collision occurs, the user detects it with a timer, selects a random back-off time, and tries again. Stability is guaranteed by limiting the number of re-transmissions. Following, after certain number of trials a packet has to be dropped, such that it will never be transmitted. This is unacceptable for high reliability requirements. For simplicity of exposition, we model the LTE RACH to operate on a slot-by-slot basis, thus as a multichannel Slotted ALOHA, whereby each slot provides M preambles.

4.1.3.2 Slotted ALOHA

Commonly, LTE RACH is abstracted as a Slotted ALOHA system with n_{a} users and M preambles (which are equivalent to channels). In a given slot, each user uniformly selects one of the M preambles for transmission with a uniform probability of $1/M$. The probability that

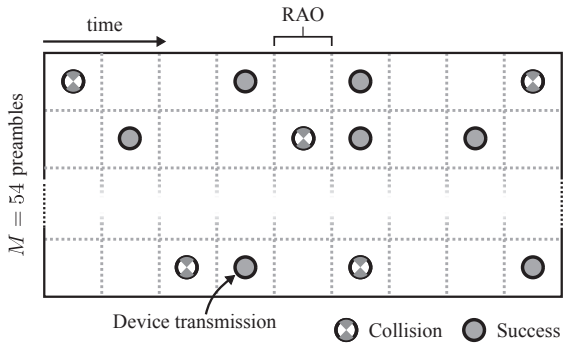


Figure 4.2: Grid depiction of the transmission of the preamble within the first step of the LTE RAP. As collided users decide which preamble to re-select in a independent manner.

Table 4.2: The variation of success probability p_s varying with number of preambles kept equal to number of users as M.

M	2	3	4	5	6	8	10	12	14	16	18	20
p_s	0.5	0.45	0.42	0.41	0.40	0.39	0.39	0.39	0.39	0.38	0.38	0.38

a given preamble contains one successful transmission is

$$\binom{n_{\mathrm{a}}}{1}\left(\frac{1}{M}\right)\left(1-\frac{1}{M}\right)^{n_{\mathrm{a}}-1}. \tag{4.5}$$

If we control the system such that the number of users n_{a} contending for transmission in a slot is equal to the number of preambles M, i.e., $n_{\mathrm{a}} = M$, the probability of successful transmission becomes

$$p_s = \left(1-\frac{1}{M}\right)^{M-1}. \tag{4.6}$$

Tab. 4.2 summarizes that the probability of successful transmission on a given preamble in a slot, and thus the throughput of the slotted ALOHA access system, increases with decreasing number of preambles M, while controlling the number of users n_{a} to match the number of preambles ($n_{\mathrm{a}} = M$).

4.1.3.3 Framed Slotted ALOHA

The slotted ALOHA works on a slot by slot basis. Each user after accessing a slot makes a decision when to re-transmit the packet. If a user decides to re-transmit immediately, the initial arrivals transmitting on the next slot interfere with this user. The interference can be avoided if coordination can be achieved between these users. A windowing approach, where users decide on a basis of n slots instead of a single slot, ensures that no new users can access the frame [Sch83a]. All the new users have to wait until the start of a new frame. This adds an initial delay on average half of the size of the frame. However, variations in the user activity

is smoothed and if the window size is large enough, a similar behavior in each frame can be expected. A smoother activity results in decreased variance in number of users per slot. Thus, the user behavior estimation techniques are more accurate and can be deployed.

Another way to deploy an algorithm that reacts to the number of users is to use tree algorithm.

4.1.3.4 Tree Algorithms

Tree algorithms are designed to stabilize the resource efficiency for uncoordinated uplink radio resource management. The algorithm assumes certain capabilities in the communication system such as: slotted time, collision channel and instant central feedback.

In the following, the operation of selected types of Tree Algorithms is briefly explained. The principle behind the tree algorithm is a user splitting strategy. First, users access a slot randomly. If multiple users access the slot at the same time the result is a *collision*. After the initial collision, all the initial arrivals are blocked. After the collision, users draw a random number, e.g, either 0 or 1 for Binary Tree Algorithm (BTA). Those which selected 0 are allowed to transmit in the following slot and those which selected 1 wait until at least a second slot.

The feedback message reports the outcome of a slot to all users. If the feedback is a success, then all of the users decrement their waiting counter by 1. If the feedback is a collision, users increment their waiting counter by one. When the feedback is a collision, the collided users select one or zero randomly to increment their waiting counter by that value. Each user transmits when their waiting counter is 0. The random splitting is repeated after every collision until no collision appears. Extensions from binary to Q-ary is also possible and more details can be found in [Mas81], [MF85].

The operation of a BTA can be depicted as a tree diagram, like the one shown in Fig. 4.3. In such a diagram, each group of users with the same split decision is represented by a *node*. The number inside each node reflects the number of users that have reached that node. In case of collision, that is, if the number of users in the node is greater than one, two new branches sprout from the collided node, since the users are divided into two new groups. The numbers by each branch represent the two possible choices that a user can make.

Tree algorithms are a reactive solution against user activity to regulate the use of resources through feedback. However, a proactive way against user activity is collision avoidance that can be used to proactively control users with feedback before their initial access such as in access barring.

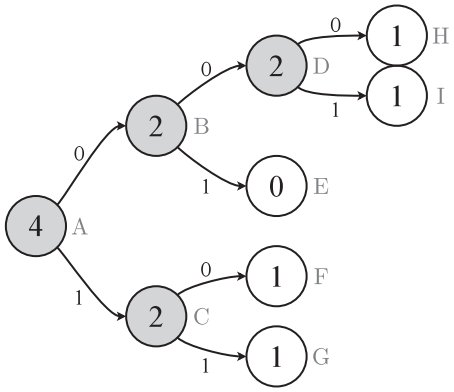


Figure 4.3: Tree representation of an example of a Binary Tree Algorithm ($Q = 2$), with 4 initial users.

4.1.3.5 Access Barring

In access barring, the eNB broadcasts an Access Class Barring (ACB) parameter between zero and one to all users. A user independently draws a random number between zero and one and compares its random number with the ACB parameter. If the user's number is smaller than the ACB parameter, then the user re-tries until it passes the access barring. If the user's number is larger than the ACB parameter, the user immediately advances to the LTE random access.

Dynamic Access Barring algorithm (DAB) [Dua+16] adjusts the ACB parameter dynamically based on the collision history and the resulting estimate of the number of user requests that will advance to the random access contention in the upcoming slot. DAB broadcasts the updated ACB parameter to the users in each slot. Pre-back-off is a similar technique that requires users to wait a certain time before their initial transmission.

Techniques used in LTE are mostly covered, in the following part the grant free access for 5G RRM is introduced.

4.1.3.6 Grant Free Access

In previous sections we have introduced contention based access to collect requests of the users. However, a contention based access for user payload is also possible. This is a logical option with smaller payloads. The payload based contention is called grant free access. The name comes from the hybrid access logic where the requests are sent on a contention basis to receive a grant for payload transmission. As this scheme disregards grants, it is named as grant free access.

Analytical models used to evaluate delay of a packet or a request is the same for random access and grant free access. The only difference is the assumption for a slot size matching

the size of a packet or a request. This also makes the payload size a part of the analysis. So assuming homogeneous payload among applications makes the analysis similar to the request based analysis.

4.1.4 Related Work

We briefly review the general area of random access strategies for M2M traffic. We then review in detail the related studies on collision avoidance and on collision resolution. Furthermore, the state-of-the-art access algorithms are evaluated against delay-constraints. Lastly, the admission control related work is included.

4.1.4.1 General Random Access Strategies for M2M Traffic

Wei et al. [WCT12; WCT13] have analyzed LTE random access for synchronous M2M traffic. In a follow-up work, Cheng et al. [Che+15] and Wei et al. [WBC15a] have proposed to estimate traffic arrivals and to dynamically allocate resources for M2M traffic. Cheng et al. [CLL11] have examined prioritization to serve M2M traffic according to its delay tolerance. Inaltekin and Wicker [IW08] have conducted a Nash equilibrium analysis of M2M arrivals with a game-theoretic perspective of selfish users reaching the random access channel. Lioumpas and Alexiou [LA11] investigated the effects of uplink scheduling for mitigating M2M arrivals. They proposed dynamic allocation of different random access resources to better accommodate the variable M2M arrivals. The general strategy of allocating additional random access resources for M2M traffic has been examined from a number of perspectives. A resource control perspective has been considered in [ST12] while a game theory based perspective has been investigated in [Pan+14]. The perspective of reporting based resource allocation has been introduced in [MSP14a] while delay requirements were considered in [Jia+13; OHL15] and a context-aware perspective has been considered in [PLW16].

Alternative strategies for the challenging synchronous M2M traffic have been explored in several studies. Based on examining the signaling for random access, data transmission schemes have been proposed in [And+13; Dhi+14; She+11]. Grouping-based radio resource management [LCL11; CZB14] has been proposed to trade off between different user groups. Prioritized random access (PRADA) [Lin+14] generalizes the grouping approach. Learning mechanisms for random access slot selection have been explored in [BMG14]. A code expansion random access strategy has been introduced in [Pra+12].

4.1.4.2 Collision Avoidance for M2M Traffic

For one-shot M2M arrivals, Duan et al. [DSW13; Dua+16] have examined Dynamic Access Barring (DAB) in an LTE system. Duan et al. dynamically modify the LTE Access Class Barring (ACB) parameter according to the collided and successful preamble transmissions. Effectively, DAB strives to modify the arrival rate so as to maintain a uniform arrival rate to the random access channel. DAB requires a new parameter transmission to all registered devices in the network at the end of each frame. Ko et al. [Ko+12] have exploited timing advance information in the random access messages to develop a novel access scheme for fixed location devices. The resolution of the timing advance information requires two contending users to be at least 156 meters apart. Wang and Wong [WW15a] merged DAB with the timing advance information approach. The parameters of both techniques were molded into a single optimization problem and the effects of the parameters were investigated under heavy M2M traffic.

Group paging, where a set of distributed user nodes need to be connected to a central eNB node, is closely related to synchronous M2M traffic models. Jiang et al. [JWD14] introduced a pre-backoff approach for group paging and analyzed the optimal paging size for a given initial number of M2M devices. However the suggested technique has a scaling problem due to the LTE backoff mechanism after a collision [Har+15]. The scaling problem is partially solved in [DW15] by forming different clusters for group paging and treating the cluster heads at a different phase as another group paging process. This solution assumes that the cluster heads are able to carry the data transmitted in their cluster within a single slot as an optimized (aggregate) forwarding operation. This aggregate forwarding operation in turn gives rise to a new scaling problem [DW15].

Further solutions for M2M overload based on access barring have been examined in a few studies. In [Phu+12; Tsa+12], access class barring variations have been investigated, while a full rejection of M2M traffic depending on the traffic load has been investigated in [KHT12]. Fast Adaptive Slotted ALOHA (FASA) with access barring was introduced in [Wu+13] and arrival estimation for access barring was investigated in [He+15]. For one-shot arrivals, Lien et al. [Lie+12] proposed a cooperative access barring scheme for multi cellular M2M communications. The multitude of picocells is exploited to offload some M2M communication between cells depending on the arrival density; thus essentially smoothing the M2M traffic load across a spatial region. Wali and Das [WD14] proposed heterogeneous networks as a solution to overload M2M traffic scenarios.

4.1.4.3 Tree Based Random Access for M2M Traffic

General stability analyses of tree-based collision resolution for bursty traffic models have been conducted in [VB04]. For simultaneous batch arrivals from an unknown number of devices, the batch resolution algorithms in [CS88; PFP05; Zan12a] estimate the number of devices and conduct random access for their requests.

Madueno et al. [MSP14b] proposed a collision resolution method using a q-ary tree splitting algorithm for synchronous M2M arrivals in LTE. The tree splitting algorithm splits the available preambles into groups. Preamble groups are assigned to the collided requests with a specific back-off timer. Although the tree splitting approach serves all the requests faster compared to previous approaches, it achieves a long run average throughput (of successful request per preamble per slot) of only 0.20. Energy harvesting aspects of tree based random access for M2M traffic have been examined in [VAA16].

4.1.4.4 Tree Resolution Algorithm

Tree resolution algorithm has been extensively studied for (non-synchronous) Poisson process arrivals. Capetanakis [Cap79a] showed that the stable throughput (of successful requests per slot) of the binary tree algorithm, that uses a dedicated binary tree to resolve the collisions, is 0.34. The optimum (dynamic) tree protocol [Cap79a, Section III-A.] with an optimal branch size for the first branch, followed by binary tree branching increases the throughput to 0.4295. Extensions of the optimum tree protocol generally exploited specific feedback information or interference cancellation techniques to achieve further throughput increases. For instance, the modified tree algorithm [Mas81] uses the data of an empty or a collided resource to estimate and resolve a wasted slot before it occurs, achieving a throughput 0.4622. The First Come First Served (FCFS) based contention resolution [Gal78] adapts the decision of joining a slot or not and the number of available slots with the resource use feedback, achieving a throughput of 0.4877. The successive interference canceling tree algorithm extracts data from collided slots to achieve a throughput of 0.693 [PVB07; YG07].

4.1.4.5 Hybrid Random Access

A hybrid random access scheme that shares resources between the random access channel and data transmission channel was investigated by Wiriaatmadja and Choi [WC15]. Wiriaatmadja and Choi and modeled human to human (H2H) traffic and M2M traffic in order to optimally allocate resource for the data channel and the random access channel. Liu et al. [Liu+14] and Verma et al. [Ver+16] combined conventional carrier sense multiple access for collision resolution with a reservation based time division multiple access. Hybrid random access

combining a contention-based and contention-free access has been examined for M2M in a home setting in [Yu+15].

Introducing the wide set of random access protocols we move on to state of the art on consideration of delay constraints for random access.

4.1.4.6 Delay constrained random access

A branch of work in delay constrained random access assumes that the *arrival distribution* is known such that the resolution can provide guarantees for that arrival setting. In this branch, the total number of devices is assumed to be known. However, the exact activation time of each device is not known. Thus, contention algorithms are optimized with the knowledge of the total number of devices. In [Ste+13], authors suggests that devices are polled to the access channel. After each passing time-slot the probability of access decreases where after some time there are only idle channels. They also suggest that probability is modified exponentially. The set of outcomes is fed to a maximum likelihood estimator to provide the total number of backlogged devices. Through the knowledge obtained from polling they allocate required number of resources for contention. However, in case polling is done periodically it can translate into added delay. Another work with known number of users is [SLP17] where authors have investigated resolution of certain number of users via successive interference cancellation capability within a certain limited amount of time. A recent work [Vil+18a] uses stochastic network calculus to provide stochastic bounds on the delay for dynamic access barring. Work [Jia+17a] derives delay distribution for the multichannel slotted ALOHA.

Previously, the request based contention is defined as random access and packet based contention is defined as grant free access. As the study also covers evaluation of 5G based access algorithms introduction to state of the art on delay constrained grant free access is necessary.

4.1.4.7 Delay constrained grant free access

Grant-free access from the system level perspective was investigated in [Jac+17], for an outdoor 3GPP urban scenario with multiple cells. The main contribution of the paper are simulation-based results that outline the setups in which the grant-free approaches outperform the grant-based ones, for the case of Poisson arrivals. Another work evaluates user activity with Bernoulli arrivals with different activation probabilities for grant-free scenarios in [Kot+18], proposing a new hybrid scheme that benefits from the advantages of both grant-based and grant-free schemes and focusing on the achievable data rates.

The capabilities of grant free access can be further increased through interference cancellation capabilities, as assumed to be provided by the 5G physical layer. Recent work have been focusing on this capability.

4.1.4.8 Delay constrained random access with interference cancellation:

A system level integration of grant free K-MPR in the 5G setup is evaluated via simulations in [SC17], comparing the effects of the channel estimation failure and contention failure on the access protocol design. In [Sin+18], K-MPR FSA is evaluated and the authors provide analytical expressions for collision probability under Poisson arrivals; we note that the extension for other arrival types is not trivial. Moreover, the resource-efficiency perspective is neglected, as the throughput is not evaluated, and there is no discussion on how increasing K affects the throughput. An extension of Irregular Repetition Slotted ALOHA (a slotted ALOHA-based scheme with successive interference cancellation) for the scenarios with multiple classes of Beta arrivals with different reliability-latency constraints is investigated in [Abb+17].

Finally, after we introduce the state of the art in admission control. The admission control before random access is totally novel up to our best knowledge. In the following, we discuss the admission control in wireless networks for scheduled access.

4.1.4.9 Admission control:

There are works that use admission control **after** random access for access grants. A work from Bell Labs [KLE95] proposes a protocol that is called Distributed Queuing Request Update Multiple Access (DQRUMA). A controller keeps track of a distributed queue. The *distributed queue* is the buffer status of multiple users. When a user has a packet, it can place a request on the random access. This request can collide and the collision is resolved with a tree resolution algorithm. This is called the *request* part of the algorithm. Through the state of the queues the controller decides whom to grant access. If a user is allocated a resource, then it can send a packet. At the end of a packet, a one bit header is added to notify that it has more packets. This approach is called piggy-backing and it *update*s the distributed queue. In general, it has a similar structure as the LTE system in terms of access grant logic. Two particular differences are that tree resolution is used and through piggy-backing the load on the random access channel is decreased. A similar adaptation for the current mobile networks is also proposed in [Lay+16]. Distributed Queuing is an adaptation of the tree resolution protocol for requests. Thus, long waiting times for long data packets are avoided through transmission of request packets. Up to the best of our knowledge stochastic delay constraint access has been neglected by the DQ protocols. Some of the most recent work on DQ have taken a load reactivity direction.

4.1.5 Notation

The sets are denoted with calligraphic capital letters $\mathcal{A}$. Sequences are denoted with bold lower-case letters $\mathbf{a}$. Sequences of sequences are denoted with bold upper-case letters $\mathbf{A}$. $\mathrm{E}[.]$ is used for expectation and e denotes the natural exponent. $\hat{(.)}$ is used for the estimated quantities. Probability mass functions are given with $p[.]$ while probability density functions are denoted as $p(.)$.

4.2 The effect of user activity on average delay constraints in cellular networks and hybrid solutions

This section evaluates a class of state of the art radio resource managements protocols for synchronous user activity called as M2M traffic. We examine collision avoidance techniques, such as pre-smoothing and access barring, with tree collision resolution to form the class of hybrid collision avoidance-tree collision resolution protocols.

As we have taken the safety-critical intra-aircraft communications for our use-case, we apply the state of the art solutions for intra-aircraft requirements, as defined in Sec. 3, to compare them. The worst case for such a machine to machine (M2M) scenario will be an alarm or the re-synchronization after a power shortage, since all of the M2M devices at that moment will try to reach the network simultaneously this will make a reliable communication almost impossible with the available LTE system, which is a candidate for the in-aircraft communications with the possibility of LTE-U chip sets [LTE15]. With the advances in wireless sensor technology and their widespread deployment on aircraft it will be important to develop effective communication protocols for synchronous M2M communication in Wireless Aircraft Intra Communications (WAIC) settings.

4.2.1 Scheduled access for M2M

A basic approach to multiple access is to use static time division multiple access scheduling, in which the devices are pre-assigned slots and thus their mutual interference is avoided. However, this approach makes sense if the activity patterns of the users are a-priori known (or can be precisely forecast) or if n_{a} is close to n_{tot}, when the base station schedules the resources to all n_{tot} devices in the cell. Otherwise, if $n_{\mathrm{a}} \ll n_{\mathrm{tot}}$ and the user activation is random, static scheduling results in a very low efficiency as can be seen in Fig. 4.4. Indeed, for the static scheduling, the number of resources needed per user, namely the throughput is simply

$$T = \frac{\mathrm{E}[N_{\mathrm{a}}]}{n_{\mathrm{tot}}} \tag{4.7}$$

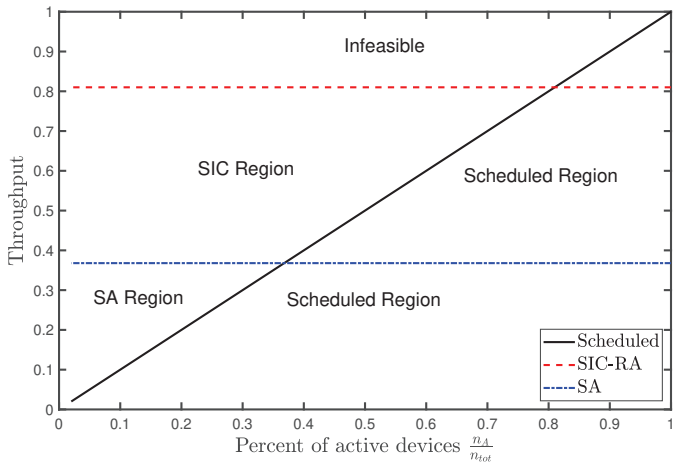


Figure 4.4: Resource efficiency (x-axis) is plotted against the device activity (y-axis). Scheduled access and random access schemes are set to separated into regions on the figure as efficient solutions in different settings. Right hand side as separated by the scheduled line defines the scheduled region. Top left hand side is currently infeasible as no algorithm demonstrated capabilities to achieve such throughput for this region. Below the successive interference cancellation (SIC) region is able to push the throughput up to levels of 0.8 from the limit of 0.367 of simple Slotted ALOHA.

as all active users are successful without any contention, i.e., $N_{\mathrm{a}} = N_{\mathrm{r}}$, and the reliability of this approach is $r(g, \mathrm{L}, n_{\mathrm{a}}) = 1$.

An alternative to scheduled scheme is to use dynamic scheduling of the slots to the active devices. In this approach, the BS first learns both n_{a} and active user identities, such that it can instruct them which slots to use. This is achieved using a separate resource reservation phase, which is standard request based access. In the resource reservation phase, the active users contend for resources with special packets containing their identifiers (i.e., metadata) in order to inform the BS of their activity and receive notification of the slot assignment for the subsequent transmission of the packets containing data.

We first summarize how the collision avoidance techniques that are introduced in the background synergies with resolution techniques.

4.2.2 Collision Avoidance

In this section we examine collision avoidance mechanisms for M2M traffic. We consider PreBO smoothing and conduct a probabilistic analysis of the PreBO dynamics. We then consider DAB as an alternative collision avoidance mechanism.

Arrival Smoothing as Enabler for Optimal Dynamic Tree Resolution Algorithm Generally, the high-performance dynamic tree resolution algorithm mechanism [Cap79a] requires accurate estimates of the number of requests colliding on a preamble in order to initiate the tree

resolution algorithm with the optimal number of tree branches. However, in the LTE system, the number of the collided users on a preamble in a given slot is not available. The direct estimation of the Beta distributed M2M request arrivals is also very challenging [Lan+13b]. Moreover, in the LTE RACH, the request arrivals are distributed over M preambles; thus, the estimation of the collided requests on a given preamble is prone to large estimation errors.

The preBO strategy can bypasses this challenging estimation of the arrivals. The pre-back-off uniformly distributes the requests over Γ back-off time slots with M preambles available in each slot. This uniform distribution shapes the request arrival process to follow approximately a Poisson distribution, as illustrated in Section 4.2.2.2.

Capetanakis [Cap79a, Section III-A.] showed that the optimal number of branches for the initial branching in a tree algorithm should be the first integer greater than the constant μ divided by 1.15. μ is the mean of the Poisson distributed number of request. Instead of configuring the number of tree branches, we configure the number of pre-back-off slots Γ as a novelty. Importantly, our number of pre-backoff slots is analogous to the number of initial tree branches in [Cap79a]. Based on this key insight, we configure the number of pre-back-off slots Γ as follows. We first note that with a total of n_{tot} UE requests and M preambles per slot, there are on average n_{tot}/M UE requests to be processed on a given preamble (across all pre-back-off slots). Noting that n_{tot}/M is a moderately large number in typical M2M scenarios, the optimal number of pre-back-off slots is

$$\Gamma = \frac{n_{\text{tot}}}{1.15M}, \tag{4.8}$$

which we round up to obtain an integer number of pre-back-off slots.

The second key insight about the arrival smoothing of pre-back-off is that the uniform random distribution of the n_{tot}/M UE requests on a preamble into Γ back-off slots corresponds to executing the initial tree branching of the optimal tree protocol [Cap79a, Section III-A.]. For the subsequent tree branching, two branches, i.e., the binary tree is optimal [Cap79a, Section III-A.].

4.2.2.1 User activity smoothing

The correspondence between the user activity smoothing with PreBO and the optimal tree protocol [Cap79a, Section III-A.] is introduced in section. M2M requests may arrive over a time period of multiple T_A slots, see Section 4.1.2. We set the number of PreBO slots

$$\Gamma = \left\lceil \frac{n_{\text{tot}}}{1.15M} \right\rceil \tag{4.9}$$

equal to the number of optimal branches. Specifically, the $n_{\text{a}},\ n_{\text{a}} \leq n_{\text{tot}}$, UE requests arriving in a given slot t are uniformly randomly distributed over the next Γ slots. That is, each UE

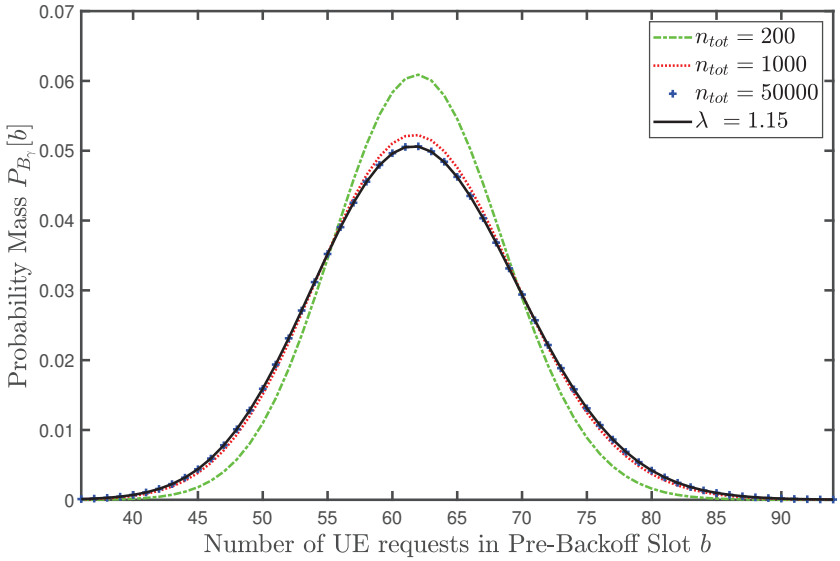


Figure 4.5: Probability mass function for number B_γ of UE requests in a given pre-backoff (PreBO) slot γ out of a total of $\Gamma = \lceil n_{\text{tot}}/(1.15M) \rceil$ PreBO slots for $M = 54$ preambles and different numbers of UE requests n_{tot}.

node independently randomly selects a given back-off slot $t+\gamma,\ \gamma = 1, 2, \ldots, \Gamma$, with uniform probability $1/\Gamma$.

4.2.2.2 User activity smoothing analysis

For the analysis in this section, we consider n_{tot} arrivals in a single slot, i.e., the Delta M2M traffic model. The motivation is to provide a best-case analysis. Later on this analysis is compared with simulations for both Delta and Beta arrivals in Section 4.2.5. Independently randomly distributing n_{tot} UEs into Γ back-off slots results in the probability mass function (pmf)

$$P_{B_\gamma}[b] = \binom{n_{\text{tot}}}{b} \left(\frac{1}{\Gamma}\right)^b \left(1 - \frac{1}{\Gamma}\right)^{n_{\text{tot}}-b}, \quad b = 0, 1, \ldots, n_{\text{tot}}, \tag{4.10}$$

for the number B_γ of UEs in a given backoff slot $\gamma,\ \gamma = 1, 2, \ldots, \Gamma$. Fig. 4.5 illustrates the pmf $P_{B_\gamma}[b]$ for the a group of sensors that can be modeled by the beta distribution. We observe from Fig. 4.5 that for the typical large number of requests n_{tot} (on the order of thousands) relative to the number of preambles M, the pmf $P_{B_\gamma}[b]$ closely approaches the Poisson distribution with mean $1.15M$. Moreover, we observe from Fig. 4.5 that the support of the pmf $P_{B_\gamma}[b]$ can reasonably be approximated by the Poisson distribution to the range $1.15M/2, \ldots, 3 \cdot 1.15M/2 = 27, \ldots, 81$ (for the considered typical number of preambles $M = 54$).

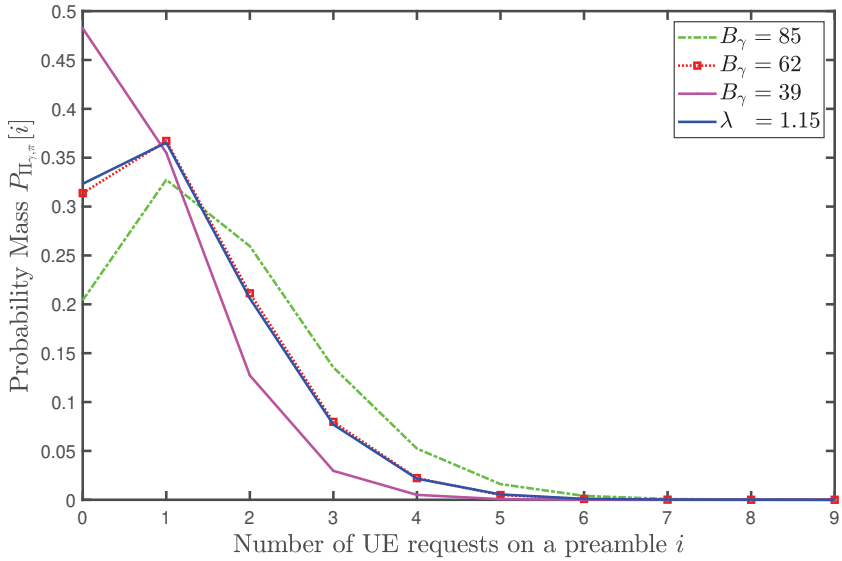


Figure 4.6: Probability mass function (pmf) for number $\Pi_{\gamma,\pi}$ of UE requests on a given preamble π within a given PreBO slot γ for different given numbers B_γ of UE requests in a given backoff slot γ. The pmf of $\Pi_{\gamma,\pi}$ approximates the pmf of a Poisson random variable with mean 1.15.

Through conditioning on the number B_γ of UEs in a given back-off slot γ, we obtain the pmf for the number $\Pi_{\gamma,\pi}$ of UEs on a given preamble $\pi,\ \pi = 1, 2, \ldots, M$, within a given pre-back-off slot $\gamma,\ \gamma = 1, 2, \ldots, \Gamma$, as

$$P_{\Pi_{\gamma,\pi}}[i] = \sum_{b=i}^{n_{\text{tot}}} P_{B_\gamma}[b] \binom{b}{i} \left(\frac{1}{M}\right)^i \left(1 - \frac{1}{M}\right)^{b-i}, \quad i = 0, 1, \ldots, n_{\text{tot}}. \tag{4.11}$$

Alternatively, pre-back-off can directly distribute the n_{tot} UE requests uniformly randomly over the ΓM preambles covered by the Γ back-off slots. Then,

$$P_{\Pi_{\gamma,\pi}}[i] = \binom{n_{\text{tot}}}{i} \left(\frac{1}{\Gamma M}\right)^i \left(1 - \frac{1}{\Gamma M}\right)^{n_{\text{tot}}-i}, \ i = 0, 1, \ldots, n_{\text{tot}}. \tag{4.12}$$

Inserting Γ from Eqn. (4.8) gives

$$P_{\Pi_{\gamma,\pi}}[i] = \binom{n_{\text{tot}}}{i} \left(\frac{1.15}{n_{\text{tot}}}\right)^i \left(1 - \frac{1.15}{n_{\text{tot}}}\right)^{n_{\text{tot}}-i}, \ i = 0, 1, \ldots, n_{\text{tot}}. \tag{4.13}$$

In order to avoid notation clutter, we abbreviate $P_{\Pi_{\gamma,\pi}}[i]$ to P_Π. Notice that the mean of $\Pi_{\gamma,\pi}$ is 1.15. Also, notice that for large M2M systems, i.e., for $n_{\text{tot}} \to \infty$, the binomial distribution in Eq. (4.13) converges by the Poisson Limit Theorem [Bil12] to the Poisson distribution with parameter 1.15, i.e., to $P_\Pi[i] = e^{-1.15} 1.15^i / i!$.

Fig. 4.6 illustrates the pmf P_Π for the typical range of numbers B_γ of requests in a given backoff slot ranging from $B_\gamma = 27$ to $B_\gamma = 81$. We observe from Fig. 4.6 that with probability

of roughly one third there is only one request on a given preamble, resulting in successful random access. We also observe that if there are multiple requests on a given preamble, i.e., a collision occurred, then the number of collided requests is typically small, in the range from two to five. A higher number (of six or more) collided requests occurs only rarely.

4.2.3 Tree based Random Access (TRA) for M2M Traffic

In contrast to the prior studies on tree resolution algorithm, we combine the collision avoidance mechanism from Section 4.2.2 with tree resolution algorithm to efficiently serve bursty M2M traffic.

More specifically, prior studies on dynamic tree resolution algorithm access employed estimation methods with Poisson traffic. The bursty M2M traffic would require novel traffic estimation methods for the dynamic tree resolution algorithm. The effect of user activity also motivates the subsequent sections in this chapter. The challenging estimation problem is bypassed by executing a collision avoidance mechanism prior to feeding the UE requests into the TRA. As we demonstrate in Section 4.2.2.2, the pre-back-off smooths the bursty arrivals to approximate a Poisson distribution. Thus, after the pre-back-off, any of the tree algorithms can be employed to achieve high performance reliable random access in LTE. However, signaling and the decision timings have to be adapted to fit the LTE system.

4.2.3.1 Specification of the TRA Protocol

After the collision avoidance mechanism, the tree-based random access (TRA) protocol sequentially processes the UE requests on a given preamble. In particular, in case of the PreBO collision avoidance, the $\Pi_{\gamma,\pi}$ UE requests on preamble π in a given pre-back-off slot γ are processed as follows. If $\Pi_{\gamma,\pi} = 1$, then the single UE request is successful. If $\Pi_{\gamma,\pi} \geq 2$, then we sequentially launch a binary (two branches) tree resolution algorithm (collision resolution) for these UE requests. Specifically, these UE requests are directed to a preamble group consisting of two preambles in an upcoming slot. Note that with M preambles there are at most $M/2$ preamble groups for binary tree collision resolution in a slot. The process of directing UE requests to a preamble group is repeated until all of the UE requests have been served.

4.2.4 Hybrid Collision Avoidance-Tree Resolution Protocols

4.2.4.1 Pre-Backoff Tree-based Random Access (PreBOTRA)

Protocol Specification The PreBOTRA protocol is a concatenation of the pre-back-off (PreBO) specified in subsection 4.2.2.1 and the tree resolution algorithm (TRA) specified

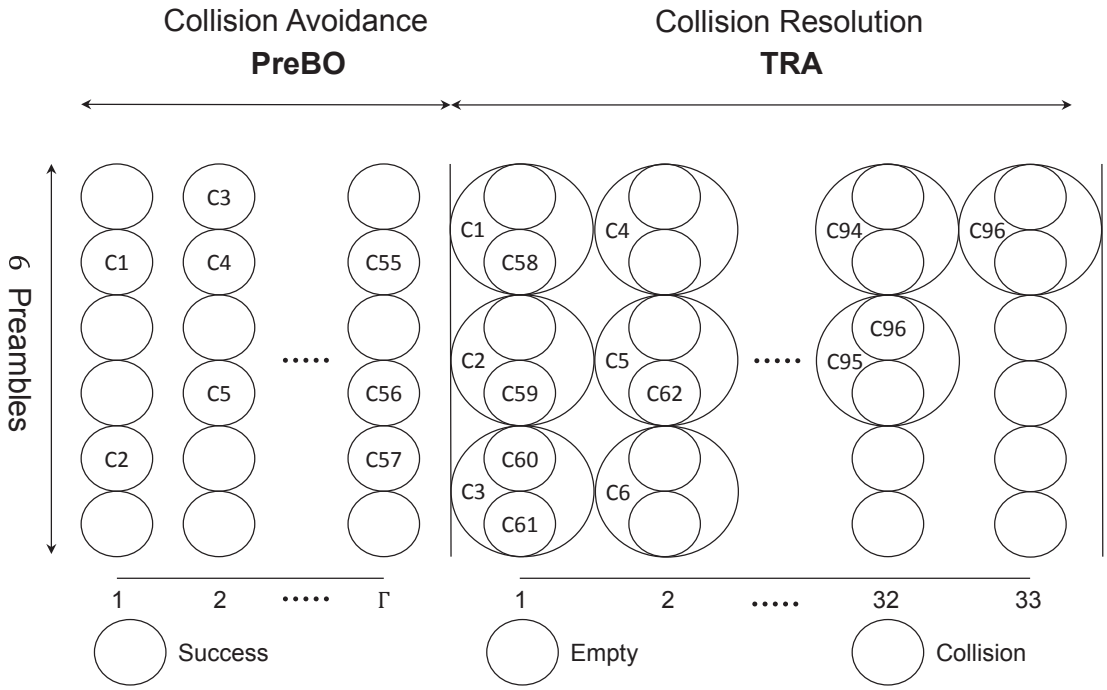


Figure 4.7: Example illustration of PreBOTRA operation for $M = 6$ preambles: 57 preambles out of the 6Γ preambles during the PreBO collision avoidance phase have a collision, i.e., were randomly selected by a set of $\Pi_{\gamma,\pi} \geq 2$ UE requests. Collisions are resolved through binary TRA, i.e., each set of collided UE requests is directed to a group of two preambles.

in Section 4.2.3.1. The operation of the resulting PreBOTRA protocol is illustrated for an example with $M = 6$ preambles in Fig. 4.7. In the illustrated example, 57 preambles (out of the total of 6Γ preambles available for PreBO) experience a collision, i.e., contain two or more UE requests. Starting with slot $\Gamma + 1$, the binary TRA is launched for each of the 57 collisions. For the binary TRA with the $M = 6$ preambles, there are $M/2 = 3$ preamble groups in each slot during the TRA collision resolution. In the illustrated example, collision 1 from the first PreBO slot results in one success and one collision (C58) in the first TRA slot. Such new collisions are then recursively resolved with TRA. In the illustrated example, a total of $96 - 57 + 1 = 40$ new collisions occur during the TRA.

Generally, when synchronous M2M arrivals are detected (see Section 4.1.2.1), PreBOTRA initiates the PreBO with Γ (see Eq. (4.9)) slots. UE requests arriving for any slot during the activation time period T_A, $T_A \geq 1$, are distributed by the PreBO over slots T_A, $T_A + 1, \ldots, T_A + \Gamma - 1$, as illustrated for Delta arrivals with $T_A = 1$ in Fig. 4.7. For simplicity of the protocol, the TRA collision resolution always starts Γ slots after the beginning of the synchronous M2M arrivals, as illustrated in Fig. 4.7, even when $T_A > 1$. Thus, for $T_A > 1$,

the PreBO phase of late arriving UE request overlaps with the TRA phase of early arriving UE requests. In this case, the newly (late) arriving UE requests, distributed by the PreBO, simply randomly content for individual preambles; while these preambles are also utilized in groups of two for the binary TRA.

4.2.4.2 Throughput and Delay Analysis

We analyze the throughput of PreBOTRA by evaluating the number of slots required to successfully complete all n_{tot} UE requests. PreBOTRA distributes the n_{tot} requests over Γ back-off slots, each with M preambles. Thus, the number $\Pi_{\gamma,\pi}$ of UE requests on a given preamble π in a given back-off slot γ is characterized by the pmf P_{Π} (4.11), resp. (4.13). The binary tree-based collision resolution is executed for these $\Pi_{\gamma,\pi}$ UE requests.

The expected number $f_{\text{ST}}(n_{\text{a}})$ of slots required to complete the static binary tree resolution algorithm for n requests is given as [MP93a]:

$$f_{ST}(n_{\text{a}}) = \begin{cases} 1 & \text{if } n_{\text{a}} = 0 \\ 1 & \text{if } n_{\text{a}} = 1 \\ 5 & \text{if } n_{\text{a}} = 2 \\ 7.667 & \text{if } n_{\text{a}} = 3 \\ . & \\ . & \\ 2.88n & \text{if } n_{\text{a}} = n_{\text{a}}. \end{cases} \tag{4.14}$$

The expected number $f_{\text{ST}}(n_{\text{a}})$ of slots covers the initial transmission attempt of a given set of $\Pi_{\gamma,\pi}$ UE requests on preamble π in back-off slot γ, plus the slots required for the resolution if a collision occurs, i.e for $n_{\text{a}} > 2$. Thus, the expected number of slots required to serve the set of $\Pi_{\gamma,\pi}$ UE requests on a given preamble π in a given pre-back-off slot γ is $\sum_{n_{\text{a}}=0}^{n_{\text{tot}}} P_{\Pi}[n_{\text{a}}] f_{ST}(n_{\text{a}})$. Noting that a total of ΓM sets of UE requests (on the ΓM preambles across the Γ PreBO slots) need to be served, a mean number of $\Gamma M \sum_{n_{\text{a}}=0}^{n_{\text{tot}}} P_{\Pi}[n_{\text{a}}] f_{ST}(n_{\text{a}})$ slots are required to serve all sets of UE requests on one preamble. Each slot provides M preambles. Thus, a mean number of

$$S_{\text{PreBOTRA}} = \Gamma \sum_{n_{\text{a}}=0}^{n_{\text{tot}}} P_{\Pi}[n_{\text{a}}] f_{ST}(n_{\text{a}}) \tag{4.15}$$

slots are required for the last set of UE requests to complete PreBOTRA. On average, a given UE request will experience slightly more than the half of this time duration as PreBOTRA delay, as a request is more likely to be resolved towards the end of the algorithm. But the

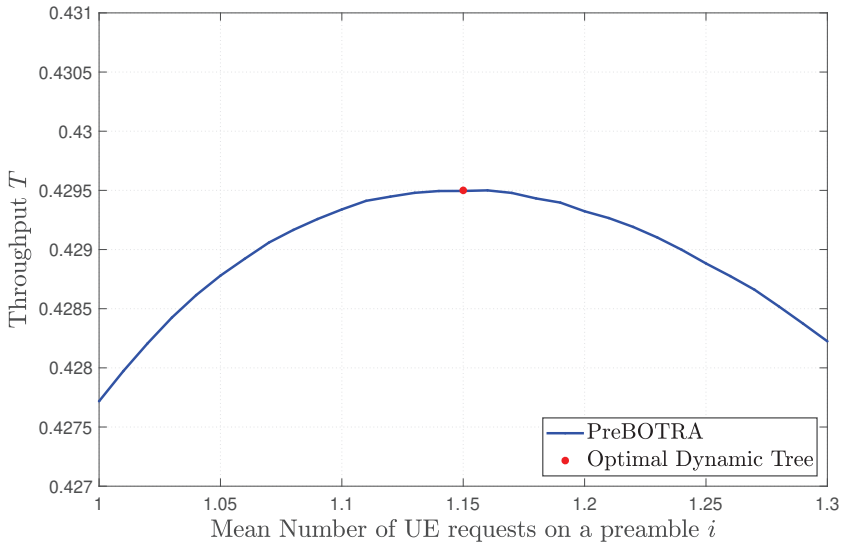


Figure 4.8: Expected PreBOTRA throughput Eq.(4.17) as a function of the mean of $\Pi_{\gamma,\pi}$, i.e., of the mean number of UE requests on a preamble in a PreBO slot. Fixed parameters: $M = 54$ preambles, $n_{\text{tot}} = 10,000$ nodes. Setting the number Γ of PreBO slots according to Eq.(4.8) to achieve a mean number of 1.15 UE requests per preamble in a PreBO slot achieves the maximum expected throughput of 0.4295 (successful UE requests per preamble per slot) of the optimal dynamic tree [Cap79a].

success probabilities do not vary drastically. Thus, a lower bound to the average UE request delay in slots is

$$D \geq \frac{\Gamma}{2} \sum_{n_{\text{a}}=0}^{n_{\text{tot}}} P_{\Pi}[n_{\text{a}}] f_{ST}(n_{\text{a}}). \tag{4.16}$$

In order to derive the expected PreBOTRA throughput, we neglect that UE request are dropped when exceeding the number of transmission attempts permitted by LTE. As evaluated in Section 4.2.5, this is a reasonable assumption as the drop probabilities are typically very low. Thus, we consider that n_{tot} UE requests are successfully served during the total PreBOTRA time span of S_{PreBOTRA} slots. Each slot has M preambles available. Hence, the expected throughput in successful UE requests per preamble per slot is

$$T = \frac{n_{\text{tot}}}{\Gamma M \sum_{n_{\text{a}}=0}^{n_{\text{tot}}} P_{\Pi}[n_{\text{a}}] f_{ST}(n_{\text{a}})}. \tag{4.17}$$

which simplifies with Eq. (4.8) to

$$T = \frac{1.15}{\sum_{n_{\text{a}}=0}^{n_{\text{tot}}} P_{\Pi}[n_{\text{a}}] f_{ST}(n_{\text{a}})}. \tag{4.18}$$

In Fig. 4.8, we plot the expected throughput as a function of the mean number of UEs per preamble, i.e., we vary the mean of $\Pi_{\gamma,\pi}$ around 1.15. The plot verifies that setting the

Table 4.3: Summary of evaluation benchmarks with theoretically expected maximum throughput and added complexity with respect to standard LTE.

Protocol	Description	Exp. Tpt.	Added Complexity
Benchmarks			
LTE+PreBO	LTE with PreBO [Har+15]	0.367	Synch. arrival detection
LTE+DAB	LTE with DAB [DSW13; WW15a]	0.367	AB update in each slot
LTE+TRA	Tree for each collision [MSP14b]	0.2	Collision based feedback
Hybrid Collision Avoidance-Tree Resolution			
PreBOTRA	PreBO /w tree, Sec. 4.2.4.1	0.4295	LTE /w PreBO + Tree
DABTRA	DAB /w tree, Sec. 4.1.3.5	0.4295	LTE /w DAB + Tree

number of PreBO slots Γ according to Eq. (4.8) to achieve a mean of $\Pi_{\gamma,\pi}$ equal to 1.15 achieves the maximum expected throughput of 0.4295 successful UE request per slot. The plot also illustrates that small deviations of the mean of $\Pi_{\gamma,\pi}$ from 1.15 cause only minuscule throughput degradations.

4.2.4.3 DABTRA

Analogously to PreBOTRA, the overall DABTRA protocol is a concatenation of the DAB collision avoidance specified in [Dua+16] (and briefly summarized in Section 4.1.3.5) and the TRA collision resolution specified in Section 4.2.3.1. For protocol simplicity, similar to PreBOTRA, we let the TRA always start Γ slots after the beginning of the synchronous M2M arrivals. Alternative DABTRA variations could vary the start of the TRA phase, e.g., the TRA could start right away as soon as UE requests that have passed the DAB experience a collision on a preamble. As the name suggests DABTRA provides a dynamic adaptation to the arrivals. In that sense, it is sensitive to instantaneous changes in user activity while it is self adapting to smooth changes in user activity.

4.2.5 Performance Evaluation

In this section we compare the performance of two forms of the proposed hybrid collision avoidance-tree resolution algorithm, namely PreBOTRA and DABTRA, against existing benchmarks listed in Tab. 4.3 that employ either collision avoidance or tree resolution.

4.2.5.1 Evaluation Setup

We conducted the evaluations through discrete event simulations of a standard LTE RACH system in MATLAB with a maximum of eight transmission attempts for a given UE request and $M = 54$ preambles. For each combination of M2M traffic model and number of UE

requests n_{tot}, we simulated 100 independent replications, resulting in 95 % confidence intervals that are smaller than 5 % of the respective sample means. The confidence intervals are not plotted to avoid clutter.

4.2.5.2 Benchmarks

We have compared three benchmark protocols, summarized in Tab. 4.3, with the PreBOTRA and DABTRA protocols. Generally, all protocols operate with the underlying standard LTE random access procedure with at most eight transmission attempts. The protocols without DAB do not employ access barring. The protocols without TRA back off uniformly over at most Γ slots after a collision. The LTE with PreBO benchmark invokes the pre-back-off when synchronous arrivals are detected. This LTE with PreBO benchmark represents the pre-back-off collision avoidance focused approaches, as for instance examined in [Har+15]. The LTE with DAB benchmark employs dynamic access barring for collision avoidance, as examined in [DSW13; Dua+16; WW15a]. The LTE with tree benchmark resolves collisions with a binary tree. This LTE with tree benchmark represents the tree resolution focused approaches, as for instance examined in [MSP14b].

4.2.5.3 Performance Metrics

We define the mean throughput as the number of UE requests n_{tot} divided by the mean time period in slots from the starting instant of the synchronous M2M arrivals until all n_{tot} requests are either successful or dropped. We define the mean delay as the mean time period in slots required to serve a UE request from the time instant of request generation (activation) until the request is either successfully transmitted or dropped. We define the UE request drop probability as the ratio of the number of dropped UE requests to the total number of n_{tot} UE requests.

4.2.5.4 Evaluation Results

In Figs. 4.9, 4.10, and 4.11, we plot the performance metrics obtained from simulations as a function of the number of UE requests n_{tot}. We also plot the expected throughput (see Eq. (4.18)) and delay (see Eq. (4.16)) from the PreBOTRA analysis abbreviated as (ana.) in the Figure. We observe from Fig. 4.9 that for Delta arrivals, PreBOTRA essentially attains the theoretically expected maximum mean throughput of 0.4295. On the other hand, for Beta arrivals, PreBOTRA gives throughput below 0.4 for relatively small numbers of users, while the throughput approaches the theoretical maximum for large numbers of users. As described in Section 4.2.4.1, for simplicity, PreBOTRA always employs the PreBO interval Γ specified in Eqn. (4.9), which assumes that all n_{tot} UE requests are distributed over the Γ PreBO slots.

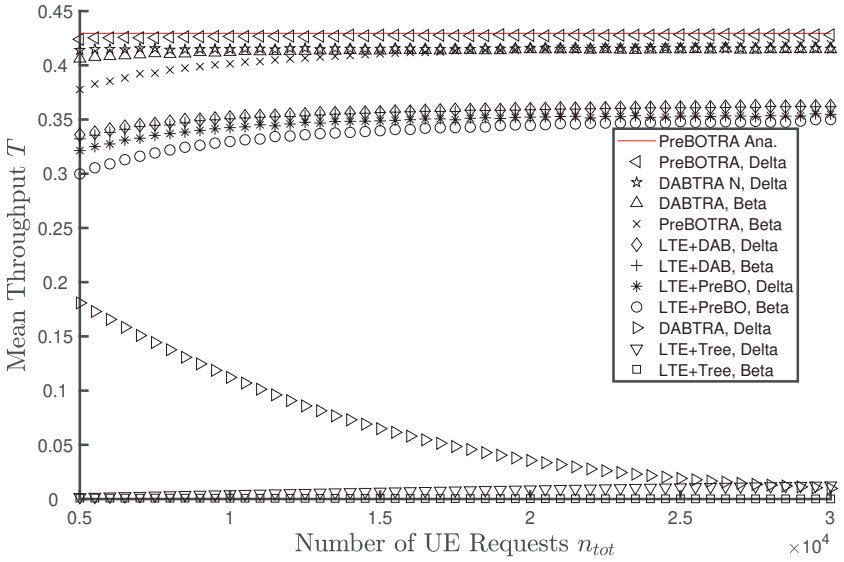


Figure 4.9: Mean throughput, i.e., mean number of successful UE requests per preamble per slot, as a function of number of UE requests n_{tot} arriving either in $T_A = 1$ slot (Delta arrival model) or over $T_A = 50$ slots (Beta arrival model). Fixed parameter: $M = 54$ preambles.

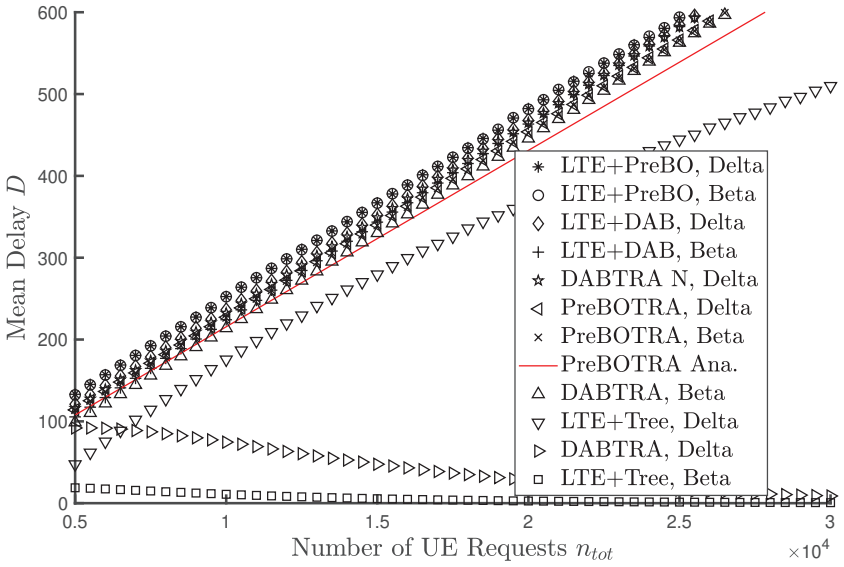


Figure 4.10: Mean delay in slots as a function of number of UE requests n_{tot}.

However, for an activation time period $T_A > 1$, the n_{tot} UE requests are distributed by the PreBO over more than Γ slots (see Section 4.2.4.1). Thus, the mean number of UE requests on a preamble is lower than the optimal 1.15, leading to the throughput degradation outlined in Subsection 4.2.4.2 and illustrated in Fig. 4.8. That is, we are operating to the left of the optimal dynamic tree point in Fig. 4.8. Increasing the number of UE requests n_{tot} for fixed

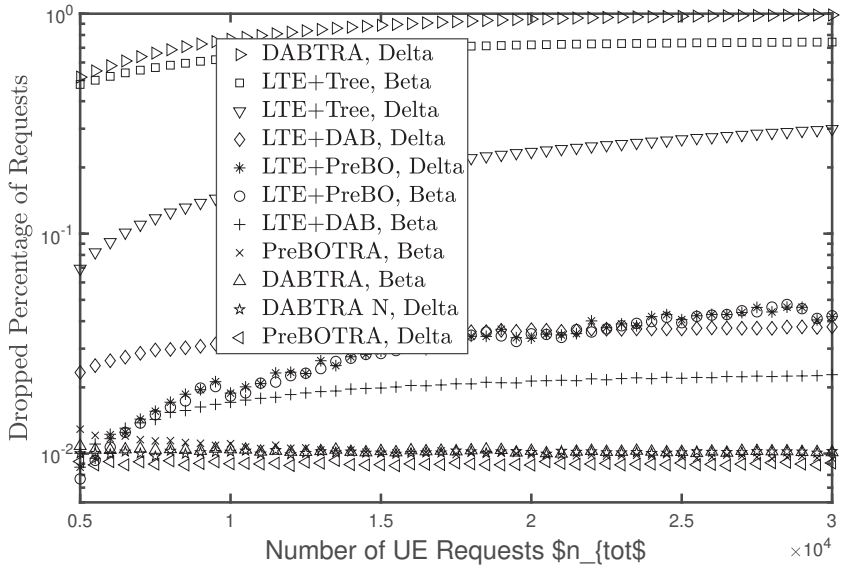


Figure 4.11: UE request drop probability as a function of number of UE requests n_{tot}.

activation time period T_A moves the operating point to the right, i.e., we are approaching the optimal dynamic tree point in Fig. 4.8 from the left.

Turning to DABTRA, we observe from Fig. 4.9 low throughput for Delta arrivals. By adapting the access barring based on the observed history the original DAB version reacts slowly to the sudden Delta arrivals, leading to excessive congestion on the preambles, and accordingly high drop probabilities (see Fig. 4.11) and accordingly low throughput. A modified DABTRA version that is provided with information about the n_{tot} Delta arrivals, denoted as "DABTRA n" in Fig. 4.9, achieves nearly the same throughput as PreBOTRA for Delta arrivals. The collision avoidance provided by DABTRA n effectively reduces the preamble congestion. However, the Delta arrivals are not optimally shaped (with the optimum Γ, see Section 4.2.4.1) for TRA, leading to slightly lower throughput than PreBOTRA. For Beta arrivals, the original DABTRA gives slightly higher throughput than PreBOTRA for low user numbers n_{tot}, while both PreBOTRA and DABTRA achieve essentially the same throughput for high user numbers n_{tot}. The more gradual Beta arrivals (compared to the sudden Delta arrivals) give DAB sufficient time to dynamically adapt the access barring, providing effective collision resolution. In practical systems, perfectly synchronized arrivals are typically due to some system reset, for which the modified DAB can be activated, or the arrivals are spread out over several slots (which have on the order of 10 ms duration). Thus, the performance of the modified DABTRA for Delta arrivals and the original DABTRA for Beta arrivals can be considered to reflect the DABTRA performance possible in practical systems.

We observe from Fig. 4.9 that the benchmarks employing only collision avoidance, followed by the conventional LTE collision resolution with uniform backoff, i.e., LTE+PreBO and LTE+DAB, achieve very similar throughput that approaches the theoretical maximum of $1/e$ for increasing number of UE requests n_{tot}. In contrast, we observe from Fig. 4.9 that the benchmark without collision avoidance, i.e., LTE+Tree gives very low throughput below 0.02. The results underscore the importance of collision avoidance for synchronous M2M arrivals. Without collision avoidance, the synchronous M2M arrivals cause excessive congestion on the preambles. The tree resolution cannot resolve the resulting excessive collisions within the LTE re-transmission limit (of eight in the considered scenario), leading to high drop probabilities (see Fig. 4.11). On the other hand, the LTE+PreBO and LTE+DAB results indicate that after collision avoidance, the standard LTE collision resolution can achieve moderate throughput levels. Adding the tree resolution after the collision avoidance to form the hybrid collision avoidance-tree collision resolution protocols significantly increases the throughput levels (compared to the benchmarks employing only collision avoidance with standard LTE collision resolution), as indicated by the PreBOTRA and DABTRA results.

Importantly, we observe from Fig. 4.10 that the increased throughput with the hybrid collision avoidance-tree collision resolution protocols (compared to collision avoidance with standard LTE collision resolution) does *not* come at the expense of increased delays. Rather, we observe that the mean delays of PreBOTRA and DABTRA are very slightly lower than the delays of the collision avoidance benchmarks LTE+PreBO and LTE+DAB. This is because the tree resolution optimizes the utilization of the preambles in the collision resolution.

We observe from Fig 4.11 that the hybrid collision avoidance-tree collision resolution protocols achieve consistently low drop probabilities around 10^{-2} in the considered scenario with an LTE limit of at most eight transmission attempts. In order to assess the effects of increasing the number of transmission attempts on PreBOTRA we have run simulations with $n_{\text{tot}} = 10000$ Delta arrivals and for increasing number of transmission attempts. In particular, for 8, 12, 16, and 20 transmission attempts, the mean drop probabilities are 0.009, $5 \cdot 10^{-4}$, $5 \cdot 10^{-5}$, and $2 \cdot 10^{-6}$, respectively, while the mean delays are 227, 230, 231, and 231 slots. These results indicate that the required level of reliability can be set via the maximum number of permitted transmission attempts. The results also indicate that this adjustment of the reliability level has negligible impact on the mean delay.

4.2.6 Summary

In this section the class of hybrid collision avoidance-tree resolution protocols for the medium access control (MAC) of machine-to-machine (M2M) traffic is proposed. Specifically, the PreBOTRA protocol combines pre-backoff with tree collision resolution, while the DABTRA

protocol combines dynamic access barring with tree collision resolution. These protocols are based on the knowledge of the total number of communicating machines (devices), which is commonly available for wireless communication systems on board of an aircraft.

The knowledge for the number of devices is available but the unpredictable activation times representing as beta arrivals affects the throughput and reliability (the drop rate in this study). This emphasizes the impact of the burst arrival effect on the radio resource efficiency and capability.

Another assumption of this study is the detection of the start of the burst arrival used to switch from conventional access mode to burst access mode. Even though application layer estimation techniques are assumed to be available to trigger this switch, incorporation of a cross layer solution can be complex. Instead a MAC layer activity estimation can be investigated to guarantee radio resource efficiency and burst arrival detection.

In the following section we investigate the question of whether we can deal with the over-dimensioning of resources through estimating the instantaneous activity of the user. If the activities of the users are known, radio resources can be allocated dynamically i.e. more efficiently.

4.3 Evaluation of throughput with user activity estimation in cellular networks

As the resource efficiency is the main goal of radio resource management it is used as the main utility and it is maximized after the delay-reliability constraints are fulfilled. In the previous section we demonstrated that most state of the art techniques sacrifice the resource efficiency to fulfill these constraints. The reason for the lack of resource efficiency is the unexpected activity of users. Thus, the radio resource manager has to be conservative and allocate resources as if the worst-case scenario would happen.

As the root cause of the over-dimensioning of the resources is the unexpected activity of the users, we can estimate the user activity to overcome this problem. In this section we investigate this new perspective.

One important remark here is we investigate whether the estimation would improve the resource efficiency. We actually do not estimate the arrivals in the scope of this section, we assume a precise estimation is available. The implementation of an estimator and its effects are investigated in the following section.

The cellular networks, currently LTE, are characterized by reservation-based access, which involves a random-access based, signaling-intensive connection-establishment pro-

cedure [Tya+17]. This approach is highly inefficient when short packets are sporadically exchanged, which is characteristic for IoT use-cases [LAA14b; Mad+16]. Moreover, each stage of the connection-establishment has the potential to compromise reliability and increase latency [Eri+18; Pop+19]. Thus, 3GPP has decided to standardize a grant-free access method, alongside the existing resource-reservation, in which the users will contend with their data packets in random-access fashion [18]. In short, the random access channel problem has morphed into a grant free access problem in the scope of 5G as the payload is used for contention based access.

In this section, we investigate the throughput maximization of the grant-free access from the perspective of medium-access control considering the overhead of estimation. Specifically, we analyze how the knowledge of number of arrived (i.e., contending) users can be used to boost the throughput, providing the following contributions:

- We give a formal definition of reliability under predefined latency constraint for the batch arrival, developing it for the cases when the number of arrived users is exactly known, or given by a certain arrival distribution. We also formally define throughput for both cases, providing insight on the role of the knowledge of the number of arrived users.
- We derive throughput under the reliability-latency requirements for framed slotted ALOHA (FSA) with K-multipacket reception (MPR), i.e., we assume that the operation of physical layer can be represented with successful reception up to and including K packets that occur simultaneously in a slot.
- We instantiate the analysis for the cases of Poisson and Beta arrivals, which are standard models of IoT traffic, and evaluate the impact of K-MPR and the knowledge of the number of arrived users. We show that increasing K as well as the knowledge of the number of arrived users pay off in throughput.
- The last insights suggest that estimation of the number active users in grant-free access can be beneficial, as this information is typically not readily available. In this respect, we investigate the impact of the potential estimation errors on the throughput, showing that for high reliability-latency requirements the gains are still considerable, even with high error levels.

This section is organized as follows: Section 4.3.1 introduces the system model. Section 4.3.2 defines the analytical framework composed of reliability-latency requirement and throughput maximization, which is then applied to FSA with K-MPR in Section 4.3.3. Sec-

tion 4.3.4 evaluates the performance of FSA with K-MPR under Poisson and Beta arrivals and the effect of user activity estimation.

4.3.1 Model

As in the previous section we consider a single cell scenario with a single base station (BS) and a total of n_{tot} users. At any time-instance only n_a of the users are active and access the resources in a random fashion.

N_a is a random variable denoting the number of active users at a time instance. The distribution is modeled as either the Poisson and or the Beta distribution as introduced in Sec. 4.1.2. This is specified separately for each analysis. N_a is also the random variable that is estimated to improve the resource efficiency.

We assume that the available slots are grouped in K-*superslots*: a K-superslot is dimensioned such that if there are up to and including K simultaneous transmissions occurring in it, all of them are successfully received (and the corresponding users become resolved). Otherwise, if there are more than K transmissions occurring in a K-superslot, none of them can be successfully received. In other words, we assume that the physical layer operation can be represented by K-MPR. The signaling overhead required to obtain the channel estimation needed to enable K-MPR is evaluated in [Gür+18] and in this study we assume the channel state information is available. We also assume that a K-superslot contains K slots in order to achieve the K-MPR capability and note that the linear increase in the superslot size with K is a reasonable assumption, cf. [GSP18; MDA17]. Finally, we assume that the users are aware of the superslot boundaries. This type of synchronization could be achieved via means of a downlink control channel, which is the typical scenario in cellular systems. Fig. 4.12 shows an example of 6-superslot.

For $K = 1$, the above model reduces to the standard collision channel model. Moreover, although simplistic, this model of K-MPR can be used as an approximation for systems in which other sources of diversity are employed to achieve multipacket reception, like the use of spreading codes, or multiple antennas.

The access decision of users is regulated via a grant-free access algorithm, whose goal is to ensure a predefined level of reliability $\mathrm{R(L)}$ of user resolution under a predefined latency constraint L in time units, see Fig. 4.12. We denote this requirement as the *reliability-latency* requirement in further text. Note that in the proposed setup, the number of frequency channels g assigned to the access procedure is the degree of freedom that can be optimized such that the target reliability $\mathrm{R(L)}$ is achieved. This reflects a typical radio resource management problem.

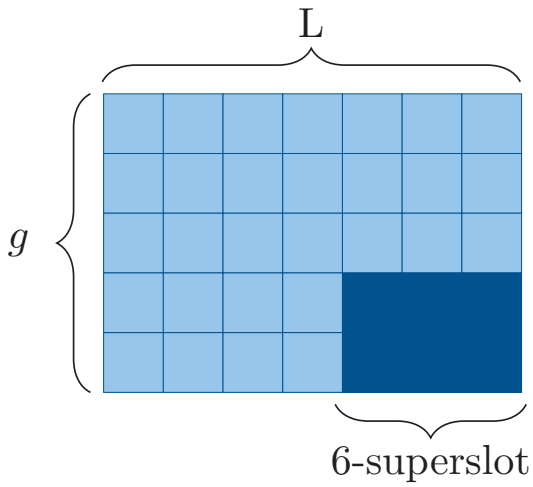


Figure 4.12: The resource grid comprising L time slots and g channels; L is given by the latency budget and g is optimized such that target performance is achieved. The figure also shows a 6-superslot defined over 2 channels and 3 time slots.

4.3.2 Performance Parameters

4.3.2.1 Reliability-latency

Formally, denote by ξ_u the event that an active user u becomes resolved and by L the maximum allowed latency of the resolution in time-units. In case of batch arrivals, the access algorithm should satisfy the following reliability-latency definition

$$r(g, \mathrm{L}, n_\mathrm{a}) = \Pr[\xi_u, L \leq \mathrm{L} | g, n_\mathrm{a}, n_\mathrm{tot}] \geq \mathrm{R(L)} \tag{4.19}$$

for all active users u in the batch, where it is assumed that the realization n_a of N_a is known. Assuming that the access algorithm does not output false positives in (i.e., the noise can not be decoded as a packet from an inactive user), the above condition can be expressed as

$$\begin{aligned} r(g, \mathrm{L}, n_\mathrm{a}) &= \sum_{k=1}^{n_\mathrm{a}} \frac{k}{n_\mathrm{a}} \Pr[N_\mathrm{r} = k, L \leq \mathrm{L} | g, n_\mathrm{a}, n_\mathrm{tot}] \\ &= \frac{\mathrm{E}[N_\mathrm{r}]}{n_\mathrm{a}} \geq \mathrm{R(L)} \end{aligned} \tag{4.20}$$

where N_r is the number of resolved users, and $\frac{k}{n_\mathrm{a}}$ is the probability that active user u is among the k resolved ones. In the assumed system model, $g = g(n_\mathrm{a})$ should be chosen such that the condition (4.20) becomes satisfied.

If the realization n_a is not known, the use of condition (4.20) is not possible. However, if the probability mass function (pmf) of N_a is known, the reliability-latency condition could be

defined as:

$$r^*(g^*, \mathrm{L}) = \sum_{n_\mathrm{a}} r(g^*, \mathrm{L}, n_\mathrm{a}) \Pr[N_\mathrm{a} = n_\mathrm{a}] \geq \mathrm{R}(\mathrm{L}) \tag{4.21}$$

where g^* should be chosen such that condition (4.21) is satisfied.

It is natural to assume that for any reasonable access algorithm, the following holds:

$$r(g, \mathrm{L}, n) \leq r(g+1, \mathrm{L}, n),\ \forall n \tag{4.22}$$

i.e., increasing the number of frequencies (which increases the total number of resources) will not lower chances to fulfill the reliability-latency condition. Under this assumption, it could be shown that there exists minimal values for $g(n_\mathrm{a})$ and g^*, for which (4.20) and (4.21), respectively, hold. Along the same lines, one can formulate optimization problems according to which these minimal values can be found, respectively

$$g_{\min}(n_\mathrm{a}) = \arg\min_g (r(g, \mathrm{L}, n_\mathrm{a}) : r(g, \mathrm{L}, n_\mathrm{a}) \geq \mathrm{R}(\mathrm{L})) \tag{4.23}$$

$$g^*_{\min} = \arg\min_g \left(r^*(g, \mathrm{L}) : r^*(g, \mathrm{L}) \geq \mathrm{R}(\mathrm{L})\right). \tag{4.24}$$

For the sake of brevity and with a slight abuse of notation, in the rest of the text we will assume that $g(n_\mathrm{a}) = g_{\min}(n_\mathrm{a})$ and $g^* = g^*_{\min}$, respectively, i.e. always the minimum number of channels is used.

4.3.2.2 Throughput

The number of resources dedicated to the resolution is $g\,\mathrm{L}$, see Fig. 4.12. We define the throughput as the expected number of resolved users vs. the number of resources

$$\mathrm{T} = \mathrm{E}\left[\frac{N_\mathrm{r}}{g\mathrm{L}}\right]. \tag{4.25}$$

In the proposed model, N_r is determined by the employed reliability-latency condition and the throughput is maximized by minimizing g. Specifically, when (4.20) is used, the throughput becomes

$$\mathrm{T} = \frac{\mathrm{R}(\mathrm{L})}{\mathrm{L}} \sum_{n_\mathrm{a}} \frac{n_\mathrm{a}}{g(n_\mathrm{a})} \Pr[N_\mathrm{a} = n_\mathrm{a}], \tag{4.26}$$

where we recall that $g(n_\mathrm{a})$ is chosen such that $r(g, \mathrm{L}, n_\mathrm{a}) = \mathrm{R}(\mathrm{L}),\ \forall\ n_\mathrm{a}$. In case when (4.21) is used, the throughput is

$$\mathrm{T}^* = \frac{\mathrm{R}(\mathrm{L})}{g^*\mathrm{L}} \sum_{n_\mathrm{a}} n_\mathrm{a} \Pr[N_\mathrm{a} = n_\mathrm{a}] = \frac{\mathrm{R}(\mathrm{L})\,\mathrm{E}[N_\mathrm{a}]}{g^*\mathrm{L}}. \tag{4.27}$$

where we also recall that g^* is chosen such that $r^*(g^*, \mathrm{L}) = \mathrm{R}(\mathrm{L})$. Obviously, there is a difference between T and T^*, which will be further investigated in Section 4.4.6.

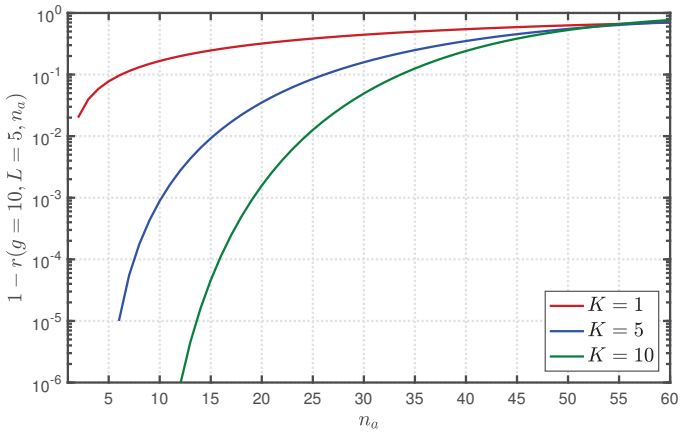


Figure 4.13: Reliability-latency performance of framed slotted ALOHA with K-MPR for varying number of users n_{a} and K, when $g = 10$ and $\mathrm{L} = 5$.

4.3.3 Grant-Free Access with FSA

In the considered scenario, the frame consists of $g\,\mathrm{L}$ slots grouped in K-superslots. Thus, there are $\lfloor \frac{g\,\mathrm{L}}{K} \rfloor$ K-superslots. In FSA, each of the active users transmits its packet in a uniformly randomly chosen K-superslot of the frame.[2]

For the given n_{a}, L, g and K, the reliability of FSA can be calculated as

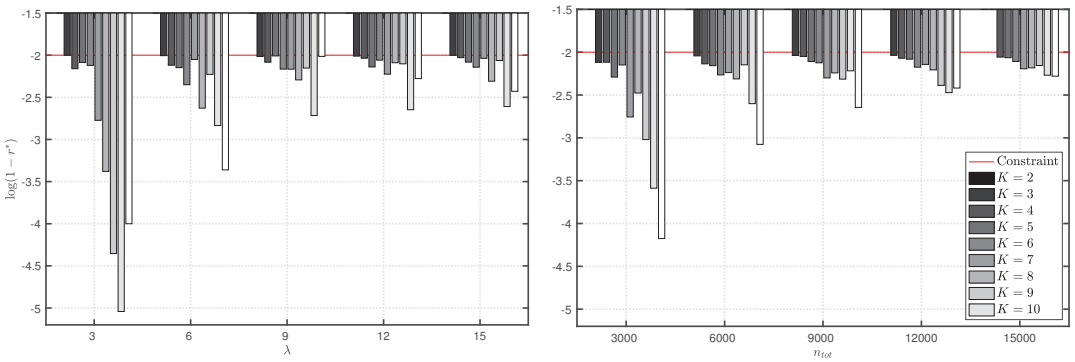
$$r(g, \mathrm{L}, n_{\mathrm{a}}) = (1-\pi)^{n_{\mathrm{a}}-1} \sum_{i=0}^{K-1} \binom{n_{\mathrm{a}}-1}{i} \left(\frac{\pi}{1-\pi}\right)^{i} \tag{4.28}$$

where $\pi = \frac{1}{\lfloor \frac{g\,\mathrm{L}}{K} \rfloor}$ is the probability of choosing a certain K-superslot. The proof of (4.28) is given in Appendix A. Using (4.28), as well as substituting it into (4.21), one can find the values of $g(n_{\mathrm{a}})$ and g^* through (4.23) and (4.24), as well as of T and T^* through (4.26) and (4.27), respectively.

In order to illustrate the effect of increasing K on reliability-latency performance, we consider an example where $g = 10$ and $\mathrm{L} = 5$, i.e., there are $g\,\mathrm{L} = 50$ slots available. Fig. 4.13 shows how $r(g, \mathrm{L}, n_{\mathrm{a}})$ behaves when n_{a} is varied in such setup. Obviously, for $n_{\mathrm{a}} \leq 50$, increasing K has a beneficial effect on $r(g, \mathrm{L}, n_{\mathrm{a}})$; such trend would continue until K reaches 50, when $r(g, \mathrm{L}, n_{\mathrm{a}})$ would become 1. On the other hand, the optimal K for $n_{\mathrm{a}} > 50$ depends on n_{a}. Nevertheless, note that for $n_{\mathrm{a}} > g\,\mathrm{L}$, the achievable levels of $r(g, \mathrm{L}, n_{\mathrm{a}})$ are small, as there are more arrived users than the number of slots.

In the next section, we turn to the throughput maximization for FSA with K-MPR when the reliability-latency requirement is fixed to $\mathrm{R}(\mathrm{L})$.

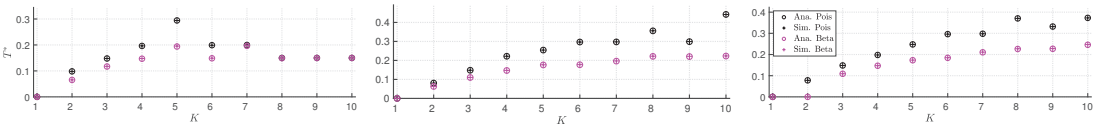
[2] A related analysis to the one considered here is made in [Sin+18], where the authors derived only T^* for the case when K is fixed to 8, but there is neither investigation of the behavior of T, nor the impact of varying K.



(a) Poisson arrivals

(b) Beta arrivals

Figure 4.14: Scenario: $1 - r^*(g^*, \mathrm{L})$ for different values of K, and (a) Poisson and (b) Beta arrivals, when $\mathrm{L} = 5$ and $\mathrm{R}(\mathrm{L}) = 0.99$. The reliability constraint is demonstrated with a red line, and the simulation results are illustrated with bar plots. It is demonstrated that the reliability constraint is satisfied with different K levels. However, with increasing K the constraint tends to be more overshot.



(a) $\lambda = 3$ and $n_{\text{tot}} = 3000$

(b) $\lambda = 9$ and $n_{\text{tot}} = 9000$

(c) $\lambda = 15$ and $n_{\text{tot}} = 15000$

Figure 4.15: T^* for different values of K and Poisson and Beta arrivals, when $\mathrm{L} = 5$ and $\mathrm{R}(\mathrm{L}) = 0.99$; the subplots depict results obtained for the same expected number of arrived users per frame.

4.3.4 Evaluation

In this section, we compare the throughput T and T^*, given by (4.26) and (4.27), respectively, for varying K and fixed L.

In order to take into account the non-stationarity of Beta arrivals, we adapt (4.21) in the following way

$$r^*(g^*, \mathrm{L}) = \sum_{t_s=1}^{T_A} \sum_{n_\mathrm{a}} r(g^*, \mathrm{L}, n_\mathrm{a}) \frac{\Pr[N_\mathrm{a} = n_\mathrm{a} | t_s]}{T_A/\mathrm{L}} \geq \mathrm{R}(\mathrm{L}), \tag{4.29}$$

where $\frac{1}{T_A/\mathrm{L}}$ is the probability to select any of the T_A/L intervals, used according to the law of total probability to calculate the expected number of users per interval. Finally, in the rest of the text, we assume that duration of L is equal to 10 ms for the Beta arrivals.

4.3.4.1 Comparison of analysis and simulations

In order to validate the analysis, we have implemented a discrete-time Monte Carlo simulator in MATLAB. We used Poisson arrivals and Beta arrivals with matching expected number of

arrived users per slot (i.e., average access load), varying K, fixing $\mathrm{L} = 5$ and $\mathrm{R}(\mathrm{L}) = 0.99$.[3] For the sake of precision, we have run 10^6 iterations for each scenario. At the start of each simulation run, the algorithm is provided $g \in \{1, \cdots, 40\}$ channels to select the minimum from, using (4.23) and (4.24); the number of channels is limited to demonstrate a realistic scenario. If no value of g is able to fulfill the reliability-latency constraint for the given K, the algorithm outputs $g = 0$ and that scenario is not simulated.

Fig. 4.14 shows simulated $r^*(g^*, \mathrm{L})$ of FSA with K-MPR when g^* is chosen according to (4.24), for Poisson and Beta arrivals and $\mathrm{R}(\mathrm{L}) = 0.99$ (note that $K = 1$ can not fulfill the requirements in the considered scenario). The results reveal that, as K increases, $r^*(g^*, \mathrm{L})$ becomes larger than the requirement $\mathrm{R}(\mathrm{L})$. This is due to the fact that with increasing K, i.e., increasing size of K-superslots, reflects in the granularity of the choice in g^* (recall that L is fixed), and consequentially in the granularity of potential values of $r^*(g^*, \mathrm{L})$. This overshooting of the reliability requirement also influences the throughput performance, as discussed next.

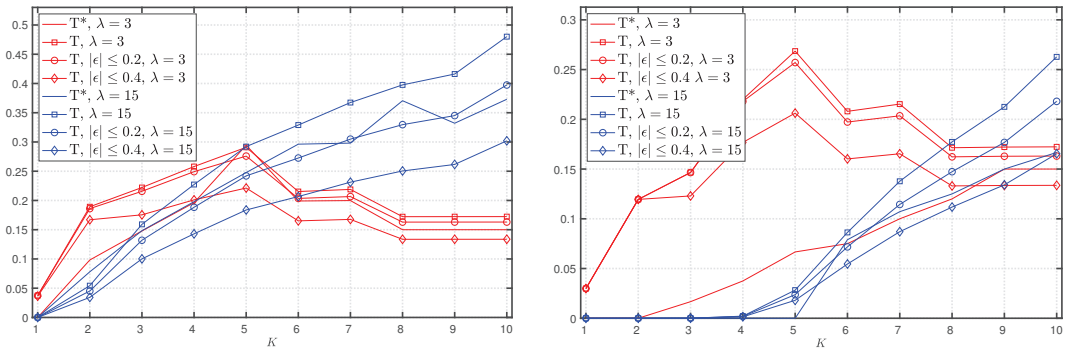
Fig. 4.15 shows throughput performance for the same settings as in Fig. 4.14. For the analytical solution, the Eq. (4.21) or Eq. (4.29) is used to calculate the g^*_{min}. This is followed by inputting the g^*_{min} into Eq. 4.27 to calculate the analytical throughput. The circles and pluses denote the simulation and analytical results, respectively; obviously, the results match. It can be seen that higher throughput can be achieved for Poisson arrivals, which can be expected due to the bursty behavior of Beta arrivals. Further, increasing K benefits the throughput in general. However, depending on the interplay between the values of the average load, K, $\mathrm{R}(\mathrm{L})$ and L, it may turn out that throughput drops after K exceeds some value; this is shown in Fig. 4.15(a), where the optimal K is 5 and not 10, which reflects the identified overshooting of the reliability-latency requirement shown in Fig. 4.14. The similar effect also exists in Figs. 4.16 and 4.17.

4.3.4.2 Comparison of throughput with and without estimation

We now compare throughput given by (4.26) and (4.27), to analyze the sensitivity to the exact knowledge of the number of arrived users. We assume that $\mathrm{R}(\mathrm{L}) \in \{0.99, 0.99999\}$, $\mathrm{L} = 5$, and investigate throughput performance for Poisson arrivals with $\lambda \in \{3, 15\}$ and Beta arrivals with $n_{\text{tot}} \in \{3000, 15000\}$, outputting the same average arrivals per slot with 1000 slots.

Fig. 4.16 is dedicated to resource efficiency with Poisson arrivals and latency-reliability constraints and shows that the estimation based throughput T outperforms the no-estimation

[3]We recall that duration of L is assumed 10 ms for Beta arrivals, which is equal to 5 generic time units in this section.



(a) $R = 0.99$ (b) $R = 0.99999$

Figure 4.16: Comparison of the throughput (y-axis)T and T* for for different reliability constraints as a function of multi-packet reception capability K (y-axis), $L = 5$, Poisson arrivals.

throughput T*, as it could be expected. This effect is more pronounced for the higher value of R(L).

Since the number of arrived users is typically not known a priori, but has to be estimated, we investigated the impact of the estimation error on the throughput performance. Specifically, we assume that the relative estimation error ϵ is bound as $|\epsilon| \leq \epsilon_{\max}$, e.g,. if the error bound is $\epsilon_{\max} = 0.2$ and the realization is $n_a = 10$ users, the estimation never estimates that there is $\hat{n_a} = 13 > n_a \cdot (1 + \epsilon_{\max})$ active users. The algorithm for selection of the number of frequency channels in (4.23) over-provisions by assigning $g_o(n_a) = g(\lceil n_a \cdot (1 + \epsilon_{\max}) \rceil)$ frequency channels. In the rest of the text we use ϵ instead of the maximum $\epsilon_{\max}$. The impact of the estimation error (and the related over-provisioning) on T is also depicted in Fig. 4.16, assuming that $\epsilon_{\max} = \{0.2, 0.4\}$, which may be considered as quite high values. Obviously, the over-provisioning plays it's role by decreasing T, such that for R(L) = 0.99, T becomes similar or worse than T* Nevertheless, for higher R(L), T with over-provisioning may significantly outperform T*, as shown in Fig. 4.16(b).

With an estimation error ϵ, we force the algorithm to allocate more channels $g(\lceil n_a \cdot (1+\epsilon) \rceil)$, compared to $g(n_a)$. T* behaves similar as T with $\epsilon = 0.2$ for Poisson arrivals with low reliability.

In Fig. 4.16(b) with a higher reliability constraint we see that for certain K values, the throughput T* is 0. This is due to an infeasible reliability constraint. However, we have a feasible solution with estimation and throughput T* is not zero.

Fig. 4.17 corresponds to the case of Beta arrivals, showing that, in comparison to Fig. 4.16, the gains in performance of T are more pronounced. For instance, T with $\epsilon_{\max} = 0.2$ fares better than T* when R(L) = 0.99. When R(L) = 0.99999, T is better than T* even with $\epsilon_{\max} = 0.4$. We also note that both Fig. 4.16 and Fig. 4.17 show that estimation is able to

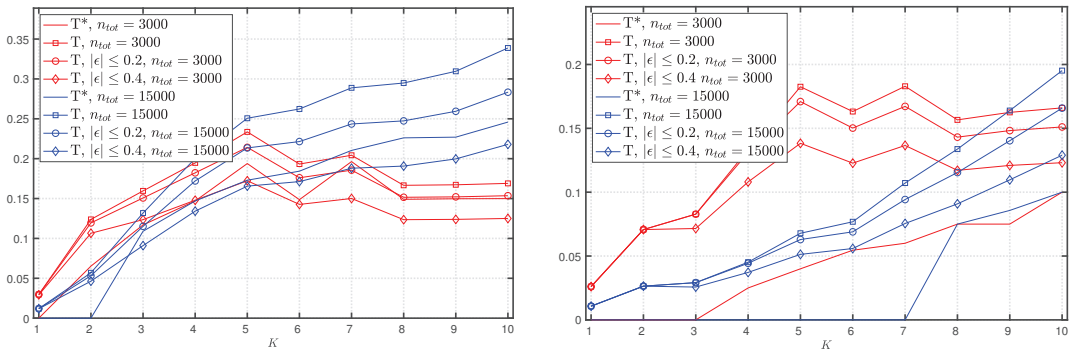
Scenario	Arrival	$K=1$	$K=3$	$K=5$	$K=10$
R=0.99	$\lambda=3$	0.0374	0.0743	0.0036	0.0224
	$\lambda=15$	∞	0.0731	0.1793	0.2870
	$n_{\text{tot}}=3000$	∞	0.3650	0.2049	0.1274
	$n_{\text{tot}}=15000$	∞	0.2095	0.4491	0.3779
R=0.99999	$\lambda=3$	∞	7.7919	3.0286	0.1488
	$\lambda=15$	∞	∞	∞	0.5767
	$n_{\text{tot}}=3000$	∞	∞	3.5643	0.6601
	$n_{\text{tot}}=15000$	∞	∞	∞	0.9521

Table 4.4: Normalized throughput gain $\frac{\mathrm{T}-\mathrm{T}^*}{\mathrm{T}^*}$. The cases that are infeasible for (4.21) and feasible for (4.20), are denoted by ∞.

"unlock" the use of lower values of K, with respect to case when only the knowledge of the arrival distribution is used to dimension g. For instance, in Fig. 4.17(b), one cannot have $K=1$ in the latter case, as $\mathrm{R}(\mathrm{L})$ constraint cannot be satisfied and the throughout T^* is 0. On the other hand, $K=1$ can be used if the estimation of the number of arrived users is performed, even with a high relative estimation error ϵ. The overall conclusion that when we are dealing with high number of users, it makes sense to estimate the number of users rather than to allocate resources based on the arrival distribution. One more specific conclusion is when $K=1$ estimation can be really crucial for certain scnearios. So in systems where $K>1$ cannot be deployed, estimation can enable high reliability constraints.

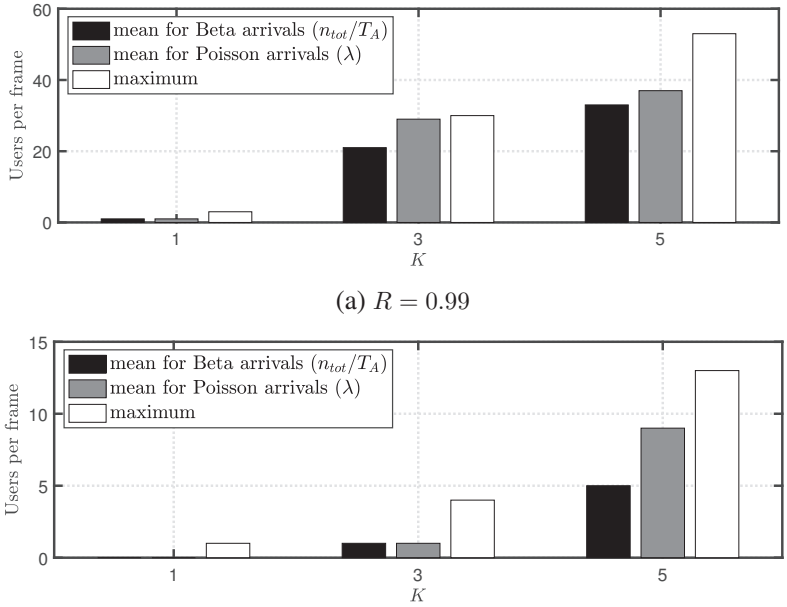
In Tab. 4.4 we shared the values of the ratio $\frac{\mathrm{T}-\mathrm{T}^*}{\mathrm{T}^*}$, which could be understood as the measure of the normalized gain in throughput if the number of arrived users is known. We see that the gain increases with increasing $\mathrm{R}(\mathrm{L})$ and increasing average load. Also, the gain for Beta arrivals is higher, as the distribution has a higher variance compared to Poisson arrivals.

Finally, in order to additionally illustrate the benefits of the knowledge of the number of arrived users, we consider a scenario in which the number of resources, i.e., L and g are fixed to $\mathrm{L}=5$ and $g=40$. Further, we assume that $\mathrm{R}(\mathrm{L})\in\{0.99, 0.99999\}$, and investigate for FSA with K the following: (i) what is the maximum number of users that can be admitted in the system at any given moment, (ii) average number of users that can be admitted in case of Poisson arrivals, and (iii) average number of users that can be admitted in case of Beta arrivals, Fig. 4.18 shows the corresponding results. Obviously, with increasing K, the gap between Poisson and Beta arrivals increases (i.e., λ becomes increasingly larger than $\frac{n_{\text{tot}}}{T_A/\mathrm{L}}$), and the trend holds for the maximum number of users versus the Poisson and Beta arrivals. This implies that, if one is able to estimate the actual number of arrived users, then one could use the existing resources much better, by letting in the system more users besides the ones that belong to the Poisson/Beta arrival process.



(a) $R = 0.99$ (b) $R = 0.99999$

Figure 4.17: Comparison of the throughput (y-axis)T and T* for for different reliability constraints as a function of multi-packet reception capability K (y-axis), L $= 5$, Beta arrivals.



(a) $R = 0.99$

(b) $R = 0.99999$

Figure 4.18: Number of users supported per frame (y-axis)) for different reliability constraints as function of multi-packet reception capability K (y-axis), $g = 40$, L $= 5$.

4.3.5 Summary

In this study, we have evaluated grant-free access scheme with reliability and latency constraints. We based our analysis on framed slotted ALOHA with K-MPR. FSA represents a single shot transmission algorithm, i.e. without any re-transmissions, while K-MPR is an abstraction of non-orthogonal multiple access schemes, seen as potential multiplexing solution in the coming 5G systems.

To stress the radio resource management perspective, we provide a definition of throughput for reliability-latency constrained grant-free access. The throughput is evaluated given different information on user activity i.e., the knowledge of the actual number of arrived users or just the knowledge of the arrival distribution. We have shown that with increasing reliability-latency requirements, the knowledge of the number of arrived users becomes more beneficial for the throughput performance. In this section we have assumed that the estimation for number of active users is available without any cost.

Contrary to our assumption in this section, the information about the number of arrived users is typically not available in grant-free access, but could be obtained using an estimation algorithm. In turn, the estimation algorithm involves an estimation error and also requires time-frequency resources for its execution, where it is reasonable to assume that the estimation error decreases as the number of resources dedicated to the estimation increase. However, such effects have to be analyzed.

Estimation algorithms are analytically characterized for their limitations in [Mag+18]. We take a more practical approach and develop a low-complexity estimation algorithm and compare the delay-constraint performance with the state of the art in the following section. The estimation algorithm empowers an admission control. Most importantly, the use of the admission control before resolution is critical to guarantee delay constraints.

4.4 A system level solution for stochastic delay constraints with user activity estimation: AC/DC-RA Admission Control based Delay Constraint-aware Random Access

In the previous section it is demonstrated that the use of estimation improves the resource efficiency. However, a perfect estimation is assumed with no cost in terms of radio resources. In this section we provide a practical example of an estimation algorithm and compare it with state of the art algorithms in terms of resource efficiency.

As there are limited number of resources in a practical system, more resources cannot always be allocated for more active users. Some of these users have to be rejected by the system. To overcome this practical limitation we introduce an admission control before the random access and after the user activity estimation process.

This can be achieved if the base station can use a means of direct intervention before an initial access of a user. This is similar to admission control in networks, where users are rejected if no capacity is left. Adaptation of the admission control for random access is enabled through separation of the initial access and the re-transmissions. The users are only

allowed to re-transmit, after the initial access, if an admission control allows them to use the resolution capacity.

To this extent, in this section we present Admission Control based Delay Constrained Random Access (AC/DC-RA) protocol, that provides stochastic delay bounds. *Stochastic Delay Constraint* is enabled through an *Admission Control* decision that is based on a novel collision multiplicity estimation algorithm.

Our contributions are four-fold:

1. We use a novel admission control decision, that takes place before the contention resolution (Sec. 4.4.4). This enables guarantees for traffic agnostic stochastic delay constraints for random access.

2. A novel user activity estimator is provided that is based on the famous Coupon Collector's Problem (Sec. 4.4.3.2).

3. We make use of a Parallel Multi-Channel Tree Resolution that re-arranges exploration of the contention slots in order to achieve stochastic delay bounds (Sec. 4.4.3.4).

4. We provide a dimensioning model for the suggested AC/DC-RA protocol which provides optimized use of system resources.

4.4.1 System Model

This section inherits the system model description in Sec. 4.1.1 and only new aspects are introduced here.

We define Quality of Service (QoS) as the reliability ($\mathrm{R(L)}$) that a packet is received at the destination within a certain delay constraint (L) after it is generated. We denote a set of sensors that have the same QoS requirement as class j and its reliability requirement as $\mathrm{R(L)}_j$ and delay constraint as L_j. The delay L incorporates delay stemming from re-transmissions due to collisions and reflect the performance of the random access channel. Any delay stemming from channel fading is not considered in this paper and only a radio resource perspective is evaluated.

We will use the term outer protocol for the traffic shaping part of the protocol that is achieved via the admission control and the term inner protocol for the contention resolution part.

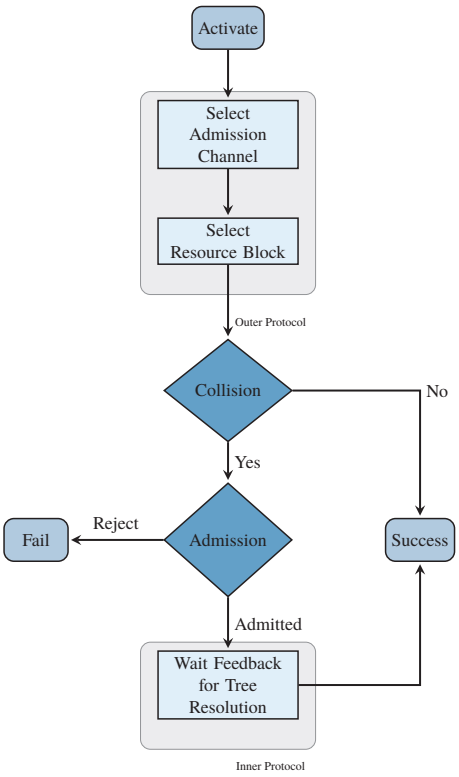


Figure 4.19: AC/DC-RA Flow Diagram - Sensor perspective

4.4.2 Proposal

We propose an outer protocol that separates the initial arrivals from backlogged users. An inner protocol that resolves each set of backlogged users in an isolated manner. The outer protocol is used for initial arrivals only. Users may collide through the use of the outer protocol. An admission decision is given for the collided users through the outer protocol. If admitted, the collided users access the inner protocol. The outer protocol uses an Admission Channel (AC) and the inner protocol uses a Resolution Channel (RC).

The admission is based on the stochastic delay constraint of the user, the collision multiplicity and the available capacity of the resolution channel. In the following, we explain in detail how this decision is taken.

The set of resources $\mathcal{M}_{AC}$ and $\mathcal{M}_{RC}$ form Admission Channel (AC) and Resolution Channel (RC) respectively where $M_{AC} + M_{RC} = M$ being the total number of resources with $M_x = |\mathcal{M}_x|$ denoting the cardinality of the set. There may be multiple admission channels with respect to each QoS class j denoted as $\mathcal{M}_{AC_j}$ and $\sum M_{AC_j} = M_{AC}$.

This protocol can be summarized with a flow diagram as given in Fig.4.19. When an event notification is received, the user is activated and starts using the outer protocol. It

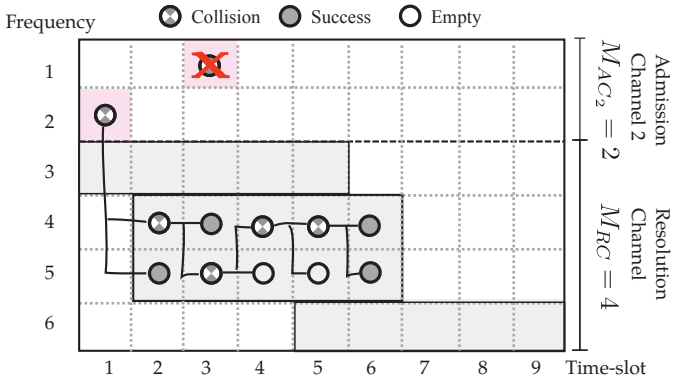


Figure 4.20: Resource separation between the inner and outer protocol of AC/DC-RA and the story of a set of requests that selected the same resource through AC/DC-RA protocol.

selects the **admission channel** that is appropriate for the QoS class. There are more than one admission channel so that the system can infer the QoS from the channel. Then it selects one of the *resources* in that channel. This selection is done with pre-set probabilities known to the user. It transmits the packet using that resource. This terminates the outer protocol. The outcome of the transmission can be a success or a collision. The central entity observes the outcome for that resource. If it is a success, the user is informed via a broadcast and it goes back to sleep mode. If a collision occurred, then an **admission control decision** is taken for that resource by the central entity. All users that have used that resource are either rejected or admitted and informed via a broadcast feedback. In case of a rejection, a user may have another radio interface. Or the sensor can report the failure to higher layers and trigger higher layer solutions e.g. switch to local control. In case of an admission, the inner protocol is initiated. The inner protocol used is a binary tree algorithm such that after each collision users have to re-select one of two new resources. The users are informed about the resources via a broadcast feedback. The feedback and the allocation method of these resources guarantee that all admitted users are successfully resolved by the inner protocol before the delay constraint.

An example for the resource allocation is illustrated in Fig. 4.20. The illustration shows the allocation of Admission Channel and Resolution Channel resources on the resource grid, where the horizontal axis represents time and the vertical axis represents frequencies. Resource use is colored in pink for initial access and in gray for backlogged access. The boxes depict the resources allocated to the respective resolution in RC. The rejection of an initial access is depicted with a red cross. For clarification of the example ternary outcome $(0, 1, e)$ of the resource use is illustrated with different symbols. A user that is required to report a fire within 100 ms with 0.99 reliability selects the admission channel 2 that represents its QoS requirement in this case. It transmits the data packet on frequency two at time-slot 1 which is a resource of that Admission Channel. The outcome is a collision as other users have selected

the same resource. Then the number of users that accessed this resource is estimated. The admission control decides that the resolution is possible within the delay constraint (100 ms) with the given reliability. The delay constraint is represented with 6 time-slots in the example. The admission control calculates the number of frequencies as 2 needed for parallelization. Then it checks if it is possible to allocate 2 frequencies in the resolution channel for that resolution. As there is available capacity in RC, the user in our example and all the other users that have collided with it are admitted to RC and resolved with a tree resolution. The allocated resource grid for the resolution is illustrated as a gray box surrounded with black lines. Within this gray box the collided users make a random selection on each time-slot bound to frequency 4 and 5. On time-slot 2, three of the users have selected the frequency 4 while one user has selected the frequency 5. In this case outcome on frequency 4 is a collision and on frequency 5 is a success. The collided users re-select one of the frequencies randomly again on the time-slot 3. This time one user has selected the frequency 4 and two users have selected the frequency 5. This results in another success. Two of the users still need to be resolved. Thus, this process continues until time-slot 6 where both of the users have selected their own resource. The resolution is completed before the delay constraint as guaranteed by the admission decision. In the meanwhile other sensors that collided at time-slot 3 are rejected since required capacity is not available in the RC to resolve these sensors in time.

4.4.3 AC/DC-RA - Outer Protocol

The outer protocol is used for the initial access of the devices. We do not use any collision avoidance mechanism to avoid delay before any user can reach the system.

Our proposal is based on two design choices. First, there are multiple Admission Channels and the user should select the one that is appropriate for the *Quality of Service class*. Second, we customize the resource selection probabilities within any of the Admission Channels to enable collision multiplicity estimation for arbitrary number of active users. Lastly, the Outer Protocol is terminated when this information is transferred to the admission control which is the gateway between two sub-protocols.

4.4.3.1 Separate Admission Channels - QoS Information

We assume that all the users have gone through an initial connection establishment or have overheard a broadcast. Through this information exchange, each user is aware of the appropriate admission channel for the required QoS.

There are multiple ACs for the initial access for different QoS classes, such that all the users in the same AC require the same delay bound and reliability. The AC is a set of resources

$\mathcal{M}_{AC_j}$ e.g. for QoS class j. Sum of all the resources orthogonal to time in the admission channels results in cardinality of the admission channels M_{AC}. As detailed in Section 4.4.1 a resource represents a single cell in the resource grid.

4.4.3.2 Resource Selection Probabilities - Collision Size Estimator

We assume that a set of users of size n_{a} at time instant t, selects randomly one resource from a set of resources in the admission channel at the same time. Depending on this selection a user may collide or be successful. Also some resources maybe unoccupied. The central entity can only observe the ternary outcome (0, 1, e) of these resources. From this outcome it has to make a collision size guess.

A similar estimation problem has already been investigated in the state of the art for RFID tag readings [KN06] for throughput optimization. However, the estimation time scales at best linearly with the number of users n_{a}. However, the work relies on Poisson approximation that is valid only with high number of resources. Usually, such resources are scarce and costly in terms of delay. To solve this problem, another work has considered the resource selection probabilities as a design parameter trading off precision for estimation speed [Ste+13]. Here, we aim at generalizing such an estimation to any number of active sensors and map it to the well-known Coupon Collector's Problem (CCP).

Coupon Collector's Problem There are M unique coupons that are obtained through independent draws from an urn with replacement. The problem is to find the expected number of draws until all M coupons are collected. Coupons may have equal or unequal selection probabilities. We will refer to selection probability of the i^{th} coupon as p_i such that,

$$1 = \sum_{i=1}^{M} p_i. \tag{4.30}$$

This problem is solved for equal and unequal coupon selection probabilities [AOR03]. We do not focus on expected number of draws until all M coupons are collected, but we will focus on the expected number of draws given a certain set of uniquely drawn coupons $\mathcal{M}_{s+c}$. Thus, we are guessing the expected number of draws that have been made given that a certain set of unique collected coupons.

Analogy to Collision Size Estimation We define a contention in a single time-slot t as an experiment. Suppose there are n_{a} users selecting M resources randomly on a contention basis at time-slot instance t. We observe the outcome of the contention on these M resources. We define the outcome on a contention resource i as o_i where a sequence of outcome is $\mathbf{o} =$

Dist.	Geom	Pois.	$p^0 = 10^{-2}$	$p^0 = 10^{-3}$	$p^0 = 10^{-4}$
$\hat{N}_{max}$	10^2	$2 \cdot 10^3$	$1.5 \cdot 10^2$	$1,1 \cdot 10^3$	$9 \cdot 10^3$

Table 4.5: Expected number of draws for Coupon Collector's Problem with $M = 18$ for various distributions.

$(o_1, o_2, o_3, o_4) = (1, 0, e, 1)$ for an example with $M = 4$. The ternary outcome $o_i \in \{0, 1, e\}$ of the contention for resource i is converted to the set of coupons collected. We consider idle resources as not-selected coupons, i.e., the $\mathcal{M}_{s+c}$ can be defined as,

$$i \begin{cases} \in \mathcal{M}_{s+c} \text{ if } o_i \neq 0 \\ \notin \mathcal{M}_{s+c} \text{ if } o_i = 0. \end{cases} \tag{4.31}$$

Using this set we calculate the expected number of draws $E[Z]$, corresponding to the estimated number of users at time-slot t $\hat{n_{\text{a}}}$. Then, the set of selected coupons can be written as $\mathcal{M}_{s+c} = \{1, 3, 4\}$ since resource 2 is idle. Using the probability of selecting any of the M resources.

The estimated number of active users $\hat{n_{\text{a}}}$ is given with expected number of draws given a set of uniquely drawn coupons with unequal probabilities

$$\hat{n_{\text{a}}} = E[Z|\mathcal{M}_{s+c}] = \sum_{z=0}^{\infty} \left(1 - \prod_{i \in \mathcal{M}_{s+c}} (1 - e^{-p_i z}) \right), \tag{4.32}$$

where the probability that a user did not select a resource i is multiplied for each resource for z users. Then this is subtracted from one to calculate the probability that all of these resources are selected at least once. Then the expectation is taken over z. The sum is up to infinity to calculate the probability of an outcome given there are up to infinite users. For large enough z, probability that a resource is not selected converges to 0. So Eq. (4.32) gives us the expected number of users given the outcome. Further explanation for Eq. (4.32) is given in App. B.

It is clear that each different selection $\mathcal{M}_{s+c}$ may give a different result in terms of number of users. We define the highest expected number of users as $EZ|\mathcal{M}_{s+c} = \mathcal{M} = \hat{N}_{max}$ for, $\mathcal{M}$, the outcome of the complete set, i.e., $\forall\ i\ o_i \neq 0$, where we have a collision or success on all resources. The estimation range for the number of active users n_{a} is up to $\hat{N}_{max}$. Therefore, the resource selection probabilities p_i should be adjusted, such that $\hat{N}_{max}$ is larger than the worst case number of active users. On the other hand it is intuitively clear that adjusting p_i to increase $\hat{N}_{max}$ results in further decrease in precision of the estimation. Otherwise we can decrease p_i to increase N_{max} to infinity.

In Table. 4.5 we have summarized $\hat{N}_{max}$ with different distributions of p_i. We have used the constraint in Eq. (4.30) in order to calculate p_i for various distributions. The p_i for each

distribution is as follows: (1) for geometric distribution with a fixed p we set the selection probability as $p_i = (1-p)^i \cdot p$, (2) for Poisson distribution with a mean λ we set the selection probability as $p_i = \frac{\lambda^i e^{-\lambda}}{i!}$, (3) for power series, defined the selection probability as $p_i = p^0 \cdot \alpha^i$. We have to set p^0 and adjust α accordingly. We then used the Eq. (4.32) to calculate $\hat{N}_{max}$. In Table. 4.5 we see that $p^0 \approx \frac{1}{N_{max}}$. Thus, using the power series we can easily adjust the estimation.

Collision Size Estimation After we have the estimated number of active devices $\hat{n_{\text{a}}}$, we will use the maximum likelihood to partition these devices into each resource. In the following parts we will use n_{a} instead of $\hat{n_{\text{a}}}$ for ease of reading.

The problem is now to partition n_{a} users to M bins. The partitioning is constrained with the outcome $\mathbf{o}$, i.e., collision on resource 2 and success on resource 5 translates in to $o_2 = e, o_5 = 1$. Possible guesses $\mathbf{g}$ will be sequences that fulfills the outcome constraints. The guess of resource i in the x^{th} sequence is g_i^x. We also use g_i for a guess for resource i, and $\mathbf{g^x}$ as the guess sequence x. Now we can write the constraints

$$g_i \begin{cases} = 0 & \text{if } o_i = 0 \\ = 1 & \text{if } o_i = 1 \\ \geq 2\,, \; \leq \left(n_{\text{a}} - \sum_{j=1}^{i-1} g_j\right) & \text{if } o_i = e. \end{cases} \tag{4.33}$$

We define the guess set $\mathcal{G}$ such that it involves all guess sequences fulfilling a given outcome sequence $\mathbf{o}$ and the number of active devices n_{a}. For example, with $M = 3$ and a outcome sequence of $\mathbf{o} = (o_1 = 1, o_2 = e, o_3 = e)$ where we have $n_{\text{a}} = 7$ we will have $\mathcal{G} = ((1,2,4),(1,3,3),(1,4,2)) = \{\mathbf{g^1}, \mathbf{g^2}, \mathbf{g^3}\}$, such that $g_3^2 = 3$ and $\mathbf{g^2} = \{1,3,3\}$.

We can calculate the probability of a correct guess as in

$$p[\mathbf{g}] = \prod_{i \in \mathcal{M}} \left(\binom{n_{\text{a}} - \sum_{j=1}^{i-1} g_j}{g_i} (p_i)^{g_i} \right). \tag{4.34}$$

This will enable calculation of the most likely partition, to have an estimate on how many users contend each resource as

$$\hat{\mathbf{u}} = \arg\max_{\mathbf{g}} p[\mathbf{g}], \forall \; \mathbf{g} \in \mathcal{G}, \tag{4.35}$$

where $\hat{\mathbf{u}}$ is the sequence for the collisions size estimation for all resources. The equation is complex to calculate with increasing dimensions of $\mathbf{g}$ as it is a combinatorial maximum likelihood calculation. It depends on N_{max} and cardinality $|\mathbf{g}|$ such that ${N_{max}}^{|\mathbf{g}|}$ cases may

be evaluated depending on the feedback. For practical implementations a heuristic estimator can be used an example is as such

$$\hat{u_i} = \begin{cases} \lceil p_i \cdot \hat{n_\mathrm{a}} \rceil & o_i = e \\ o_i & o_i \neq e \end{cases}, \tag{4.36}$$

where $\hat{n_\mathrm{a}}$ is the total number of user estimation given by Eq. (4.32) that uses the outcome sequence $\mathbf{o}$ and N_{max} as input

Comparison As a comparison for our estimation technique, we choose two maximum-likelihood estimators (MLE). First one is based on the observation of non-idle resources only $M_{s+c} \triangleq \sum_{i=1}^{M} \mathbb{1}_{o_i \geq 1}$ (that is, without knowledge of the number of idle resources), where $\mathbb{1}$ is the indicator function . The MLE operates on the following exact probability of observing M_{s+c} non-idle resources, given a total of M resources and a total of n_a sensors:

$$P_{\text{MLE}}[M_{s+c}|M, n_\mathrm{a}] = \frac{\left\{ {n_\mathrm{a} \atop M_{s+c}} \right\} M!}{M^{n_\mathrm{a}}(M - M_{s+c})!},$$
$$\hat{n_\mathrm{a}} = \arg\max_{N_t} P_{\text{MLE}}[M_{s+c}|M, n_\mathrm{a}] \tag{4.37}$$

where $\left\{ {n_\mathrm{a} \atop M_x} \right\}$ are the Stirling number of the second kind.

Second comparative technique is adaptation of the work from Zanella [Zan12b] on the RFID collision set estimation. The work is based on observing the number of collided M_c and successful M_s resources, and, using the approximation of the exact expression, computes the maximum-likelihood n_a by finding the roots of the expression, i.e. finding the number of resources that maximizes the idle likelihood while minimizing the collision likelihood as in:

$$\frac{n_\mathrm{a} - M_s}{M_c} = \frac{\frac{n_\mathrm{a}}{M}(e^{\frac{n_\mathrm{a}}{M}} - 1)}{e^{\frac{n_\mathrm{a}}{M}} - 1 - \frac{n_\mathrm{a}}{M}}. \tag{4.38}$$

The average collision size is then computed from $\hat{n_\mathrm{a}}$ as in $\frac{\hat{n_\mathrm{a}} - M_s}{M_c}$. It has to be noted that, since neither of MLE approaches vary the resource selection probabilities (i.e., both use uniform probabilities), none of them can give a reliable estimate above a certain total number of active devices N_{max}, i.e., whenever $M_c = M$ is observed.

We have conducted Monte Carlo simulations in MATLAB for comparing the estimators. The resource selection probabilities are set with respect to power distribution calculated in section 4.4.3.2 for CCP and N_{max} values are set as 500, 1000 and 2000. The resource selection probabilities are set uniformly for the baseline case. The reason for this selection is that the state of the art uses the Poissonization of the outcomes which is a valid approach

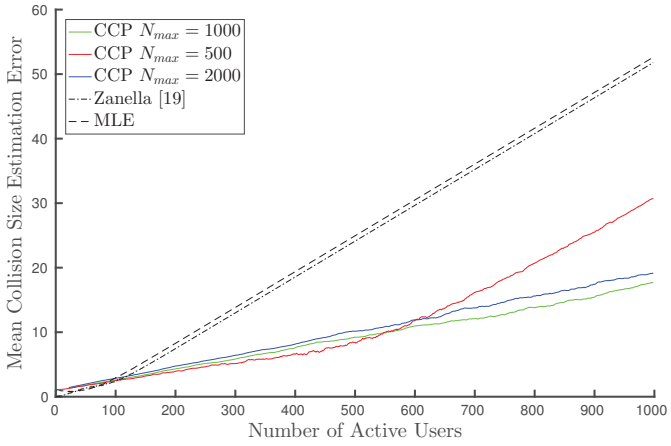


Figure 4.21: Mean collision size estimation error

only with equal resource selection probabilities. In Fig. 4.21 collision size estimation error is plotted with 18 resources M. The absolute estimation error is calculated as $|\hat{u}_i - u_i|$ taking the difference between estimated and actual number of users per resource i, which is then averaged as in $\mathrm{E}[|\hat{u}_i - u_i|]$ over multiple runs and multiple resources. CCP is compared against the state of the art with varying the number of active users from 1 to 1000. Each number of active users are simulated for 1000 runs. The limitation of uniform resource selection is observed from the results. The MLE estimator saturates with $M = 18$ after 100 users since the observation is always a set of collisions when the resource blocks have equal probability to be accessed. Thus, the MLE estimates 100 users with full collision set and the error linearly grows with the number of active users. In CCP, with increasing number of users an idle occurs and this enables scalability up to N_{max} active users.

We have also evaluated an error in the setting of N_{max} and how such a wrong setting will affect the system in Fig. 4.21. The N_{max} set to 500 represents the case where we may have more users accessing the medium than the allowed maximum. We see that the absolute estimation errors are almost the same up to 500 active users. After this point the estimation error grows linearly with increasing number of users similar to the state of the art. On the other hand the case where N_{max} is set to 2000 represents that we always have a higher limit for maximum number of users compared to active number of users. This has a less critical effect compared to setting a lower maximum limit. This can be observed in Fig. 4.21 where the absolute error has increased slightly but is in general lower compared to the previous case. Thus, it can be concluded that a relatively high N_{max} can be selected to avoid the saturation effect.

The scalability comes with the cost of precision loss with low number of active users. Even though, the mean error difference is approximately 1 user up to 200 active users, the state of the art is better than the CCP.

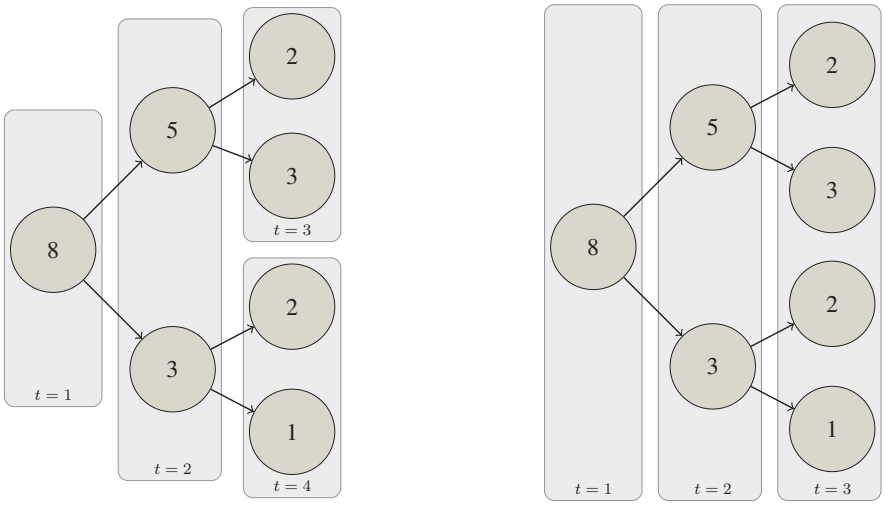
The precision of the estimation is evaluated on average. Thus, the strictness of the stochastic delay constraints provided through the use of the estimator is valid on a set of realizations, but not for each realization of the random process. Also, the stochastic delay constraint would be valid if the number of arriving users is upper-bound so the exact estimation can be converted to an upper-bound for reliability. We enable this via adding the mean estimation error $\mathrm{E}[|\hat{u}_i - u_i|]$ from the analysis as a pessimism factor on top of the collision multiplicity estimate u_i. This makes sure that the stochastic delay constraint is not violated due to estimation error. We have evaluated the results for the guarantees where the estimator is integrated in the system in Sec. 4.4.6.

The outcome of the estimation and the QoS requirement is obtained from the initial access of the users to the admission channel. Given these information the delay of the contention tree resolution can be obtained through stochastic analysis. This information enables the admission control decision. In the following section we investigate the stochastic delay analysis of the inner protocol.

4.4.3.3 AC/DC-RA - Inner Protocol

In this subsection, we first shortly introduce the inner protocol and we investigate the stochastic analysis for delay constraints. The details of the inner protocol and the analysis can be found in Chapter 5.

The resolution algorithm used is a version of binary tree resolution for each collision. The algorithm uses a breadth-first exploration of the tree. An advantage of breadth-first is a possible exploration of multiple contention slots simultaneously. We call this *parallelization* of the resolution and M_P denotes the number of parallel allocated resource for a resolution. An example with two possible tree algorithm parallelizations is given in Fig. 4.22. The resolution starts with 8 users and with the first split 3 users select one resolution slot while the remaining 5 select the other resolution slot. The users are resolved with a parallelization of 2 and 4. In the case of parallelization of $M_P = 2$ the resolution needs a capacity of 2 frequencies for a duration of 4 time-slots to schedule all resolution slots. However, with a parallelization of $M_P = 4$ the resolution needs a capacity of 4 frequencies for a duration of 3 time-slots. Thus, required capacity increases since higher amount of parallel resources are blocked for faster resolution.



(a) Parallelization with 2 frequencies $M_P = 2$ (b) Parallelization with 4 frequencies $M_P = 4$

Figure 4.22: Example of parallel exploration of trees.

4.4.3.4 Delay Constrained Resolution

In this section, we investigate how the analysis for the inner protocol can be used for the admission control.

For a stochastic delay constraint L and reliability $\mathrm{R(L)}$, e.g., $\mathrm{R(L)} = 0.95$, means that the delay constraint L should be achieved 95 percent of the time.

Stochastic delay bounds for MP-CTA are given in [GAK17b] for different number of users. These values can be placed in a look up table (LUT) for varying N number of users, L the delay, for a specific reliability $\mathrm{R(L)}$ as in Tab. 4.6. The LUT then outputs the minimum number of parallelization M_P required to fulfill the stochastic delay constraint of all the devices in the contention resolution. If it is infeasible then it returns zero. For example, given 10 users and a delay constraint of 5 slots, it is infeasible to achieve a resolution where all users are resolved with 0.95 reliability. This is denoted as $M_P = 0$. However, a delay constraint of 10 slots is achievable with a parallelization of $M_P = 3$.

We use this analysis and define a function f that outputs the number of resources M_P given the required reliability and delay constraint with the number of users,

$$f(L_i, R_i, N) = \begin{cases} 0 & \text{if infeasible} \\ M_P & \text{if feasible.} \end{cases} \tag{4.39}$$

Infeasibility is invoked when allocation of all the M_{RC} frequencies in the resolution channel cannot achieve the required delay then $f = 0$ is returned.

Delay Constraint	5	10	15	20	25	30	35
Users							
5	0	1	1	1	1	1	1
10	0	3	2	1	1	1	1
15	0	4	2	2	1	1	1
20	0	6	3	2	1	1	1
25	0	8	4	3	2	2	2
30	0	9	4	3	2	2	2
35	0	11	5	4	3	2	2
40	0	13	6	4	3	3	2

Table 4.6: The parallelization M_P, given in table, needed to resolve certain number of users for varying delay constraints L_j and a reliability level $\mathrm{R(L)}_j = 0.95$. The reliability level is not a dimension of the table.

We can check a concrete example using the values shared in Tab. 4.6[4]. An example would be for a delay constraint of 15 slots with 20 users and a reliability of 0.95 percent. We can read the cross-section of these values to see the required parallelization. This can be formulated as $f(15, 0.95, 20) = 3$ such that we can use a parallelization of 3 to achieve the stochastic delay constraint in an efficient manner. The required parallelization is 2 for 10 users, and 4 for 25 users. Thus, we can allocate just the right number of resources to achieve the stochastic delay constraint.

In this section we have shown that a delay constrained resolution is achievable through the MP-CTA. In the following section we explain how the information provided via the outer protocol will enable guarantees though use of the inner protocol, this leads us to the admission decision.

4.4.4 AC/DC-RA - Admission Control

AC/DC-RA is not improving the throughput of random access but limiting delay for a resolution. Thus, dealing with increasing number of users is still an issue. In order to investigate the scaling problem, we have to consider the capacity of the Resolution Channel.

We define *capacity* as a set of resolution resources. The resources for resolution is fixed in terms of frequency and time. For instance, the **capacity** required to resolve a collision given in Fig. 4.20 is a 2 frequency 5 time-slot grid. The 2 frequencies are blocked for 5 time-slots. The capacity of the Resolution Channel is also defined in terms of frequencies M_{RC}. It is clear that not all collisions will fit in the RC. Thus, to guarantee that users admitted to the system are always served within the stochastic delay constraint, we have to reject some of

[4]The values shared in the table are calculated using the analysis in Chapter 5.

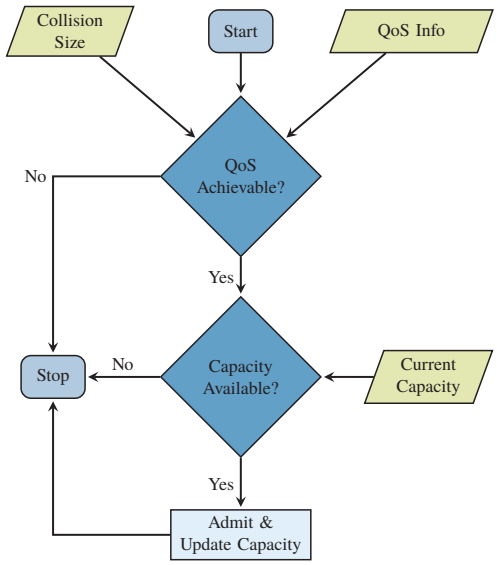


Figure 4.23: Admission control decision state diagram

the users. The decision whether to reject the users or to admit them to RC is done by the admission control of AC/DC-RA.

4.4.4.1 Admission Control

The admission is decided through evaluation of QoS information, collision size information against the resolution channel capacity. We zoom in the admission block from Fig. 4.19. We provide another flow diagram for the admission control decision in Fig. 4.23. The **QoS information** is extracted in terms of L_j and $\mathrm{R(L)}_j$ from the selected admission channel index j. The collision size estimation returns the vector $\hat{\mathbf{u}}$ where $\hat{u}_i$ is the **collision size estimation** for the i^{th} resource. We add the mean estimation error for the expected N_{max}, calculated with $\mathrm{E}[|\hat{u}_i - u_i|]$ as a pessimism factor to each collision multiplicity estimation that gives $\hat{u_i}^{\dagger}$. As the realization n_{a} is unknown, the estimation $\hat{u_i}^{\dagger}$ has to be used in this case for the delay constrained resource allocation calculation. The admission control feeds this information to the stochastic tree analysis $f(L_j, R_j, \hat{u_i}) = M_{P_{j,i}}$ to obtain the number of required resources. In case the QoS is not achievable, i.e., $M_{P_{j,i}} = 0$, the devices are directly rejected. If not, the requested number of resources are compared against the available number of resources in RC. If there is enough capacity the users are let into the system for resolution or else are rejected. The admission decision $D_{j,i}$ that is given for all users in resource i of the admission channel j can be summarized as in,

$$D_{j,i} = \begin{cases} \text{Reject} & \text{if } M_{P_{j,i}} = 0 \text{ or } M_{P_{j,i}} > M_{RC}^t \\ \text{Accept} & \text{if } M_{P_{j,i}} <= M_{RC}^t, \end{cases} \tag{4.40}$$

where M_{RC}^t is the number of available resources in the resolution channel at time-slot t and is updated as $M_{RC}^t \leftarrow M_{RC}^t - M_{P_{j,i}}$ after an accept decision. It is initialized as $M_{RC}^t = M_{RC}$ and after each resolved contention, the freed resources are added back.

The system will operate in a resource limited environment such that allocation of resources to admission channel and resolution channel will impact the behavior of the system. In order to analyze this trade-off we propose an analytical model.

4.4.5 Analysis

We propose a Markov Chain model to analyze the system given in Fig. 4.24. We simplify the system to five different states. Initial state is an *Off* state that represents the user activation characteristics with respect to the application. When active with the probability p_{on}, the user goes to the transmission state *Tx*. This state is the initial access state, and the user selects one of the resources, i, in the j^{th} admission channel $\mathcal{M}_{AC_j}$ and transmit a packet with that resource. This selection is done on the appropriate admission channel for QoS class.

The initial access is a success with probability $1 - p_c$. Then the user may go to success state *Suc*. After the transmission is completed it goes back to *Off* state. If the initial access results in a collision it goes to the admission state A_R with probability p_c. In this state the number of collided users with that specific user is estimated and a decision whether resolution time is within QoS class of the user is given.

After initial access, the user is admitted with probability $1-p_r$. After successful contention resolution it proceeds to the *Suc* state. If the user cannot be admitted then it is rejected with probability p_r and goes to the fail state *Fail* where it informs higher layers before going to the *Off* state.

We can extract the state probabilities in terms of state transition probabilities as,

$$P_{Off} = \frac{1}{1 + 3p_{on} + p_{on}p_c\left(1 - p_r\right)} \tag{4.41}$$

$$P_{Tx} = \frac{p_{on}}{1 + 3p_{on} + p_{on}p_c\left(1 - p_r\right)} \tag{4.42}$$

$$P_{A_R} = \frac{p_c p_{on}}{1 + 3p_{on} + p_{on}p_c\left(1 - p_r\right)} \tag{4.43}$$

$$P_{Suc} = \frac{\left(p_r p_c - p_c p_r\right) p_{on}}{1 + 3p_{on} + p_{on}p_c\left(1 - p_r\right)} \tag{4.44}$$

$$P_{Fail} = \frac{\left(1 - p_r p_c + p_c p_r\right) p_{on}}{1 + 3p_{on} + p_{on}p_c\left(1 - p_r\right)}. \tag{4.45}$$

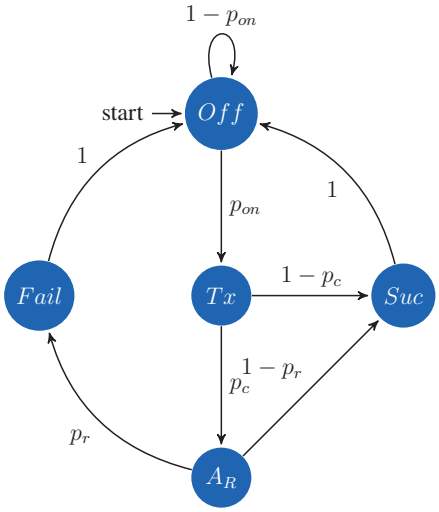


Figure 4.24: Markov Chain for AC/DC-RA

We investigate the state transition probabilities as follows: The activation probability depends on the application. For the sake of steady state analysis we consider Poisson arrivals in this scenario, which is usually assumed for machine activity [Dem+09] and [3GP12]. To provide an average dimensioning we assume the probability that a user generates any packet between two random access opportunities, and the total mean arrival rate as λ with activation probability $p_{on} = 1 - e^{-\lambda}$.

4.4.5.1 Collision Probability p_c

For the calculation of the collision probability we suggest modeling the problem as a bins and balls problem with unequal probabilities that gives the final equation for $p_c[u]$ probability of collision with u users as $p_c[u = n] = \sum_{j=1}^{J} \sum_{x \in \mathbf{S}_x^J} \left(P_J^x[n] \cdot j \right)$, where partitions with size u are counted $P_J^x[u]$ up to J repetitions and weighed accordingly for partitioning N users in M bins. Further proof of this equation is given in App. B.2. Thus, we can calculate the probability of a collision p_c as, $p_c = 1 - p_c[1] - p_c[0]$.

After the Transmission State then we move to the Admission State.

4.4.5.2 Admission Rejection Probability p_r

Collisions are resolved with the tree algorithm. Each of these resolutions occupy $M_P \cdot \mathrm{L}$ resources where M_P is selected with respect to the number of collided users and L is the delay constraint in terms of time-slots. As we have finite resources in our system, allocating resources to the resolutions can be considered as a serving process. Thus, we model the serving of a resolution as a queue, Random Access Queue (RAQ), process, where each

resolution resource is a server and arrivals are collisions to be served. It is a queue with no buffer since the admission decision is given instantly. In this section, we investigate the RAQ model in order to analytically provide the blocking probability in such a queue, that will be representing an admission rejection probability p_r decision due to insufficient amount of resources in RC.

We model each collision as an arrival to the RAQ. Since we expect a collision on all resources to use admission channel effectively, the average number of collisions can be written as $\lambda_{RAQ} = M_{AC_j}$ for class j. Thus, on heavy load, we expect a collision on all AC resources, i.e., deterministic arrivals. With low load, we expect probabilistic number of collisions thus, a Markovian number of arrival to the RAQ. The analysis is based on the assumption of deterministic arrivals.

In order to guarantee the resolution time we reserve frequencies during the resolution. For serving time we have a deterministic value $h_{RAQ_j} = \mathrm{L}_j$, such that the serving time for each QoS class depends only on the delay constraint.

The number of available resources is converted to the number of servers. The number of servers is determined by the expected parallelization of the resolution in the system. We can calculate the expected level of parallelization as in,

$$E[M_P] = \sum_{u=0}^{N_{max}} f(\mathrm{L}_j, \mathrm{R(L)}_j, u) P_c[u], \tag{4.46}$$

where $P_c(u)$ is the probability that a collision with size u occurs and given with Eq. (4.4.5.1). Thus, each resolution needs on average M_P resources. And we have M_{RC} resources in total. Via dividing the total number of resources to the average number of resources per resolution we can calculate the expected number of on-going resolutions as

$$M_G = \left\lfloor \frac{M_{RC}}{E[M_P]} \right\rfloor \tag{4.47}$$

where M_G represents the average number of servers in the resolution channel.

Through this we can write the admission rejection probability p_r as

$$p_r = \frac{\frac{\left(L_j \cdot M_{AC_j}\right)^{M_G}}{M_G!}}{\sum_{o=1}^{M_G} \frac{\left(L_j \cdot M_{AC_j}\right)^{o}}{o!}}. \tag{4.48}$$

The proof is driven using call blocking probabilities in [Gim65] and details are given in Section B.3. Finally, we have all the parameters required to analyze the protocol.

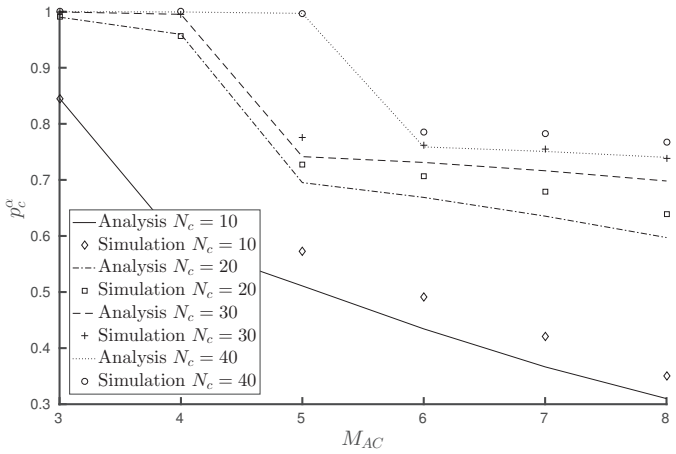


Figure 4.25: Evaluation of p_c^α with varying the resources in admission channel M_{AC}, Poisson arrivals with distinct means of $N_c = 10, 20, 30, 40$ users are evaluated.

4.4.6 Evaluation

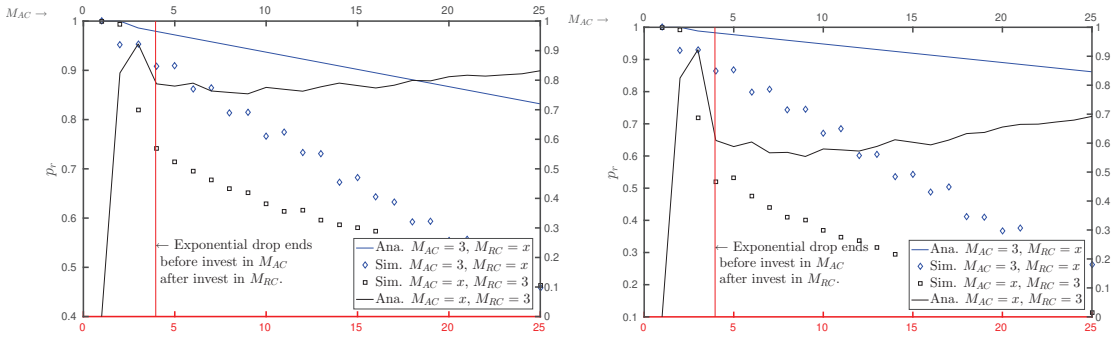
In this section we first evaluate the suggested algorithm in a prioritization scenario. Following this we compare our analysis with simulation results to show that the analysis provides a reasonable estimate to enable analytic dimensioning of the system. All the simulations are done in a MATLAB based discrete time simulator.

We make the following assumptions in our simulator. There is zero propagation time. A collision channel model is implemented based simulator on MAC layer and perfect channel conditions are assumed. Costless and immediate feedback is assumed which is necessary for both tree and access barring based solutions. A single cell scenario for uplink traffic is investigated. Resources are organized in time and frequency.

4.4.6.1 Comparison with Analysis

We investigate the behavior of the protocol with various resource separation decisions and to show validity of the analysis we simulate the AC/DC-RA with varying number of resources for admission channel M_{AC} and resolution channel M_{RC} and compare with our analysis. While varying the size of one channel we fix the other to $\{15, 25, 45\}$. We assume a Poisson arrival rate with average of 30 users per time-slot.

We compare the analysis of number of arriving collisions with simulations in Fig. 4.25 where we plotted the varying number of resources in admission channel M_{AC} against the collision probability. The collision probability is the expectation with a given Poisson arrival. Since in simulations we use Poisson arrivals, we use the law of total probability over the



(a) Fix one to 25 (b) Fix one to 45

Figure 4.26: Comparison of AC/DC-RA analysis with simulations varying the amount of allocated resources to M_{AC} and M_{RC} with average 30 users per slot.

probability of observing different number of devices as

$$p_c^{\alpha}[N_c] = \sum_{i=0}^{\infty} e^{-N_c} \frac{(N_c)^i}{i!} p_c[i], \tag{4.49}$$

where $p_c^{\alpha}[N_c]$ is the probability adjusted for Poisson arrivals with mean N_c.

We see that the with the pre-selected resource selection probabilities we can trace the collision probabilities with the given analytic. It is important to emphasize that since the complexity of the calculation grows exponentially it can only be used for offline dimensioning of the algorithm. Another observation is that the power series has a higher success rate than expected when it comes to sacrificing the throughput on the admission channel. For instance with 4 resources and 10 users, only 60% of the resources have seen a collision on average. Since we expect 1 out of 4 resources to be free so that the estimation works on the edge, we expect around 75% collisions.

In Fig. 4.26 we varied number of resources in resolution channel and admission channel on the x-axis and we plotted the admission rejection probability on the y-axis. The analysis is given with a solid line while the simulation is marked with data points. Due to the unequal resource selection probabilities, the users will forcefully collide. Through this, we have the same number of collisions as the number of resources in the admission channel for high arrival rates. Thus, the assumption of deterministic number of collided resources is valid for low number of resources for admission channel. However, this assumption does not hold if we a have high number of resources in the admission channel M_{AC} such that more than 1 slot may be empty. Thus, the analysis should rather be used with low number of resources in admission channel M_{AC}.

For varying the number of resources in resolution channel M_{RC}, almost a linear behavior is observed for the rejection ratio. This is expected, since a better parallelization is enabled and resolutions with high number of users are almost linearly parallelizable.

The results for varying M_{RC} in Fig. 4.26: the increase in parallelization results in decreased rejection ratio as expected. For the really low M_{RC} region, the curve has a better fit as explained previously. After the deterministic behavior for the number of arrivals vanishes, the curve deviates. Here we can emphasize a take out message for p_r. With low M_{AC}, increasing the M_{AC} exponentially decreases p_r then after a certain number of resources it saturates to a linear decrease. This behavior is similar to that of a queue close to the stability limit. For M_{RC} we have a linear decrease with a greater pace compared to the linear region of M_{AC}. Thus, we can conclude that a rule of thumb for dimensioning the resources for M_{AC} and the M_{RC} is: (1) allocate enough resource to M_{AC} such that exponential p_r behavior is overcome and (2) all the remaining resources are allocated to M_{RC}. The exponential region limit can be determined Eq. (4.48) and taking the dip of the waterfall region as observed from Fig. 4.26. For example in Fig. 4.26b, $M_{AC} = 5$ and in Fig. 4.26a $M_{AC} = 4$ should be selected and all other resources should be allocated to M_{RC}.

Through the provided insights for the dimensioning of the algorithm, we now set the resources of AC/DC-RA accordingly and compare with the state of the art.

4.4.6.2 Comparison with Baseline

We select the Dynamic Access Barring (DAB) algorithm as a baseline [WW15b]. This algorithm is an improved version of the access class barring algorithm currently used in LTE RACH. Through a backlog estimation the barring factor is updated dynamically. The barring of users enables optimal saturation throughput of Slotted ALOHA. It is also used with multiple QoS classes such that one class is prioritized over other and the no-priority class is fully barred when there are requests from the prioritized class. A dynamic barring factor is still applied to the prioritized class to guarantee optimal throughput.

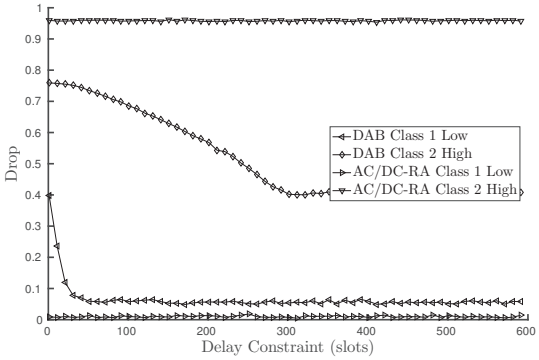
For AC/DC-RA we allocate 4 resources for each admission channels for each Class 1 and 2 and 12 resources for the resolution channel allocating 20 resources in total. For DAB algorithm we also allocate 20 resources to have a fair comparison. We use a deadline to refer to the delay constraint for comparison.

For AC/DC-RA we enable such prioritization through admission control, where one class is only accepted after the other class is fully admitted. Since we want to emphasize the priorities and the guarantee aspects, we use the same requirements for both classes. In order to show that the system can outperform the state of the art in extremely critical situations, we assume a Beta distributed arrival scenario representing bursty arrivals of M2M communications [3rd00]. We have an activation time of $T_A = 100$ slots for the beta arrival and we have other parameters of the distribution set as in the reference. The total number of users accessing the system is denotes as n_{tot}. There is an *imbalance* between different traffic classes. The imbalance reflects

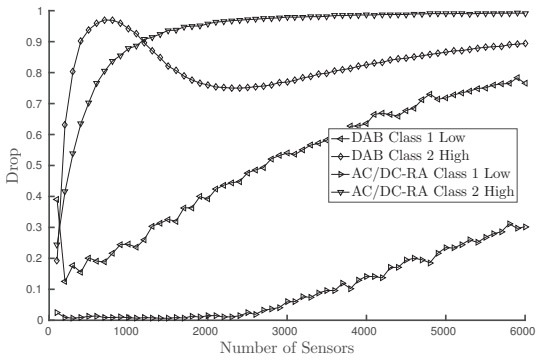
a population ratio difference between traffic of two classes. We keep the naming as Class 1 and Class 2 where Class 1 denotes the prioritized class. However, an adjective is added to the classes to point out the traffic imbalance situation. These adjectives are Low and High, where the High class has 10 times more users than the Low class.

In Fig. 4.27 we have plotted the AC/DC-RA against the baseline with the traffic imbalance. In Fig. 4.27a we have plotted the drop plus rejected ratio for varying delay constraints for Low Class 1 with $n_{\text{totlow}} = 200$ users and High Class 2 with $n_{\text{tothigh}} = 2000$ users. The decrease in the number of users in the Low Class 1 results in an increased percentage of serviced users and most of the High Class 2 is blocked out. In Fig. 4.27c we have plotted the drop plus rejected ratio for varying delay constraints for Low Class 2 with $n_{\text{totlow}} = 200$ users and High Class 1 with $n_{\text{tothigh}} = 2000$ users to represent a more scarce scenario. The Low Class 2, that uses DAB, achieves a lower drop ratio thanks to low number of users. Some of the users from High Class 1 cannot be resolved in time even though the delay constraint is large. This is due to limited resource in RC that cannot react to burst arrivals. Interestingly for DAB, with larger delay constraints both classes achieve lower drop ratios compared to AC/DC-RA. This stems from the fact that obtaining delay and multiplicity information in the scenarios with relaxed delay constraints is not necessary for timely resolution. And the loss in resources to obtain this information cannot be made up with increased efficiency in the resolution channel. In Fig. 4.27b and Fig. 4.27d the drop plus rejected ratio is plotted against varying number of users. The x-axis depicts the number of users for High Class 2 in Fig. 4.27b and High Class 1 in Fig. 4.27d. For Low Class 1 almost no user is rejected and resolution is optimized with respect to multiplicity information. Thus, the information obtained from AC is used. However, the multiplicity information obtained for High Class 1 cannot help as there is not enough capacity in the resolution channel and these users have to be rejected irrespective of their multiplicity. This guides us to an important conclusion that if there is low amount of resolution channels and relaxed delay constraints the state of the art protocols can perform better. In Fig. 4.27f and Fig. 4.27e we have enabled admission of the users to a later available resource such that a certain waiting is enforced before accessing the resolution channel. We observe that this improves the performance but as the number of resources are limited all of the Class 2 users cannot be served.

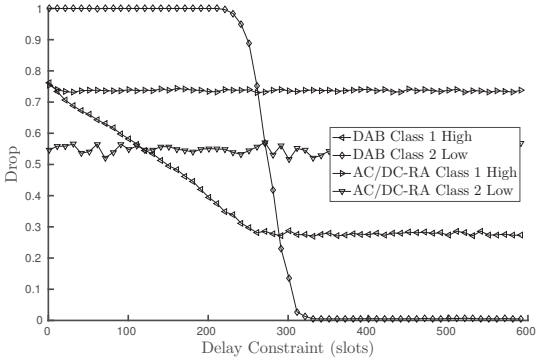
In this section we introduce a new system level protocol AC/DC-RA - Admission Control based Traffic-Agnostic Delay-Constrained Random Access. This protocol changes the random access paradigm with an addition of an *admission control decision*. The admission control decision is based on a novel user activity estimation and analytical modeling of the resolution. This estimation enables an accurate guess for the delay of a contention resolution and as



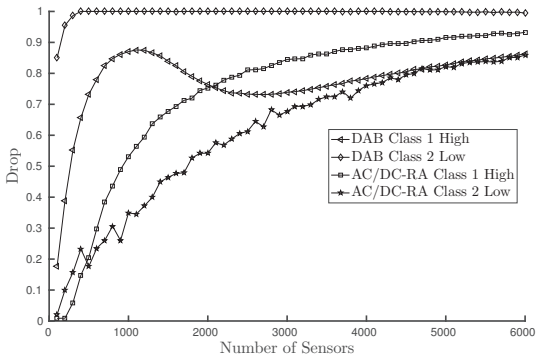
(a) Number of sensors for Class 2 $n_{\text{tothigh}} = 2000$ and for Class 1 $n_{\text{totlow}} = 200$.



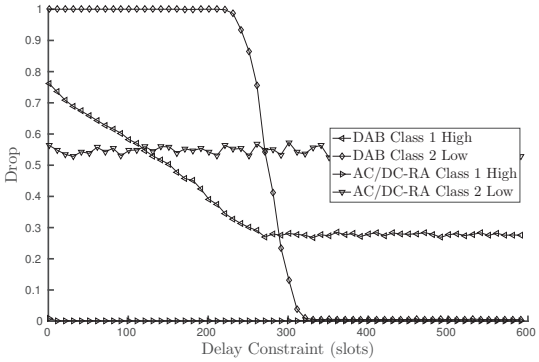
(b) Deadline L $= 10$, Class 1 has 10 % population of Class 2.



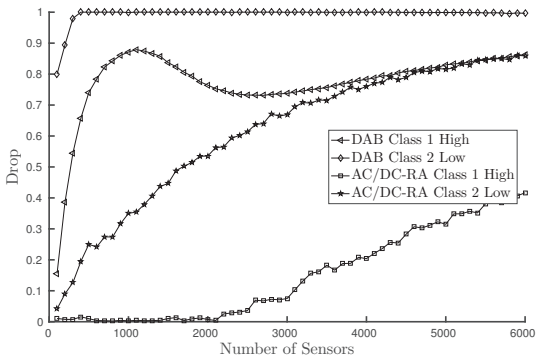
(c) Number of sensors for Class 1 $n_{\text{tothigh}} = 2000$ and for Class 2 $n_{\text{totlow}} = 200$.



(d) Deadline L $= 10$, Class 2 has 10 % population of Class 1.



(e) Number of sensors for Class 1 $n_{\text{tothigh}} = 2000$ and for Class 2 $n_{\text{totlow}} = 200$ with AC/DC-RA (buffering).



(f) Deadline L $= 10$, Class 2 has 10 % population of Class 1 with AC/DC-RA (buffering).

Figure 4.27: Comparison of AC/DC-RA with DAB varying the delay constraint L and fixing the total number of devices n_{tot}. The traffic imbalance is introduced the one class has the number of users depicted as on the x-axis (denoted as High) while the other class has 10 % of these users (denoted as Low).

demonstrated improves the resource efficiency with respect to algorithm that do not work with estimation.

4.5 Summary

In this chapter, we show that access protocols can double their efficiency if the distribution of the user activity is known. We later on showed that, given the distribution of the user activity, it is beneficial to deploy an on the fly estimation technique. Lastly, we have considered a system level integration of such an estimation through an admission control and we have given a practical example of an estimation algorithm. Finally, we have validated through simulation that such a system improves resource efficiency versus state of the art algorithms for delay constraints.

In the following chapter we focus on capabilities of the access algorithm in terms of latency-reliability constraints. We modify and analyze the access algorithms to provide stochastic and deterministic delay-constraints.

Chapter 5

Delay Constrained Reliable Access for Cellular Networks using Tree Algorithms

In the Chapter 4 we have shown that user activity estimation improves the radio resource efficiency. In this chapter we focus on the access algorithms. We investigate algorithmic improvements for access algorithms to support delay constraints. We assume the knowledge of user activity is available, and evaluate the performance of access algorithms under delay-constraints and inspect how they can be improved. Two main insights we provide is: (1) algorithms cannot be easily adapted to delay-constraints by adding more resources in the system and there is a performance loss. This loss is related to the reduction of feedback that would be normally sent after each time slot but reduced due to parallelization. Algorithmic adaptations are required to guarantee minimal loss. (2) the successive interference cancellation based tree algorithm has the potential to outperform scheduled access irrespective of the activity pattern.

The performance for access algorithms have been characterized via taking the expectation of the traffic model, giving the performance in terms of delay and throughput. However, with reliability and delay constraints, expectation is not sufficient any more. A realization based analysis is needed instead of the expectation.

The section is organized as follows: The Sec. 5.1 explains the need for an access algorithm with feedback, i.e. Tree Algorithms. The Sec. 5.2 introduces the background and the state of the art related to the Tree Algorithms. The Sec. 5.3 introduces a novel tree algorithm and up to authors best knowledge the first algorithm to guarantee full resolution capability to achieve delay constrained reliable access. This section is based on our work in [GAK17b]. The Sec. 5.5 considers a more complicated receiver capability, namely successive interference cancellation. The work extends the previous work [YG05] done for tree algorithms for interference cancellation and through use of identities of the users provides deterministic

delay bounds with increasing resource efficiency. This section is based on our work in [GGK19].

5.1 Motivation: Tree Resolution Algorithm the prominent candidate for Reliability and Delay constraints

Uplink access algorithms can be mainly separated in two categories depending on whether they use feedback or not. The main reason to avoid feedback in an access algorithm is that it requires downlink resources and failure in feedback can cause inconsistencies in the coordination of multiple users. The failure in feedback can be avoided with error detection methods and the effects can be limited if deadlocks are avoided. The downlink resource requirement is then the main reason of avoiding the feedback but can be compensated if the resource efficiency is improved on the uplink more than the feedback would consume on the downlink.

Information theoretic value of feedback in terms of exponential decrease in error probability have already been pronounced in the literature [Kra69] [GN10]. The multiple access channel model is used to evaluate the improvement through the intermittent feedback and these have been adapted to many practical schemes [SDG05], [SG10]. As required by the information theoretic framework, the results abstract many practical limitations away such as transmit to receive switching times of the hardware and are not directly applicable for radio resource management. Thus, gains of feedback for radio resource management (RRM) is not analyzed.

In RRM, binary acknowledgement (ACK) or non-acknowledgement (NACK) feedback is assumed. The feedback is used to inform the transmitter so the transmitter can declare success or repeat the transmission against channel failures. If resources are not dedicated but contended, same feedback informs the failure due to a contention [Abr70]. For design of contention algorithms, a static channel behavior with channel coding is assumed to avoid failures due to fading. The role of feedback in this case is limited to report contention failures.

The frequency of the feedback varies for contention algorithms. In the very first versions feedback is optionally included [Abr70]. This has been sufficient for low load cases. But caused problems for the packets that are lost due to contention. As the transmitter is unaware of the outcome the packets, it just transmitted the next packet. On later algorithms feedback after each contention slot is assumed [KL75], this increased the success rate as retransmissions are enabled if the packet is lost. However, the retransmissions increased the load and this increased the contention probability.

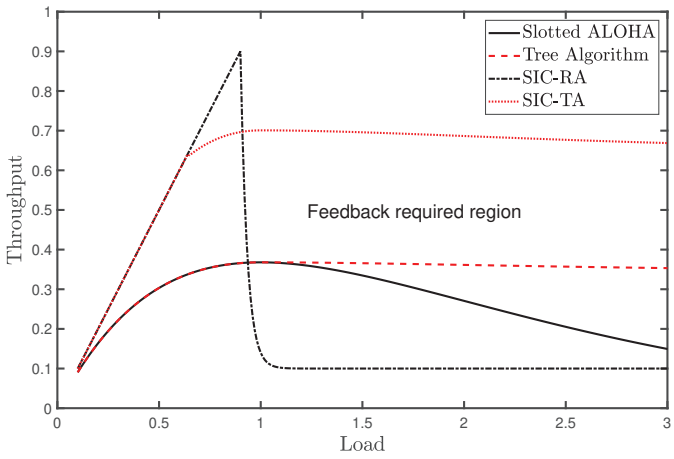


Figure 5.1: The throughput versus increasing load for algorithms with and without feedback.

So retransmissions are needed to increase the reliability, but too many retransmissions cause system to become unstable, since too many retransmissions meant that the average transmission per contention slot surpassed a certain value. So retransmissions should be limited to keep the average number of users per contention slot to an acceptable average value. Capetanakis [Cap79a] found the solution by setting different retransmission rules for the current contention and the previous contention with tree resolution algorithm. In tree algorithms, the current contenders are prioritized and on the next contention slot only the current contenders are allowed to re-transmit. The feedback, hibernates the other transmissions to a following contention slot until the current contention is resolved. Even though the delay still increased exponentially after a certain load, as the users just kept being postponed, the modification to the feedback guaranteed a stable contention behavior with increasing load. The reasoning behind the stable behavior is a *binary search like* logic, where n users are asked to select one of the two following contention slots until only 1 user selects the following contention slot. This guaranteed a success in finite number of contention slots, more precisely on average each $\log_2 n$ contention slots.

In Fig. 5.1 the throughput on y-axis versus load on x-axis is illustrated for different random access algorithms with and without feedback and also with and without successive interference cancellation. It is clear that in low load regions feedback does not have any effect for both cases. But in high load regions it guarantees that the throughput can be sustained. It is important here to mention as with increasing load an non-increasing throughput the delay increases.

Later works [Sch83b], [Bin00] have demonstrated that guaranteeing a success within finite number of slots is also possible through the use of estimation techniques. Capetanakis have previously shown in [Cap79a] that estimation also decreases the average contention slots

required to reach a success. However, with an inaccurate estimate the algorithm still reaches a success at worst on average each $\log_2 n$ contention slots. This is not true for other algorithms and stability is lost with errors on the estimation [Kos16].

In short, the stability, in terms of success in finite number of contention slots, of the tree algorithm is guaranteed even against load estimation errors as the *binary search like* logic acts as a built-in estimator. The structured feedback that acts as a built-in estimator makes the tree resolution algorithms the most prominent candidate for reliability and delay constraints. In the following sections we will use the structured properties to analytically define the worst-case performance of tree algorithms.

5.2 Background, State of the Art

The intuitive logic of the tree algorithms have been introduced in Chapter 4. In this Chapter a detailed background on tree algorithms is discussed.

5.2.1 Scenario

We consider a star topology where the central entity is called the gateway and leaf entities of the star are called devices. We consider an uplink scenario where only devices transmit a packet to the gateway. There are n_{tot} devices attached to the gateway. Considered resources in the system are slots of a single channel with a TDM scheme.

For the first section we assume a collision channel and on the second section we assume a perfect SIC channel model where perfect cancellation is possible if clean packets are received. These assumptions are common in MAC layer research to focus on a layer 2 based solution. Each device is synchronized perfectly to the slots defined by the TDM structure. The devices are randomly and sporadically activated and the number of active devices at any slot is n_{a}, such that $n_{\text{a}} < n_{\text{tot}}$. The devices have a homogeneous radio latency constraint L[1] and reliability constraint R. We investigate the multiple access problem of maximizing throughput that we abstract as maximizing number of successfully used slots.

5.2.2 Background on Tree Algorithms

Here we introduce a small discussion about the naming of the algorithm to give a historical lookout on tree algorithms. A reader that is only interested in the analysis can skip this part.

The naming of the contention tree algorithms has been a problem within the medium access community. Capetanakis called the algorithm, the *tree algorithm* [Cap79b]. This made

[1]For simplicity we assume that the constraint can be expressed in terms of slots.

distinguishing of the algorithm hard with other tree algorithms that are used in fields such as computer science. Gallager, to tree algorithms as the *splitting algorithm* [Gal78]. This name had a wider coverage in terms of multiple access while still having the problem of overlapping with other fields. Massey, in order to solve this problem, called it *Capetanakis Tybakov Mikhailov Collision Resolution Algorithm* [Mas81] with respect to the authors suggesting the algorithm. But this name did not get reused in the community. Afterwards, Huang named them *Interval Searching Algorithms* [HB85] while Kaplan used *Multiple Access Trees* [KG85] during the same era. The former was indeed too general for tree resolution while the latter did not encapsulate the resolution perspective. Finally, Jansen coined the term *contention tree algorithms* [JJ00] which we believe is reverting back to Massey with gracefully dropping the contributor names. We use the name Tree Algorithms for brevity in our work.

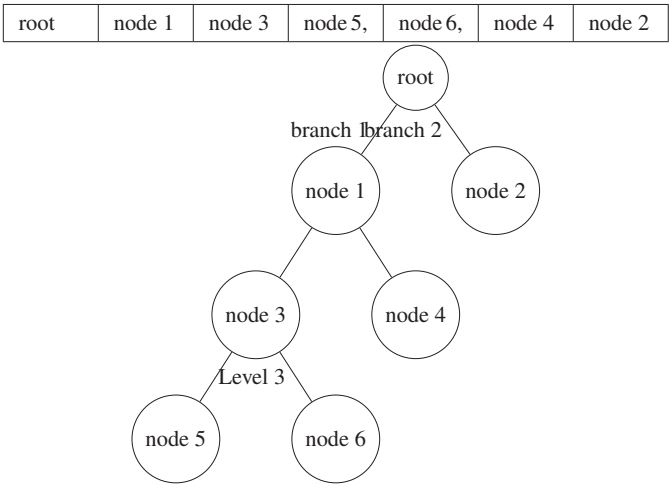


Figure 5.2: An exemplary binary tree

5.2.2.1 Preliminaries

At this point we introduce the terminology for graph theoretic trees and contention based access. The tree lexicon is important to visualize different steps of the algorithm. An examplary tree is illustrated in Fig. 5.2. The source of the tree is called the *root*. Each element in the tree except the first one is called a *node*. The maximum number of branches stemming from a *node* is called the branching factor and denoted as Q. A level of a node is the distance, i.e, the minimum number of nodes until the root of the tree and the level of the root is zero.

In contention tree algorithms, each collision except the initial collision stems from a previous collision. This dictates the tree analogy. The initial collision is referred to as the *root*. Each node in the tree will be also referred to as a *contention slot* as one is the more

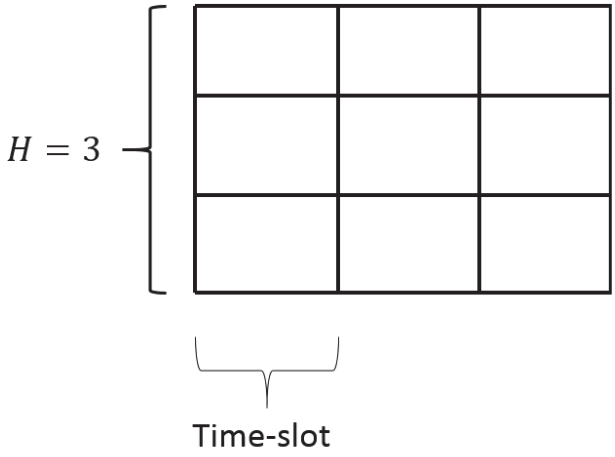


Figure 5.3: Time-slot and channels

specific name used for contention tree algorithm. Immediate children of the same contention slot are called *contention frames* as a group. A contention frame will contain at maximum Q contention slots.

The user will be referred to as contender. Channel is time-slotted. The physical layer behavior of a channel is abstracted with the collision channel. We assume immediate feedback from BS to all contenders. There are multiple channels that are available to be used in parallel. A bundle of H channels for a single time instance is called a time slot as illustrated in Fig. 5.3.

5.2.2.2 State of the art on Analysis for Adaptive Multichannel Contention Algorithms for Delay Constraints

The adaptivity of an access algorithm to delay constraints can be achieved in two ways: First, the physical layer algorithms can use extra bandwidth for each resource that enables k-multipacket reception (K-MPR) capability [GVS88] or the bandwidth can be used to build orthogonal channels that can be deployed by the medium access layer algorithm. Initially, we investigate the possibility to use orthogonal decoding techniques where the extra bandwidth is distributed in a medium access control fashion, leaving the physical layer exactly the same.

Multichannel access algorithms have been in use for a long time [BM82]. However, the delay constraint analysis for multichannel algorithms has only recently provoked interest. This is due to upcoming machine type communication. The initial solution has been to analyze the capability of current cellular networks. Currently deployed access algorithm for the cellular networks can be abstracted as a Multichannel Slotted ALOHA (MC-SA) system. The delay distribution given the number of users using MC-SA for varying number of channels is ana-

lyzed in [Jia+17b]. This work is extended in [Vil+18b] through shaping capabilities provided with Dynamic Access Barring with a stochastic network calculus framework. However, none of these algorithms can sustain high throughput with high reliability like tree algorithms. This motivates the use of multichannel tree algorithms for high reliability requirements. We first explain in detail the state of the art in the tree algorithms.

5.2.3 State of the art on Tree Algorithms

We summarize the types of tree algorithms under delay constraints and throughput scope. As the main contribution of this study is on multichannel aspects of tree algorithms, the state of the art on this aspect is summarized.

5.2.3.1 Multichannel Tree

A prior work for the multichannel use is investigated in [Cho85]. The author assumes a resource grid with H channels. The algorithm proceeds as a Q-ary tree algorithm in time perspective. On top of that at each time instance users select one of each H channels randomly. So even though they call it a Q-ary tree, effectively it is a $H \cdot Q$-ary tree as two splits are enforced after each collision. The authors have shown the algorithm decreases latency as well as the resource efficiency with increasing number of channels. The work in [Cho85] is the first work that enables H channel parallelization of the tree, however with the cost of changing the tree from a Q-ary tree to a $H \cdot Q$-ary tree. In our work, we enable a H channel parallelization without losing the high resource efficiency of tree algorithms.

A solution to parallelization of trees is allocating different channel resources for each collision in a centralized fashion. The central allocation is investigated in [Gür+17b] and [GAK17a]. Both works show that the problem becomes a scheduling problem of resolution resources, where the trade-off is decreasing the minimum latency with increasing maximum latency or vice-versa. The problem of allocating resources for multiple tree resolutions in a multichannel environment is demonstrated in our own work [Gür+17b], that is described in Chapter 4. Furthermore, different centralized resource allocation for multichannel tree algorithms is discussed in our own work [GAK17a]. Centralized algorithms comes with the cost of increased feedback message size. Even though these proposals showed improvement in terms of reliability, none of these provided a distributed algorithm and analytical evaluation of the performance. A distributed algorithm is necessary to avoid inconsistencies in an uplink access scenario.

In our algorithm we make use of a binary tree and parallelize it over H channels. This makes sure that the efficiency does not change with the number of channels deployed. It is well-known the efficiency is related to the selection of the branching factor Q [Mas81].

5.2.3.2 Delay Analysis

Delay is the time required to resolve a contender, i.e., number of *time slots* from the root of the tree to the successful slot of the contender. Each time slot is composed of H channels. The delay is bound to the number of **simultaneously usable channels** and the **exploration technique** of the tree. The exploration technique dictates that a node in the tree gets allocated a certain slot. The information of the total number of contenders n_a has been a common assumption in most of the previous work [JJ+00]. We believe that using this assumption extends the analysis open to many arrival distributions and we will also use this approach.

Single Channel ($H = 1$) In a *single channel* tree algorithms the delay maps to the total number of nodes (contention slots). For instance, the probability of a success at i^{th} contention slot maps to the probability of having a delay of i slots for a user in single channel tree algorithm. The probability mass function of delay in single channel tree algorithm can be found in [MP93b].

Q Channels ($H = Q$) In a Q *channel* tree algorithm a contention framecan be explored in parallel at the same time slot. In this case, the delay will map to the probability distribution of success in a contention frame. Allocating each contention frame to one time slot keeps the tree structure intact. Similar recursive analysis to the one used for a success in node can be used for contention frame. In [KG85] details for this analysis can be found.

Infinite Channels ($H \geq Q^m$) The case where the number of channels is larger than the number of contention slots in any level of the tree is refered as *infinite channel* case. In this case, all the contention slots in one level of the tree can be explored at the same time-slot. Thus, the probability of a success in m^{th} level of the tree can be used as delay of one user in the tree. The probability of success of a contender in level m conditioned on the initial number of contenders for Q-ary trees are given in [JJ+00]. The authors provide the probability that the tree terminates at level M conditioned on the initial number of contenders. Even though the use of infinite channels results in the minimum delay any Q-ary tree can achieve, it is not practical as the number of channels required for each time slot grows exponentially per level with respect to Q, i.e., Q^m at level m.

Arbitrary number of Channels In a practical scenario, the number of channels used by a tree algorithm may vary. The state of the art does not cover this scenario. This will be discussed in the Sec. 5.3.

5.2.4 Query Tree Algorithm

The query tree replaced the random decisions of the tree algorithm with the ID of the devices. The idea of limiting the size of the tree to introduce a maximum latency is initial introduced by Capetanakis [Cap79b]. As the tree algorithms are mostly used in RFID solutions, the practical implementation of this idea has been also in the scope of this community [LLS00]. Practically, the RFID related work exchanges the binary feedback with queries since RFID tags are powered with a query device. Theoretically, the same algorithm can be constructed with the binary feedback instead of a query.

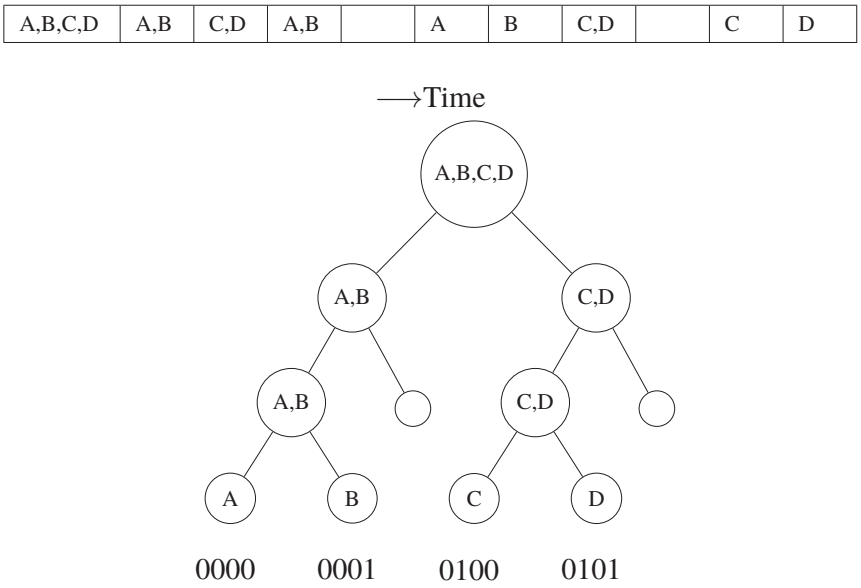
Query Tree Algorithm (QTA) is suggested in [CLL07]. In QTA every device has a unique id formed of u bits. This limits the total number of devices attached to the gateway to $n_{\text{tot}} = 2^u$. In QTA queries are used instead of feedback but the overhead is the same. In QTA devices are queried with respect to their id bits. The queries start with an empty query. A single bit is appended to the list of queries after each collision, starting from the left-most bit. For the next collision it appends the next bit. As each device has a unique id, this guarantees that two devices have a unique access decision in worst-case after u transmissions (if all previous $u-1$ bits are the same for two devices). The gateway implementation of QTA is given in Alg. 2, where the device implementation is only answering to the matching queries.

A detailed example is given for $M = 4$ in Fig. 5.4a. We have named the 4 devices as {A,B,C,D} with addresses {000,001,100,101} respectively. Each circle denotes a slot in the tree. The time-wise progression of the tree is given with slots above the tree. The address size, u is fixed to 3.

In the first slot, 4 devices transmit at the same time and collide. Next slot, the address $0xx$ is queried. Only, A and B transmit. It is again a collision. On the following slot, the query for address $1xx$ is also a collision so the algorithm moves one level down. The address $00x$ is queried and both devices transmit. The query for $01x$ result in an idle slot. Queries for address 001 and 000 is done on slot 6 and 7, respectively and both are successes. The right branch goes through a similar process.

5.2.5 Successive Interference Cancellation Tree Algorithm

The interference cancellation capability is initially introduced to a MAC layer algorithm and to tree algorithms in the work of Yu [YG05]. The inter-slot cancellation capability demonstrated



(a) QTA worst case with $M = 4$

Figure 5.4: Worst-case example for Query Tree Algorithm with $M = 4$. $u = 3$ is set such that maximum number of devices is $n_{\text{tot}} = 2^u = 8$.

that the throughput of random access algorithms is able to break the 0.5 limit reaching 0.69 packets per slot. Introducing the SIC capability for tree algorithms required the modification of the feedback. The feedback has to be extended from binary to k-ary. In case of an idle the feedback reported 0, in case of a collision it reported e and in case of a success it tried to cancel this success from previous slots and reported k. The k depicts how many packets are recovered in total including the success and the interference cancellation rounds.

5.3 Delay Constraint aware Multichannel Contention Tree Algorithm M-CTA

This section investigates a radio resource management problem for reliability and delay constraints, with the assumption that the number of users is known. The validity of this assumption have been investigated in Chap. 4. Up to our best knowledge there are only limited adaptive algorithms and respective analysis for **delay constraints** given the number of active users n_{a} and especially none for the tree algorithms. Important note is all the other algorithms have an error floor while the tree algorithms do not.

To this end the take out message for this section contribution is two-fold: First, we provide a multichannel contention tree algorithm, that can adapt the number of channels it deploys. Second, we provide a novel delay analysis for contention tree algorithms with arbitrary

number of channels. Through extensive simulations, we show that our analysis is valid and the algorithm can be used to adapt the number of channels to meet different requirements including delay.

The structure of this section is as follows. In subsection 5.3.2 we introduce our model and the analysis. In subsection 5.4 simulations are given to show that analytical assumptions match realizations.

5.3.1 Multichannel Contention Tree Algorithm (M-CTA)

The Multichannel Contention Tree Algorithm (M-CTA) is introduced here for dynamic adaptation to the delay requirements of the contenders. The algorithm makes use of up to H channels per time slot for timely resolution of contenders. This limitation is enforced such that the algorithm uses exactly the required number of channels.

Similar to the distributed contention tree algorithm that works with ternary feedback $\{0, 1, e\}$, this algorithm broadcasts a list of ternary feedback to inform the contenders about the outcome of H channels with a feedback list $\mathbf{f} = (f_1, f_2, \cdots, f_H)$ with $f_i \in \{0, 1, e\}$. The reception of the feedback does not require that contenders monitor all the channels as the concise[2] feedback can be received from a single broadcast channel. The contenders make distributed decision with the feedback list. This decision is enabled through keeping three separate counters at the contender side as summarized in Alg. 1. All the counters are updated via the feedback received from the channel and the Q-ary decisions of the contender at each level.

If the contender receives a collision feedback, it makes a Q-ary decision for level m, which is denoted as $b_m \in \{0, 1, \cdots, Q-1\}$. Thus, the list of all decisions up to and including level m is a list with $m+1$ elements $\mathbf{b} = (b_0, b_1, \cdots, b_m)$. Each contender keeps track of its decision.

The algorithm running on the contender side is given in Alg. 1 that is initialized for the values at level 0. At each time slot, contenders know which level of the tree is resolved through updating the m parameter. This is calculated after the feedback. The first collision observed in the channel denotes the level 0 of the tree. Only after the collision at level 0, an additional feedback to inform the contenders about the selected branching factor Q and parallel factor G is broadcast by the central entity. The number of total channels H can be calculated from these values as

$$H = Q \cdot G. \tag{5.1}$$

[2]Of course with increasing H the feedback will not be concise anymore. For instance, $H = 100$ results in 200 bits as 3 states are reported, and this is smaller than the expected payload of 32 bytes for machines.

Algorithm 1 The algorithm running on contender side for the multichannel tree algorithm.

Initialize $m = 0$, $t_m = 1$, $i = 1$, $q_m = 1$, $r_m = 0$.
while $f_{r_m}^{q_m,m} \neq 1$ **do**
  if $(i = q_m)$ **then**
    Transmit on channel r_m
  end if
  Receive and save feedback list $\mathbf{f}^{i,m}$.
  if $(m = 0)$ **then**
    Save G and Q
  end if
  $i \leftarrow i + 1$
  if $(i > t_m)$ **then**
    $m \leftarrow m + 1$, $i = 1$
    Update t_m, q_m, r_m
  end if
end while

The reason both values are separately announced is clarified later on. The number of time slots for each level is calculated via the feedback received from the previous level. The feedback list for time slot i and level m can be denoted as $\mathbf{f}^{i,m}$ and there are t_m such feedback lists, one for each time slot at level m. Thus, the number of time slots for level m can be calculated as,

$$t_m = \left\lceil \frac{\sum_{i=1}^{t_{m-1}} \sum_{j=1}^{H} \left[f_j^{i,m-1} = e \right]}{G} \right\rceil, \tag{5.2}$$

where $[\cdot]$ is the Iverson bracket, which returns 1 if the proposition is true. The Eq. (5.2) calculates the number of contention frames in level m from the feedback of all time slots at level $m-1$. The calculation is done by each contender such that the contender can transmit at the right channel. The contender calculates the $q_m{}^{\text{th}}$ time slot and $r_m{}^{\text{th}}$ channel for a transmission at level m of the tree via

$$q_m = \left\lceil \frac{z_m}{H} \right\rceil \text{ and with } r_m = z_m - (q_m - 1) \cdot H, \tag{5.3}$$

where z_m is the location of the contender in that level of the tree and it is given by

$$z_m = \sum_{i=1}^{m} Q^{m-i} \cdot b_i - s_m, \tag{5.4}$$

where b_i is the Q-ary decision of the contender at level i and s_m is the skipped number of slots at that level of the tree with successes and idle slots and is given by

$$s_m = \sum_{k=1}^{m} Q^{(m-k)} \cdot \left(\sum_{i=1}^{q_{m-1}-1} \sum_{j=1}^{H} \left[f_j^{i,m-1} \neq e \right] + \sum_{j=1}^{r_m - 1} \left[f_j^{q_{m-1},m-1} \neq e \right] \right) \text{ with } m > 0. \tag{5.5}$$

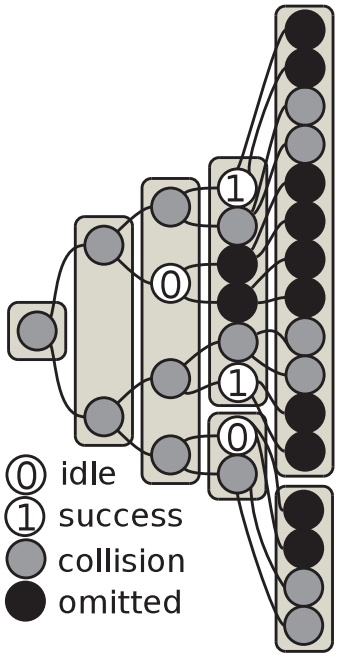


Figure 5.5: Diagram for the evolution of the tree depicted with omitted slots. After each success or idle following children slots are omitted. Using this information, a contendercan keep the structure of the full tree intact and only deduce the omitted slots to calculate on which time slotit should transmit.

These calculations may look like complicated operations but they are just counters. The detailed equations are shared for the sake of completeness. It is possible to update these counters level by level in a recursive way and the feedback lists does not need to be saved. Further optimization is an implementation issue and it will not be discussed further in this section.

In the following part of this work we consider a binary tree algorithm (BTA) thus, Q is set to 2. This means that as we may have up-to H channels per time slot we have $G = \frac{H}{2}$ contention frames per time slot. We will use G and H for multiple contention frames and channels respectively, for sake of readability.

Allocation of time slots for each level of the tree is given in Fig. 5.5. A contention tree structure with 4 levels is illustrated, where some of the slots are omitted. Each level is divided into time slots, where each time slot is composed of $H = 4$ channels. We see that level 3 and level 4 requires 2 time slots whereas level 0, 1 and 2 requires only 1. This figure points out that at maximum there are 2^m slots at each level of the tree as this is a binary tree. If a slot results in a success or an idle, no further children emerge from that slot. For instance, the idle at level 2 of the tree, would have had 2 children[3] at level 3, and 4 children at level 4. The number of omitted slots at any level of the tree can be calculated through the knowledge of a

[3]Virtual children, no real family is involved in the process.

success or idle. This information is used by the contenders to deduct how many time slots are required for each level and where exactly they would transmit.

Let us build a more specific example from Fig. 5.5. For illustrative purposes we assume that upper branch represents a selection of 0 and lower branch depicts a selection of 1. At first level of the tree some of the contenders have selected $b_0 = 0$ and the others $b_0 = 1$ as we have collisions in both slots. From the contenders that have selected $b_0 = 0$, none have selected $b_1 = 1$ as we have an idle on the lower slot branching out. But, they have selected all 0s that can be depicted as $\mathbf{b} = (b_0 = 0, b_1 = 0)$. Thus, the feedback list at time slot 1 for level 2 is $\mathbf{f}^{1,2} = (f_1 = e, f_2 = 0, f_3 = e, f_4 = e)$. This is broadcast directly after the outcome and the contenders are aware that they have to skip some slots on the following level. The contenders at the upper slot of the tree have $z_2 = 0$, with $b_0 = 0$ and $b_1 = 0$ as a result of their selections. This translates into $r_2 = 0$ for those contenders. Checking Eq. (5.5) we see that these contenders neglect all the feedback from the previous level as the summation is limited with $r_m - 1$, that is zero in this case. On the other hand, the contenders with $z_2 = 2$ have to calculate the number of skipped slots as $s_3 = 2$ with Eq. (5.5). We know that contenders with $z_2 = 2$ have selected $b_0 = 1$ and $b_1 = 0$ and if a contender has selected $b_2 = 1$, this results in $z_3 = 5 - s_3 = 3$ and $r_3 = 3$ due to the skipping. Thus, this contender transmits in time slot 1 of that level due to the skipped slots which would be time slot 2 without the skipping logic.

This concludes the algorithmic description of M-CTA. In the following sections, a complete analysis of the statistics of the number of time slots that are required to complete the tree is derived.

5.3.2 Analysis

In Table 5.1 the most relevant variables which are used in this section are presented.

The grouping of contention frames into time slots erases the recursive properties of tree algorithms. As a consequence, a recursive approach to obtain the length of the tree (in terms of time slots) is not an option. On the contrary, a level-wise approach such as the one presented in [KG85] will be the basis of the analysis. We first need the definition of the total and level-wise number of time slots.

Definition 5.3.2.0.1. Let $\mathcal{T}^N$ be the random variable modeling the number of time slots needed to complete a M-CTA given N contenders.

Definition 5.3.2.0.2. Let $\mathcal{T}_m^N$ be the random variable modeling the number of time slots at level m.

Table 5.1: Most relevant variables for computing $p_{\mathcal{T}^N}(t)$

Variable	Definition	Definition index
N	Number of initial contenders	-
G	Parallelization parameter	-
m	Level index	-
M	Number of considered levels	-
$\mathcal{T}^N$	Number of time slots	5.3.2.0.1
$\mathcal{T}_m^N$	Number of time slots at the level m	5.3.2.0.2
$\mathcal{X}_m^N$	Number of collisions at the level m	5.3.2.0.5
$\mathcal{K}_m^N$	Number of contenders at the level m	5.3.2.0.6
$\mathcal{Y}_\eta$	Number of children collisions of a parent collision with η contenders	5.3.2.0.12

From these two definitions, it follows that the total number of time slots in the M-CTA, or just *tree* for simplicity, can be expressed as:

$$\mathcal{T}^N = 1 + \sum_{m=1}^{\infty} \mathcal{T}_m^N, \tag{5.6}$$

with 1 added for the root of the tree. Our aim is to obtain the probability mass function (pmf) $p_{\mathcal{T}^N}(t)$ of the number of time slots in the tree, provided the number of contender N. In order to derive this pmf, we can use (5.6). But, we require the pmf of $\mathcal{T}_m^N$ to calculate the pmf of $\mathcal{T}^N$. $\mathcal{T}_m^N$ is a joint pmf. However, such a joint pmf of an infinite set of variables is too complex to deal with. Therefore, we have to set a limit for our variables such that the difference between the finite and the infinite result is negligible. With this in mind, we define a new, finite set of random variables with cardinality M.

Definition 5.3.2.0.3. Let $\hat{\mathfrak{T}}_M^N$ be the set of random variables $\mathcal{T}_m^N$ from $m = 1$ to $m = M$:

$$\hat{\mathfrak{T}}_M^N \triangleq \left\{ \mathcal{T}_m^N : m \in \mathbb{N} \ \wedge \ m \leq M \right\}. \tag{5.7}$$

The selection of M and its effects on the accuracy of the result are discussed in App. C.

We can now define the joint pmf of the variables in $\hat{\mathfrak{T}}_M^N$ as follows.

Definition 5.3.2.0.4. Let $p_{\mathcal{T}_1^N,...,\mathcal{T}_M^N}(t_1, ..., t_M)$ be the joint pmf of the variables in the set $\hat{\mathfrak{T}}_M^N$, that is:

$$p_{\mathcal{T}_1^N,...,\mathcal{T}_M^N}(t_1, ..., t_M) \triangleq \Pr\left\{ \mathcal{T}_1^N = t_1, ..., \mathcal{T}_M^N = t_m \right\} \tag{5.8}$$

$$= \Pr\left\{ \left\langle \mathcal{T}_1^N, ..., \mathcal{T}_M^N \right\rangle = \langle t_1, ..., t_M \rangle \right\}. \tag{5.9}$$

In (5.9), a vectorial notation was used instead of the standard notation, that is needed in the subsequent analysis.

All the statistical information of the number of time slots in the tree is contained in the joint pmf of $p_{\mathcal{T}_1^N,...,\mathcal{T}_M^N}(t_1,...,t_M)$. Therefore, if this joint pmf is known, $p_{\mathcal{T}^N}(t)$ can be calculated as in

$$p_{\mathcal{T}^N}(t) = \sum_{\mathfrak{S}} p_{\mathcal{T}_1^N,...,\mathcal{T}_M^N}(t_1,...,t_M), \tag{5.10}$$

where

$$\mathfrak{S} = \left\{ \langle t_1,...,t_M \rangle : \sum_{m=1}^{M} t_m = t-1 \right\} \tag{5.11}$$

is the set of vectors from $\hat{\mathfrak{T}}_M^N$ whose sum is $t-1$. The root is not considered in the calculation and this is where the -1 stems from. Each vector in $\mathfrak{S}$ represents the number of time slots in each level such that the total number of time slots in the tree is t. Hence, we just need to sum the probability of occurrence for all these vectors to obtain the probability of $\mathcal{T}^N = t$.

The next step is to derive an expression for $p_{\mathcal{T}_1^N,...,\mathcal{T}_M^N}(t_1,...,t_M)$ as a function of N and G. However, the derivation of this joint pmf is rather difficult, since we are facing the problem of finding out the relation among numerous variables that are all dependent from one another. Therefore, we use a Markovian re-definition that exploits the level-by-level expanding nature of the trees.

The number of time slots $\mathcal{T}_m^N$ in one level of the tree only depends on the partitioning of k contenders $\mathcal{P}^{k,m-1}$ in the previous level $m-1$ of the tree:

$$\Pr\left\{\mathcal{T}_m^N = t_m | \mathcal{T}_{m-1}^N = t_{m-1},\ldots,\mathcal{T}_1^N = t_1, \mathcal{P}^{k,m-1} = \pi^{k,m-1}\right\} \tag{5.12}$$

$$= \Pr\left\{\mathcal{T}_m^N = t_m | \mathcal{T}_{m-1}^N = t_{m-1}, \mathcal{P}^{k,m-1} = \pi^{k,m-1}\right\}. \tag{5.13}$$

By conditioning the number of time slots on the partitioning of contenders in the previous level we can re-write the Eq. (5.10) as multiplications of conditional probabilities:

$$p_{\mathcal{T}^N}(t) = \sum_{\mathfrak{S}} \sum_{\mathfrak{P}} p_{\mathfrak{P}}(\mathcal{P}) p_{\mathcal{T}_1^N}(t_1) \prod_{m=2}^{M} p_{\mathcal{T}_m^N|\mathcal{T}_{m-1}^N,\mathcal{P}^{k,m-1}}(t_m, t_{m-1}, \pi^{k,m-1}) \tag{5.14}$$

where $\mathfrak{P}$ is the set of constrained vectors due to causality relations between partitioning of contenders that will be detailed later and

$$p_{\mathfrak{P}}(\mathcal{P}) = p_{\mathcal{P}^{k,1},...,\mathcal{P}^{k,M-1}}(\pi^{k,1},...,\pi^{k,M-1}) \tag{5.15}$$

is the joint probability distribution for the partitioning of contenders in each level of the tree. We re-write it with the number of contenders at each level independently:

$$p_{\mathfrak{P}}(\mathcal{P}) = \prod_{m=1}^{M-1} p_{\mathcal{P}^{k,m}}(\pi^{k,m}). \tag{5.16}$$

However, this is only valid as long as the selected partitions obeys the *causality constraints* which will be introduced later on. This can be controlled via the set of vectors that the sum is taken over. However, to simplify the formula we group the two products by adjusting the variable m. Taking advantage of this simplification and using a non-constrained sum in order to apply the law of total probability we re-write the Eq. (5.14) as

$$p_{\mathcal{T}^N}(t) = \sum_{\mathfrak{S}} \sum_{\mathfrak{P}} p_{\mathcal{T}_1^N}(t_1) \prod_{m=2}^{M} p_{\mathcal{P}^{k,m-1}}(\pi^{k,m-1}) p_{\mathcal{T}_m^N|\mathcal{T}_{m-1}^N,\mathcal{P}^{k,m-1}}(t_m, t_{m-1}, \pi^{k,m-1}) \tag{5.17}$$

$$\cong \sum_{\mathfrak{S}} p_{\mathcal{T}_1^N}(t_1) \prod_{m=2}^{M} p_{\mathcal{T}_m^N|\mathcal{T}_{m-1}^N}(t_m, t_{m-1}). \tag{5.18}$$

This is an approximation as the sum is not taken over all possible partitions but over only a subset of these partitions that fulfill the *causality constraints*. This is a rough approximation as the *causality constraints* are not enforced strictly but stochastically. But, the each time-slot probability is conditioned on the number of time slots in the previous level and this provides a good fit as demonstrated in the Sec. 5.4 with previous analytical results and simulations. This simplification enables less complex and precise results for tree algorithms to be used for high reliability applications.

This simplification implies that the number of time slots in a given level is only influenced by the number of time slots in the previous level such that tree depicts Markovian properties. This assumption holds perfectly for all levels of the tree until a success or idle occurs in the previous level, as the partitioning of contenders is not constrained due to causality constraints. This interesting result is pointed out in the later parts of the paper. This means that the tree structure is not crucial for analysis of delay of the contention tree resolutions, up to the level where the first success or idle node is observed.

With this simplification we focus on the derivation of the conditional pmfs of the number of time slots at any level of the tree, provided the number of time slots at the previous level. We will tackle this problem by analyzing first the number of *collisions* (the number of nodes with more than one contender) at each level. The number of collisions at a certain level can be easily translated into the number of time slots at the next level, as it will be shown. But first, we need to define a variable to model the number of collisions.

Definition 5.3.2.0.5. Let $\mathcal{X}_m^N$ be the random variable modeling the number of collisions within the level m, provided N initial contenders.

The conditional probability $p_{\mathcal{T}_m^N|\mathcal{T}_{m-1}^N}(t_m|t_{m-1})$ of obtaining t_m time slots at the level m, provided t_{m-1} time slots at the level $m-1$ can be expressed as:

$$p_{\mathcal{T}_m^N|\mathcal{T}_{m-1}^N}(t_m|t_{m-1})=\begin{cases}\sum_{i=0}^{G-1} p_{\mathcal{X}_{m-1}^N|\mathcal{X}_{m-2}^N}\left(0 \,\middle|\, G\cdot t_{m-1}-i\right) & t_m=0,\\ \sum_{i=0}^{G-1}\sum_{j=0}^{G-1} p_{\mathcal{X}_{m-1}^N|\mathcal{X}_{m-2}^N}\left(G\cdot t_m-j \,\middle|\, G\cdot t_{m-1}-i\right) & t_m>0,\end{cases} \tag{5.19}$$

where $p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N}(x_m|x_{m-1})$ is the conditional probability of obtaining x_m collisions at the level m, provided x_{m-1} collisions at the level $m-1$. We know that every collision at one level produces two new nodes at the next level due to binary tree structure. One time slot contains $2G$ nodes. Thus, we can convert collisions to time slots as follows:

$$\mathcal{T}_m^N = \left\lceil \frac{\mathcal{X}_{m-1}^N}{G} \right\rceil. \tag{5.20}$$

Owing to the presence of the ceiling function, the relation is not bijective, but several values of $\mathcal{X}_{m-1}^N$ map to the same value of $\mathcal{T}_m^N$. Indeed, given $\mathcal{T}_m^N = t_m$ and $\mathcal{X}_{m-1}^N = x_{m-1}$, any x_{m-1} in the set $\{G\cdot t_m - i \;:\; 0\le i\le G-1\}$ fulfills (5.20). Hence, the conversion between the marginal probability $p_{\mathcal{T}_m^N}(t_m)$ of obtaining t_m time slots at the level m and the marginal probability $p_{\mathcal{X}_{m-1}^N}(x_{m-1})$ of obtaining x_{m-1} collisions at the level $m-1$ is just:

$$p_{\mathcal{T}_m^N}(t_m) = \begin{cases} p_{\mathcal{X}_{m-1}^N}(0) & t_m=0,\\ \sum_{i=0}^{G-1} p_{\mathcal{X}_{m-1}^N}\left(G\cdot t_m - i\right) & t_m>0.\end{cases} \tag{5.21}$$

Hence, in order to deduce the relation between $p_{\mathcal{T}_m|\mathcal{T}_{m-1}}(t_m|t_{m-1})$ and $p_{\mathcal{X}_m|\mathcal{X}_{m-1}}(x_m|x_{m-1})$, exactly the same procedure needs to be applied, but this time with two variables instead of one which proves Eq. (5.19).

Provided Eq. (5.19), the problem is to find an expression for $p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N}(x_m|x_{m-1})$. In order to calculate this pmf, we need to know the number of contenders in each of the x_{m-1} collisions of the level $m-1$. Thus, we need to consider every different possibility and then apply the law of total probability. In order to do so, we define a variable for the number of collided contenders at each level.

Definition 5.3.2.0.6. Let $\mathcal{K}_m^N$ be the random variable modeling the total number of contenders which have been involved in collisions at the level m, i.e. the number of collided contenders at the level m, provided N initial contenders.

We are interested in the statistical properties of the distribution of contenders over nodes in the tree. We can directly transform our *contenders-into-nodes* problem into an equivalent

balls-into-bins problem. This simplifies the understanding of the problem and allows us to use existing solutions from the literature.

The next two lemmas deal with the number of ways to distribute balls into bins such that some condition about the number or size of collisions is fulfilled. The results will be useful for subsequent lemmas.

Lemma 5.3.2.0.1. *The number of ways $\Psi^u_{m,n}$ to arrange u balls into m bins such that $n \leq m$ of them have more than one ball can be obtained by means of the recursion*

$$\Psi^u_{m,n} = n\Psi^{u-1}_{m,n} + (m-n+1)\,\Psi^{u-1}_{m,n-1} + \Psi^{u-1}_{m-1,n}, \tag{5.22}$$

with initial conditions $\Psi^u_{u,0} = 1$, $\Psi^1_{1,0} = 1$, $\Psi^1_{1,1} = 0$, and $\Psi^{u>1}_{1,1} = 1$.

Proof. The derivation of this recursion can be found in [WBC15b]. □

Lemma 5.3.2.0.2. *The number of ways $\Gamma^{u,v}_{m,n}$ to arrange u balls into m bins such that $n \leq m$ bins have more than one ball and that the total number of balls occupying those n bins is $v \leq u$, can be computed as:*

$$\Gamma^{u,v}_{m,n} = \Psi^u_{n+u-v,n}\binom{m}{n+u-v}(n+u-v)!, \tag{5.23}$$

where $\Psi^u_{m,n}$ was given in Lemma 5.3.2.0.1.

Proof. There are u balls, v of which are with more than one ball in their bins. This implies that $u - v$ balls are alone in their bins. Therefore, we have a total of

$$o = n + u - v. \tag{5.24}$$

occupied bins, n with more than one ball and $u - v$ with single balls. Knowing this, we can compute the number of ways to arrange u balls into o bins such that v of them have more than one ball, as given in Lemma 5.3.2.0.1. As each partition must have at least one contender, by distributing the u contenders over o bins and n bins with more than one ball we force the the number of contenders in those bins to add up to v. Finally, we just need to compute the number of ways to choose o bins out of m possible bins —$\binom{m}{o}$— and the number of ways to arrange those bins —$o!$—. As a result, our final expression is:

$$\Gamma^{u,v}_{m,n} = \Psi^u_{o,v}\binom{m}{o}o! \tag{5.25}$$

After combining (5.24) and (5.25), we obtain (5.23). □

The number of bins with more than one balls n map to number of bins with collisions x_m, the total number of bins m map to the number of nodes in one level of the tree 2^m, the total number of balls in bins with more than one ball v maps to the collided contenders k_m and the total number of balls u maps to the number of users N. The probability of occurrence for $\mathcal{K}_m^N = k_m$ collided contenders given there are $\mathcal{X}_m^N = x_m$ collisions at the level m is

$$p_{\mathcal{K}_m^N|\mathcal{X}_m^N}(k_m|x_m) = \frac{\Gamma_{2^m,x_m}^{N,k_m}}{\sum_{j=0}^{N} \Gamma_{2^m,x_m}^{N,j}}, \tag{5.26}$$

where the number of ways to have k_m collided contenders with x_m collisions is divided with the number of ways to have any number of collided contenders with x_m collisions.

The probability of x_m collisions at level m, given x_{m-1} collisions and k_{m-1} contenders at level $m-1$ is required for $p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N}(x_m|x_{m-1})$ introduced in Eq. (5.19). We will treat this probability with a number theory perspective, through which we break this problem into integer partitions as illustrated in the following example. These partitions will force our distribution to obey the *causality constraints* stochastically.

Example 5.3.2.0.1. Let us consider a scenario where $x_{m-1} = 4$ and $k_{m-1} = 12$. There are five different ways to decompose 12 contenders into 4 collisions, which are the five different partitions of 12 in 4 parts, such that every part is greater than one. Namely, these partitions are:

$$(2,2,2,6)\ \ (2,2,3,5)\ \ (2,2,4,4)\ \ (2,3,3,4)\ \ (3,3,3,3).$$

At this point, it is easy to see why it is interesting to decompose k_{m-1} into partitions. Given a certain partition of contenders at the level $m-1$, say $(2,2,3,5)$, it is immediate to compute the probability of x_m collisions at the level m. Given the number of contenders it is trivial to calculate the probability of collisions that these contenders generate. We only need to compute the probability of generating 0, 1 or 2 new collisions for each collision (i.e., for every part of the partition), which is now simple since we know the number of contenders in each one. We first need to compute the probability of each partition to appear.

Definition 5.3.2.0.7. Let $\mathfrak{P}^{k,x}$ be the set of partitions of k in x parts greater than 1. An element of $\mathfrak{P}^{k,x}$ is a partition $\pi_i^{k,s}$, such that:

$$\mathfrak{P}^{k,x} \triangleq \left\{ \pi_i^{k,x} : i \in \{1, ..., \Pi(k,x)\} \right\}, \tag{5.27}$$

where $\Pi(k,x)$ is the number of partitions of k in x parts greater than 1.

Definition 5.3.2.0.8. Let $\mathcal{P}^{k,x}$ be the random variable modeling the process of randomly selecting a partition out of set $\mathfrak{P}^{k,x}$. That partition represents the distribution of the collided balls after an uniformly random allocation of N balls into m bins.

Definition 5.3.2.0.9. Let $p_{\mathcal{P}^{k,x}}(\pi_i^{k,x})$ be the pmf of $\mathcal{P}^{k,x}$:

$$p_{\mathcal{P}^{k,x}}(\pi_i^{k,x}) \triangleq \Pr\left\{\mathcal{P}^{k,x} = \pi_i^{k,x} \mid \mathcal{X}_m^N = x, \mathcal{K}_m^N = k\right\}, \tag{5.28}$$

which represents the probability of the event of selecting the partition $\pi_i^{k,x}$ from the set $\mathfrak{P}^{k,x}$.

Definition 5.3.2.0.10. Let $\eta_{i,j}^{k,x}$ be a part of the partition $\pi_i^{k,x}$, for $j \in \{1, \ldots, x\}$. The following relations hold:

$$\eta_{i,j}^{k,x} > 1. \tag{5.29}$$

Each part is bigger than one as each part represents a collision. The summation of all collisions represent the total number of contenders at that level

$$\sum_{j=1}^{x} \eta_{i,j}^{k,x} = k. \tag{5.30}$$

Each partition can be written as a list of parts,

$$\pi_i^{k,x} = \left\langle \eta_{i,1}^{k,x}, \eta_{i,2}^{k,x}, \ldots, \eta_{i,x}^{k,x} \right\rangle. \tag{5.31}$$

Definition 5.3.2.0.11. Let $\#_{i,a}^{k,x}$ be the number of occurrences of the number a within the partition $\pi_i^{k,x}$. We can formally define this new variable as follows:

$$\#_{i,a}^{k,x} \triangleq \sum_{j=1}^{x} \left[\eta_{i,j}^{k,x} = a\right], \tag{5.32}$$

where $[\cdot]$ is the Iverson bracket, which returns 1 if the proposition inside is true.

With these definitions, we can compute the probability $p_{\mathcal{P}^{k,x}}(\pi_i^{k,x})$ as follows.

Lemma 5.3.2.0.3. *The probability to have the specific partitioning $\pi_i^{k,x}$ given k contenders and x collisions is*

$$p_{\mathcal{P}^{k,x}}(\pi_i^{k,x}) = \frac{k!}{\Psi_{x,x}^k} \prod_{j=1}^{x} \frac{1}{\eta_{i,j}^{k,x}! \cdot \#_{i,a}^{k,x}!}. \tag{5.33}$$

Proof. The derivation of (5.33) is explained in the Appendix C.2. □

We introduce an expression for the probability of generating certain number of *children* collisions, provided that we know the size of the *parent* collision.

Definition 5.3.2.0.12. Let $\mathcal{Y}_\eta$ be the random variable modeling number of child collisions of a parent collision of η contenders. Since we are analyzing a Binary Tree Algorithm, the sample space of $\mathcal{Y}_\eta$ is simply $\{0, 1, 2\}$, i.e., at most two collisions can be children of one parent collision.

Lemma 5.3.2.0.4. *The probability $p_{\mathcal{Y}_\eta}(y_\eta)$ of generating y_η children collisions, provided a parent collision of size η is:*

$$p_{\mathcal{Y}_2}(y_2) = \begin{cases} \frac{1}{2} & y_2 = 0 \\ \frac{1}{2} & y_2 = 1 \\ 0 & y_2 = 2 \end{cases} \qquad \text{If } \eta = 2, \tag{5.34}$$

$$p_{\mathcal{Y}_\eta}(y_\eta) = \begin{cases} 0 & y_\eta = 0 \\ (\eta+1)\cdot\left(\frac{1}{2}\right)^{\eta-1} & y_\eta = 1 \\ 1-(\eta+1)\cdot\left(\frac{1}{2}\right)^{\eta-1} & y_\eta = 2 \end{cases} \qquad \text{If } \eta > 2. \tag{5.35}$$

As last step, we need to stitch together all the results that we have obtained in order to get a closed-form expression for $p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N}(x_m|x_{m-1})$. The following three lemmas build upon the previous lemmas and yield such an expression.

Lemma 5.3.2.0.5. *The probability of having x_m collisions at the level m, provided x_{m-1} collisions and k_{m-1} contenders partitioned in $\pi_i^{k_{m-1},x_{m-1}}$ at the level $m-1$ is*

$$p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N,\mathcal{K}_{m-1}^N,\mathcal{P}^{k_{m-1},x_{m-1}}}(x_m|x_{m-1},k_{m-1},\pi_i^{k_{m-1},x_{m-1}}) = \\ = p_{\mathcal{Y}_{\eta_{i,1}^{k,x}}}\left(y_{\eta_{i,1}^{k,x}}\right) * \ldots * p_{\mathcal{Y}_{\eta_{i,x_{m-1}}^{k,x}}}\left(y_{\eta_{i,x_{m-1}}^{k,x}}\right), \tag{5.36}$$

where $*$ denotes the discrete convolution.

Proof. Given a certain distribution (partition) of contenders, we can use $p_{\mathcal{Y}_\eta}(y_\eta)$ to compute the probability that some collision (part) at the level $m-1$ generates 0, 1 or 2 collisions at the level m. Furthermore, since the subtrees generated by the parent collisions are not related, variables $\mathcal{Y}_\eta$ are independent from one another. Therefore, we can compute the pmf of the sum of all $\mathcal{Y}_\eta$ as the discrete convolution of all of them. □

Lemma 5.3.2.0.6. *The probability to have x_m collisions at the level m, given x_{m-1} collisions and k_{m-1} contenders at the level $m-1$ is*

$$p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N,\mathcal{K}_{m-1}^N}(x_m|x_{m-1},k_{m-1}) = \\ = \sum_{\pi\in\mathfrak{P}^{x_{m-1},k_{m-1}}} p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N,\mathcal{K}_{m-1}^N,\mathcal{P}^{k_{m-1},x_{m-1}}}(x_m|x_{m-1},k_{m-1},\pi)\cdot p_{\mathcal{P}^{k_{m-1},x_{m-1}}}(\pi). \tag{5.37}$$

Proof. This lemma is just an application of the law of total probability combining the expressions of Lemma 5.3.2.0.3 and Lemma 5.3.2.0.5. □

Lemma 5.3.2.0.7. *The probability to have x_m collisions at the level m given x_{m-1} collisions at the level $m-1$ and N initial contenders is*

$$p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N}(x_m|x_{m-1}) = \sum_{k_{m-1}=0}^{N} p_{\mathcal{X}_m^N|\mathcal{X}_{m-1}^N,\mathcal{K}_{m-1}^N}(x_m|x_{m-1},k_{m-1}) \cdot p_{\mathcal{K}_{m-1}^N|\mathcal{X}_{m-1}^N}(k_{m-1}|x_{m-1}). \tag{5.38}$$

Proof. This is again a direct application of the law of total probability that combines the expressions of Lemma 5.3.2.0.6 and Eq. (5.26). □

The result of Lemma 5.3.2.0.7 is quite important as it represents probability to have so many number of collisions at level m given the number of collisions at level $m-1$ for any binary tree. Future work that wants to apply any specific modification on the tree algorithm can make use of this result. This is also what we do here for delay constrained access that result in M-CTA. The grouping of collisions in enforced with variable g as in Eq.(5.20). Finally, we have all the required ingredients to write down a close-form expression for the pmf of $\mathcal{T}^N$, which is shown in the following theorem.

Theorem 5.3.2.0.1. *The probability of tree successfully completing with t time slots before level M given N initial contenders is*

$$p_{\mathcal{T}^N}(t) \cong \sum_{\mathfrak{S}} p_{\mathcal{T}_1^N}(t_1) \prod_{m=2}^{M} p_{\mathcal{T}_m^N|\mathcal{T}_{m-1}^N}(t_m,t_{m-1}) \tag{5.39}$$

Proof. Plugging the Eq. (5.19) in (5.18), then using Lemma 5.3.2.0.7, we conclude the proof. □

5.4 Evaluation

In this section, we compare our analysis to previous results and simulations to provide convincing outcomes that approves the use of Markovian approximation.

5.4.1 Analytical Evaluation

The suggested algorithm provides the flexible parallelization of the resolution. As there are already algorithm with fixed parallelization, we can use the analytical results of these algorithms to compare our analysis. In this part, we will take two algorithms to compare against, first one is the branch based parallelization such that $G = 1$ and the second one is the

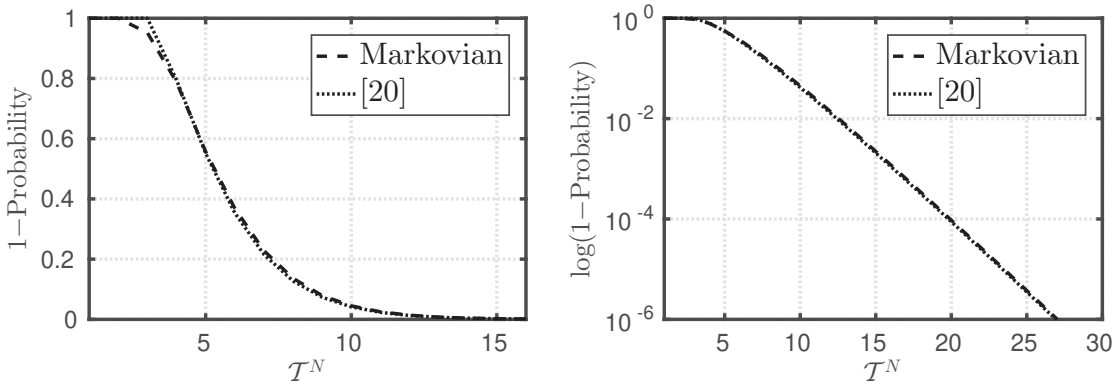


Figure 5.6: Analytical results of the inverted cumulative mass functions of the number of time slots until resolution of all contenders in a single channel tree for $G = 1$, given $N = 5$ initial contenders in log and linear. The inverted cumulative mass function represents the probability that a tree is not completely resolved with $\mathcal{T}^N$ time slots.

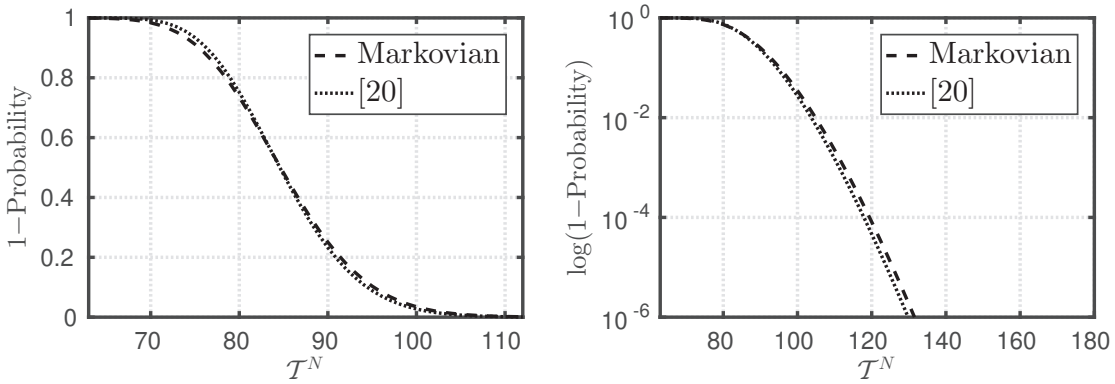


Figure 5.7: Analytical results of the inverted cumulative mass functions of the number of time slots until resolution of all contenders in a single channel tree for $G = 1$, given $N = 60$ initial contenders in log and linear. The inverted cumulative mass function represents the probability that a tree is not completely resolved with $\mathcal{T}^N$ time slots.

full parallelization with $G = \infty$. As we cannot have infinite parallelization we will select a really high G value such that the tree is always parallelized.

In Fig. 5.6 the branch parallelized tree analysis in [JJ+00], which is presumed to be exact, is compared to our analysis. The x-axis depicts the number of time slots until completion of the tree and the y-axis depicts the probability mass. The analyses are compared with $N = 5$ contenders. The values are compared in linear and logarithmic plot for the inverted CMF. The linear plot shows that the CMF obtained from the Markovian approach has up to 2.5% difference with low number of time slots and converges to the same result with increasing $\mathcal{T}$. This behavior shows the limitation of Markovian assumption. The assumption we used for analysis is the following, the number of collisions in one level of the tree is strictly depicted by the number of collisions in the previous level of the tree disregarding the number of

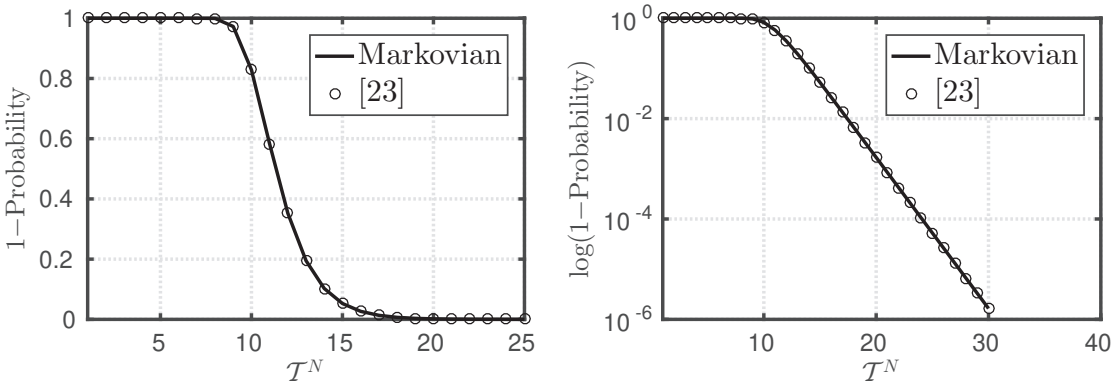


Figure 5.8: Analytical results of the inverted cumulative mass functions of the number of time slots until resolution of all contenders in a tree for $G = \infty$, given $N = 60$ initial contenders in log and linear. The inverted cumulative mass function represents the probability that a tree is not completely resolved with $\mathcal{T}^N$ time slots.

contenders. The assumption would be true if we also added the information of partitioning of contenders. As this is hard to realize in a practical system we did not consider this. We tried to obtain that information from the number of collisions in the previous level of the tree. So the partitioning of contenders is estimated. This estimation is more precise with multiple levels of the tree this we see that results converge with high number of time slots. However, if a less likely outcome occurs, in terms of partitioning of contenders, the Markovian assumption has a false estimation for partitioning of contenders and does not foresee a fast termination of the resolution. The fit of the analysis is emphasized with the logarithmic plot for higher precision levels.

Increasing number of contenders will decrease the probability of an early termination of the tree and lead to a better fit of the CMF. In order to point this out we have plotted the same comparison with $N = 60$ contenders in Fig. 5.7. Here, we see that the big difference for low number of time slots have decreased to less than 0.5% and follows the shape better. On the other hand there is slight mismatch for high number of time slots since some partitioning cases, have become highly probable with changing number of users and we see a slight difference. However, the difference is indeed small as emphasized with the logarithmic plot.

In Fig. 5.8 the full parallelized tree analysis in [KG85] is compared to our analysis. The x-axis depicts the number of time slots until completion of the tree and the y-axis depicts the probability mass. The CMF obtained from the Markovian approach seem to match the analysis in [KG85] perfectly. However, in order to focus on corner cases we take the logarithm of the inverted CMF to see that the exact match is valid also for higher order of precision. This leads us to the projection that the Markovian assumption holds perfectly with increasing

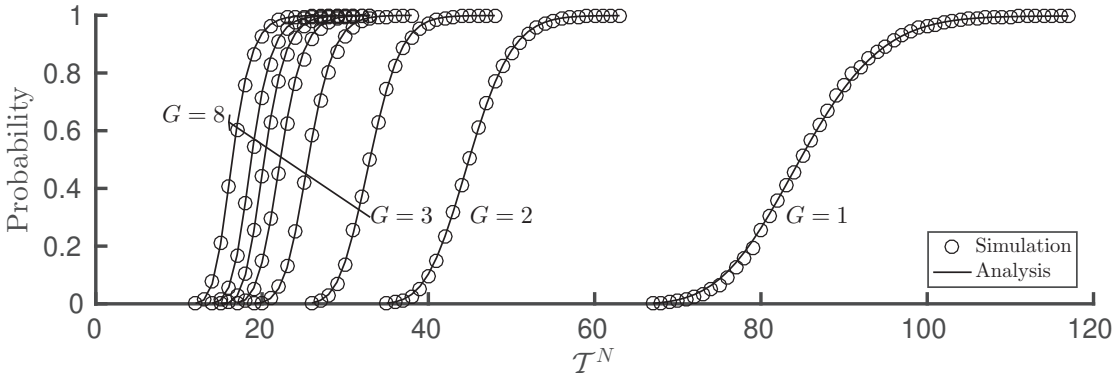


Figure 5.9: Analytical and simulative results of the cumulative mass functions of the number of time slots until resolution of all contenders in a multichannel tree for several values of G, given $N = 60$ initial contenders.

G. As the algorithm is designed for delay constrained applications it will be mostly used with high parallelized ($G > 1$) cases and use of the approximation is validated.

5.4.2 Simulation results

In order to check the accuracy of the model, simulations were performed and their results were compared with the predicted values. The simulator was written in MATLAB, and the selected parameters were $N = \{5, 60\}$ contenders and 10^5 runs for each value of N.

In the Fig. 5.9, the theoretical and the empirical CMFs of the number of time slots for the completion of the tree for $G = \{1, ..., 8\}$ are plotted together for comparison. The x-axis depicts the number of time slots and the y-axis denotes the probabilities. For $G = 1$, we see a slight but noticeable difference between the model and the actual results, as a consequence of the model. Nevertheless, this difference is rather small and the accuracy of the analytical model improves rapidly when G increases.

In the Fig. 5.10, the theoretical and the empirical inverted CMFs of the number of time slots until resolution of all contender for $G = \{1, ..., 8\}$ are plotted again but in log-scale to emphasize the match. The y-axis is limited to 10^{-3} due to the resolution of the simulations. For $G = 1$, the predicted and the actual result differ slightly but the difference is the same on each log-scale. For the remaining values of G, it can be observed that the model becomes more accurate when G increases. Thus, we conclude that the Markovian approximation is valid for reliabilities up to 10^{-3}. In the Fig. 5.11, we have also plotted the inverted mass function for $N = 5$ and we see the validity of the fit compared to simulations even with decreasing number of contenders.

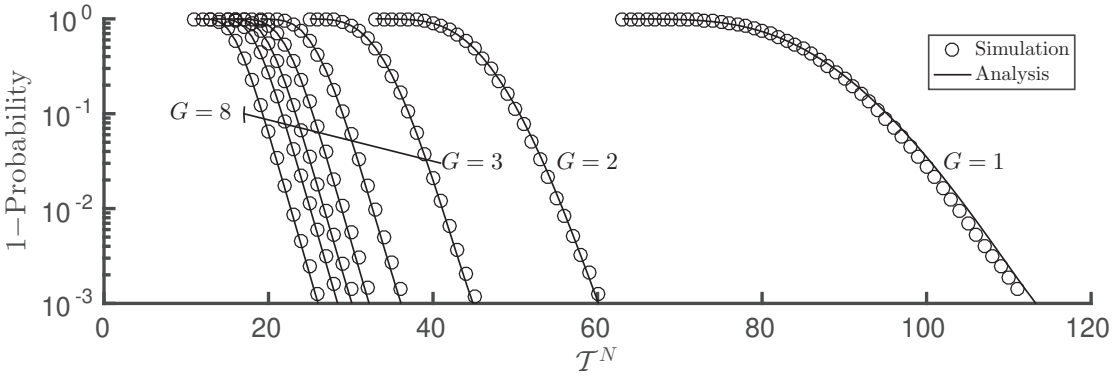


Figure 5.10: Analytical and simulative results of the inverted cumulative mass functions of the number of time slots until resolution of all contenders in a multichannel tree for several values of $G = \{1, 2, \cdots, 8\}$, given $N = 60$ initial contenders in log-scale.

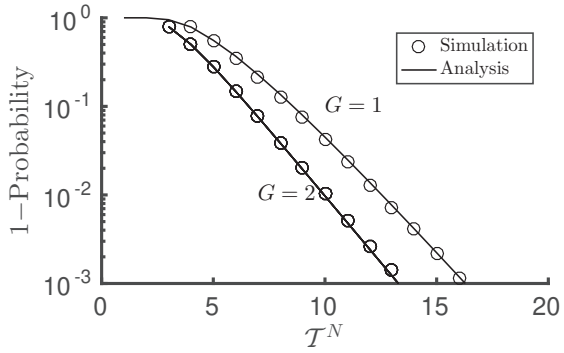


Figure 5.11: Analytical and simulative results of the inverted cumulative mass functions of the number of time slots until resolution of all contenders in a multichannel tree for several values of $G = \{1, 2, \cdots, 8\}$, given $N = 5$ initial contenders in log-scale.

Apart from the validity of the model, conclusions about the values of the access delay may be drawn as well, now that the predicted values are backed with simulations. Regarding access delay, we see how the maximum access delay for $G = 7$ might be lowered up to a 10% of the delay of a single channel Tree Algorithm, which is obtained after multiplying by two the result for $G = 1$. Hence, a tenfold reduction of the access delay can be achieved if $G = 7$, and larger reductions are possible is $G \geq 7$. Nevertheless, the higher G the lower the number of trees that can be executed in parallel if the number of channels is limited, therefore the optimum value of G needs to be carefully chosen depending on the application.

It is shown that the delay constraints can be reacted to for different levels of stochastic guarantees by adjusting the number of channels allocated to the tree algorithms. As discussed in Chapter 3 the budget of reliability used in terms of stochastic delay constraints is decreasing the budget on the other aspects of the communication such as physical layer reliability as discussed in Chapter 6 or estimation reliability in Chapter 4. It is therefore logical to investigate

methods that can guarantee deterministic delay constraints which forms the motivation of the next section.

5.5 Hard Delay Guarantees with Successive Interference Cancellation Query Tree Algorithm SICQTA

This section introduces a deterministic delay constraint algorithm that is based on tree algorithms. Interference cancellation capability is used to improve the resource efficiency. Furthermore, the space of random user decisions are limited with user identities. This limitation imposes a maximum on total number of users supported while guaranteeing deterministic delay constraints.

SIC enables recovery of overlapping packets through signal processing. This has increased the throughput of random access algorithms from 0.5 packets per slot up to 1 packet per slot with infinite number of devices, reaching the efficiency of scheduling based solutions. The trade-off is the decoding complexity. Through edge-cloud processing and distributed computing, complexity is expected to be dealt with for radio access algorithms [ElS+18].

Successive interference cancellation is initially explored for tree algorithms in [YG05]. They showed that the throughput for tree algorithms is increased from 0.35 to 0.69. SIC requires that a clean packet is received so that it can be removed from a collision. In [YG05] this is guaranteed with feedback, forcing devices to split from each other. However, too much structure is inefficient and in [Liv11] it is shown that the same trees can be built through random decisions. The random decisions are shaped with a degree distribution tailored to the number of devices. It is shown that the algorithm reaches a throughput of 1 in the asymptotic region when M goes to infinity.

Another work [SPV12] adapts the SIC work to a frameless structure where the degree distribution is replaced with setting a Binomial probability to transmit at each slot. Compared to framed structure the results show that [SPV12] has a better performance in the non-asymptotic region. However, neither of these algorithms can provide a hard guarantee on the latency. Also both of them are susceptible to varying the number of active devices. The hard guarantees can be provided via setting the decisions uniquely for each device.

This problem is initially investigated by Massey under the name "protocol sequences" for de-synchronized devices in [MM85]. These algorithms are too pessimistic to be applied to tight latency constraints as the time offset between transmission time of devices is the main issue there and it is not the main problem any more thanks to the improvement in hardware design. The unique decisions for each device for hard guarantees is investigated in a recent

work [BVT18a] under the name "access codes", where each device transmits packets with respect to a unique code. The design of these codes is of combinatorial complexity. The results are limited, as we detail in subsection 5.7. Moreover, the use of feedback is neglected in their work.

Uniqueness of the access decisions can be guaranteed through feedback to overcome the complexity of the proposed protocol. This idea is introduced with Query Tree Algorithms in [CLL07]. However, the algorithm lags behind in throughput compared to SIC-capable algorithms. The idea to use Interference Cancellation for Query Tree Algorithms is introduced in [Kum+11]. However, the explanation of the algorithm in [Kum+11] is unclear. The throughput they have shown is capped to 0.69 which have already been shown by [YG05] for TA with SIC capabilities. Hard latency guarantees are not investigated and the difference to [YG05] is left unclear.

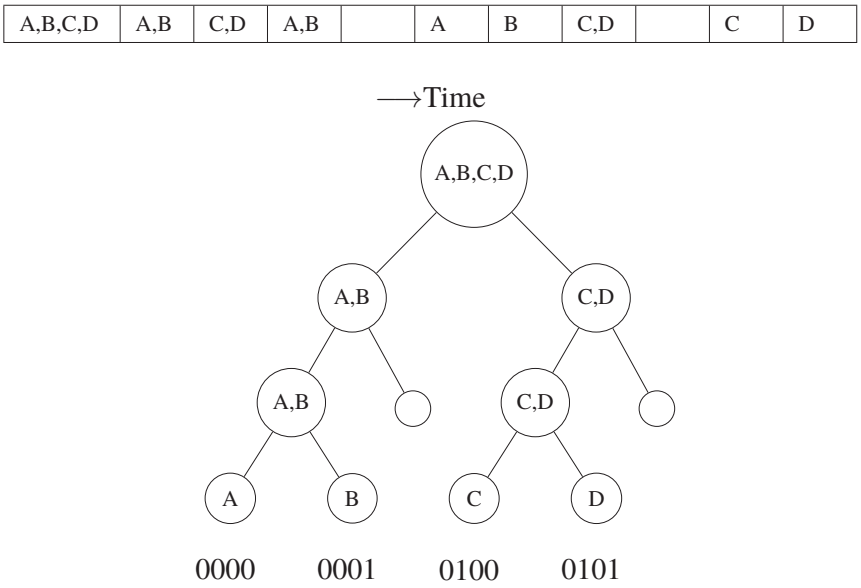
In our study we propose a novel Successive Interference Cancellation for Query Tree Algorithm, SICQTA. We provide analytical hard upper and lower bounds to and compare it with simulations to show the validity. It is shown that the algorithm easily extends to any number of active devices unlike *access codes*, and it provides a higher throughput compared to previous SIC based works. On top of that, hard latency guarantees make it a suitable candidate as a solution of the uplink resource allocation problem with unknown number of active devices.

The section is organized as follows: In subsection 5.6 we introduce shortly the Query Tree Algorithm and Successive Interference Cancellation Query Tree Algorithm. In subsection 5.7 the latency bounds are derived and we compare our solution to the *access codes* and to the simulations. Further discussions are given in subsection 5.8.

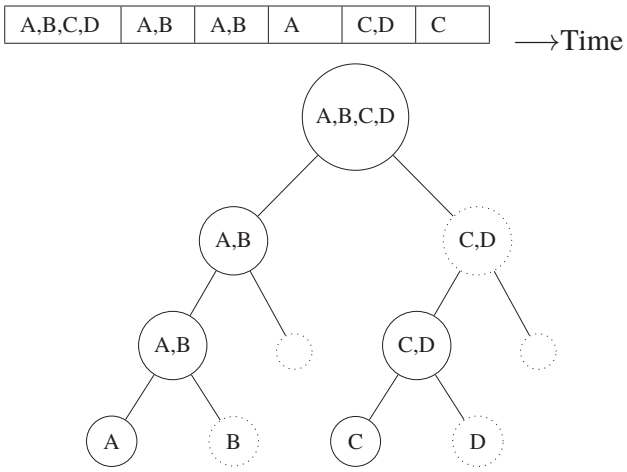
5.6 Algorithms with Feedback

5.6.1 Query Tree Algorithm with SIC (SICQTA)

SIC allows recovery of packets from a slot where a collision is observed. If for instance device A and B have transmitted a packet in slot 1, due to collision channel model, the outcome "A+B", is treated as a collision and slot is considered wasted. However, if device B has transmitted its packet in slot 2, the SIC model let us subtract B from "A+B" and enables recovery of A from slot 1. Instead of breadth first as did by QTA, the SICQTA goes depth-first. After the initial success, it checks if it can cancel the clean packet from previous collisions. If the packet is successfully cancelled then the algorithm skips the direct siblings of those slots. The algorithmic description of SICQTA is given in Alg. 3.



(a) QTA worst case with $M = 4$



(b) SICQTA worst case with $M = 4$

Figure 5.12: Worst-case example for Query Tree Algorithms with and without SIC with $M = 4$. $u = 3$ is set such that maximum number of devices is $n_{\text{tot}} = 2^u = 8$.

A detailed example for worst-case behavior of SICQTA is given in Fig. 5.12b for $M = 4$. In the first slot, all the devices are queried and it is a collision. On the second and third slot, addresses $0xx$ and $00x$ are queried, respectively. Both are collisions. The following slot, 000 is queried and it is a success. 001 is not queried, as the gateway recovered the packet from slot 2 and 3. This results in $k = 3$ as 2 slots are successfully recovered and this slot is a success. Addresses in query list Q: 001 and $01x$ is not queried and skipped. Thus, $10x$ is queried, that results in a collision. Following, 100 is queried and is a success. The gateway recovered D from slot 5 and the algorithm is terminated.

Algorithm 2 Query Tree Algorithm

```
1:  procedure GENERATE QUERY
2:      Q ← {'0','1'}                          ▷ Initialize Q list with '0' and '1'
3:      while Q is not empty do
4:          q ← Q[0]                           ▷ q is the first element of Q
5:          Transmit query at the beginning of time-slot
6:          Save received packets as r
7:          f ← |r|                            ▷ Number of received packets
8:          Q.pop                              ▷ Delete Q[0]
9:          if f = 0 or f = 1 then             ▷ Idle or success slot
10:             pass
11:         else if f > 1 then                 ▷ Collision slot
12:             Q.append('q0','q1')            ▷ Append 0 and 1 to the last query
13:         end if
14:     end while
15: end procedure
```

5.7 Analysis & Evaluation

In this section we will evaluate the latency of QTA and SICQTA and give bounds to its performance. We will also compare the performance of our work and [BVT18a] as we share the same problem definition. Finally, mean delay is compared with state of the art in tree algorithms to show that the stability region is extended.

5.7.1 QTA

An upper-bound for latency y of QTA is given in [LLS00]:

$$y \leq M\left(u + 2 - \log M\right), \tag{5.40}$$

where M is the number of active devices. This is a tight bound for $M \ll N$. For $M = N/2$ it has the most slack. Using the tree structure we can provide a tighter upper-bound for latency y as,

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2\left(u + 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor\right) - 1 \tag{5.41}$$

Similarly, the tree structure can be used to provide a lower-bound of latency as:

$$y \geq 2M - 1. \tag{5.42}$$

The proofs are given in App. D and D.2, respectively.

We explain why the example in Fig. 5.12a is the worst-case of a QTA with $M = 4$ also shedding light on the proof of the bounds. Four devices are separated into 2 groups of 2 as

Algorithm 3 SICQTA

procedure GENERATE QUERY
  $Q \leftarrow [\,], q \leftarrow$'0', $k \leftarrow$ '0' ▷ Initialization
  while $k - 1 \neq |Q|$ **do** ▷ End condition
    Transmit query at the beginning of time-slot
    Save received packets as $\mathbf{r}$
    $q_{\mathrm{b}} \leftarrow [q_1 \dots q_{n-1}\overline{q_n}]$ ▷ Invert last bit of q
    $f \leftarrow |\mathbf{r}|$ ▷ Number of received packets
    if $f = 0$ **then** ▷ Idle slot
      $q \leftarrow$'q_{b}0' ▷ Skipping collision
    else
      Q.append(q_{b})
      if $f > 1$ **then** ▷ Collision slot
        $q \leftarrow$ 'q0'
      else ▷ Cancel clean packet and skip.
        $q \leftarrow Q[-k]+$'0'
        $Q \leftarrow [Q[0], \cdots, Q[-k-1]]$
        ▷ Skip most recent $k-1$ queries thanks to SIC, $k \geq 1$.
      end if
    end if
  end while
end procedureEnd

close as possible to the root of the tree, so they cover as much as non-overlapping slots as possible. Following, devices have repeated the same collision, until the last level of the tree. We observe that for this scenario the total number of slots is $y = 11$. Using Eq. (5.40) we get 13. As expected the bound is valid and tight for this setting.

5.7.2 SICQTA

Intuitively, the efficiency of the [YG05] comes from the possibility to skip some slots in the tree. As it is shown in [YG05], the throughput of BTA is doubled. However, the throughput is the expected number of slots and this result cannot be directly translated to worst-case latency of SICQTA from QTA. We have to adapt the Eq. (5.40) for SICQTA using the skipping capability of SIC. The total number of skipped slots $\mathbf{S}$ compared to worst-case of QTA, given M active devices can be written as,

$$\mathbf{S} = \left\lfloor \frac{M}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor . \tag{5.43}$$

The proof is given in D.3.

We can use this finding to provide an upper-bound for latency of SICQTA using Eq. (5.41) and removing the skipped slots,

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor (u + 4 - \lfloor \log_2 M \rfloor) - 1 - \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor . \tag{5.44}$$

Intuitively, the algorithm needs at least M slots for M active devices and a lower-bound for latency of SICQTA can be given as,

$$y \geq M. \tag{5.45}$$

This is given without any proof, as in best-case no repetition occurs such that every slot is recoverable from another.

The upper-bound for latency can be used for the throughput calculation of the SICQTA. If number of active devices is the same as the number of total devices, i.e., $M = n_{\text{tot}} = 2^u$. Then we expect SICQTA to have a throughput of 1, as each slot in the tree should be different from one another.

Eq. (5.44) is a relaxed bound, but it becomes tight for integer values of $\log_2 M$. Plugging in $M = 2^u$ we get,

$$2^u \leq y \leq 2^{u+1} - 1 - 2^{u+1} + 2^u + 1 \tag{5.46}$$

$$= 2^u. \tag{5.47}$$

Thus, we have a throughput of 1 as expected. The proof is given in App. D.4.

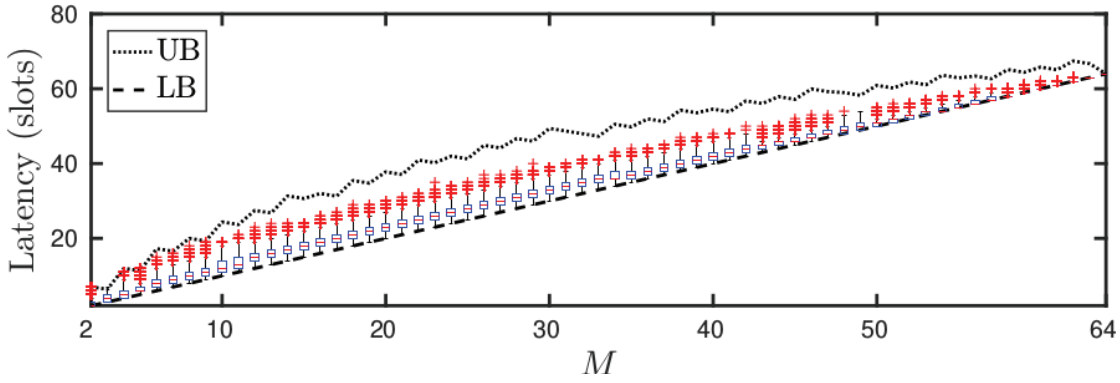
We can check the bound via the example in Fig. 5.12b. We see that in total 6 slots are used for SICQTA in the example. Using Eq. (5.44) we get 6 showing that the bound is valid and tight for this scenario.

In Tab. 5.2 we have compared the number of devices n_{tot} supported by Combinatorial Access Codes with SIC [BVT17] (CAC-SIC) with SICQTA. The number of active devices are fixed to $M = 3$ for CAC, because these are the only available results in [BVT18a]. For SICQTA, we see that with relaxed delay constraint the number of devices supported increases exponentially. And even though the results are similar for low latency constraints, the difference increases with increasing L. Also the results for SICQTA is easily extensible to other M values, while an exhaustive search is required to build codes for CAC-SIC. On the other hand effect of feedback is neglected in our analysis.

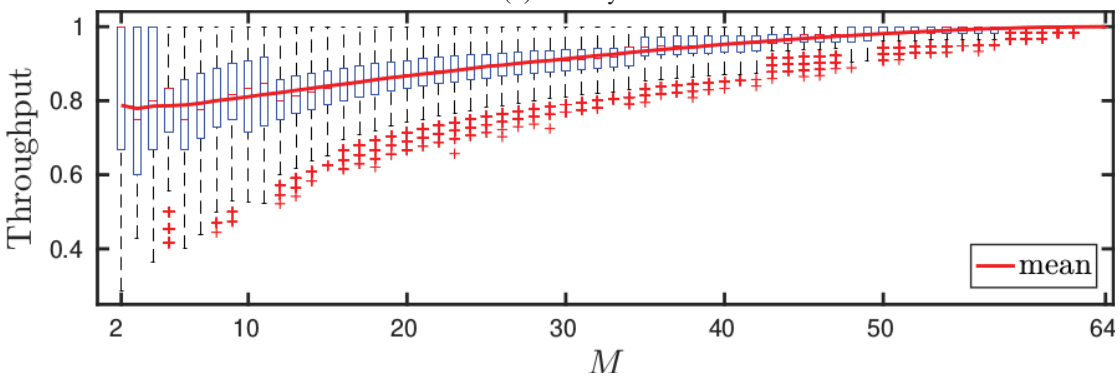
In Fig. 5.13 we have plotted the bounds versus simulation for SICQTA. x-axis depicts the varying active number of devices M and the y-axis presents the latency. We have set $u = 6$ so implicitly $N = 64$, and we have varied the number of active users M. We see that with 10^4 iterations for each data point in simulations the bounds are never surpassed and the difference between the lower and the upper bound is quite low.

Constraint	L= 4	L= 5	L= 6	L= 7
CAC-SIC [BVT18a]	7	11	–	–
SICQTA $M = 3$	8	16	32	64
SICQTA $M = 4$	4	8	8	16

Table 5.2: Number of devices supported by CAC-SIC for fixed number of active devices $M = 3$, with varying latency constraint, compared to SICQTA.



(a) Latency



(b) Throughput

Figure 5.13: Excessive simulations show the validity of the bounds. The maximum number of levels is set to $u = 6$, $N = 64$ and M is varied (x-axis).

As we deal with worst-case latency, this is the latency of the last device. In Fig. 5.13b we have evaluated the throughput with varying active number of devices M. Mean throughput is almost always above 0.8 while the tail is also quite constrained, especially with increasing M.

In Fig. 5.14 we extend the delay vs throughput comparison in [YG05] with SICQTA. In this simulation scenario continous arrivals are considered. If a device gets a packet to transmit while there is an on-going resolution, the device is queued until the end of that resolution, reflecting the setting in [YG05]. Each data point is simulated 10^6 times. We see that SICQTA enables a new throughput region that extends to throughput of 0.93 with $u = 4$. Also with $u = 6$ the throughput with stable latency is around 0.86. Of course SICQTA becomes similar

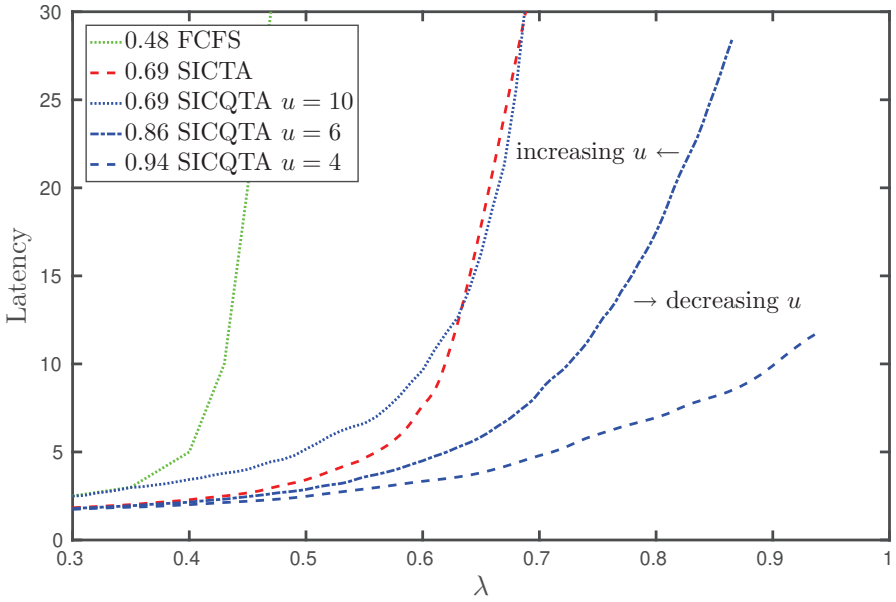


Figure 5.14: Delay vs throughput of feedback based random access algorithms.

to SICTA with increasing u value. This is logical as SICTA can be considered as a special setting of SICQTA with $u = \infty$. Here, it is shown with $u = 10$ that the behavior is almost the same as SICTA. It is worth mentioning that the average resolution time is increased as we see a shift on the y-axis compared to SICTA. We have also simulated higher values of u, i.e., $u = 16$ and did not observe any difference so they are not plotted here for clarity. For decreasing u the throughput is expected to increase further reaching 1.

5.8 Discussions

One important point for SICQTA compared to QTA is that the knowledge of number of active devices M does not improve the upper-bound of latency. The knowledge of M would be used in this case to skip to level $\lfloor \log_2 M \rfloor$. However, in the worst-case all collisions happening before this level consist of different devices, and under a SIC framework, they can all be recovered from each other to obtain useful slots. So the number of skipped slots with knowledge of M would be equal to those skipped due to SIC. However, application of knowledge of M to QTA can improve the worst-case performance and bring it close to SICQTA.

We have compared the feedback based algorithms to non-feedback based algorithms here. However, we assumed that the feedback is instantaneous and costless. In reality that is not the case. The latency incurred due to transmission and reception may even involve hardware delays such as switching from transmit to receive and vice-versa. We leave this open for future work.

5.9 Summary

This chapter investigated contention resolution for tight delay constraints from two aspects. First aspect is the use of resources orthogonal to time. We have shown that the access delay decrease is approximately inversely proportional to parallel resources. There is a loss of performance that is caused either due to structure of the algorithm or due to hardware limitations such as single radio chip on devices. We have described and then analyzed how would the parallelization work for tree algorithms. These results can be integrated to the admission controlled random access scenario proposed in Chapter 4 or upcoming 5G contention based scheduling that has to support tight delay constraints.

Second aspect is increasing efficiency of contention resolution algorithms. Recent work has shown that random access based successive interference cancellation can achieve a throughput of 1 packet per slot with high number of users. However, this is not true with low number of users. For some of the successive interference cancellation based algorithms the throughput even degrades lower than the conventional algorithms. For the throughput of 1 packet per slot, the uncorrelated activity of users is required that is why high number of users are required. The correlation is defined as the probability to select the same slot. The throughput is calculated with number of users divided by the number of slots in a frame. For instance with 2 users, to reach a throughput of one a frame with 2 slots 2 is required. As the readers can imagine with lower frame size, the correlations increase.

On the other hand compared to successive interference cancellation algorithms without feedback, e.g, Irregular Repetition Slotted ALOHA [Liv11], the tree algorithm benefits from correlations in user activity and the throughput increases with low number of users. This is because the randomness is limited and the feedback can quickly guide users to separate from each other. We have shown in Section 5.5 that even further limiting the access decisions to identities of the users, we can achieve the throughput of 1 with low number of users in tree algorithms, which is not possible with randomization based algorithms.

All in all, in this Chapter we have proposed solutions to two challenges. (1) the access algorithms have to make use of orthogonal resources to time and these resources have to be integrated in the analysis of the algorithms for delay constraints. (2) we have designed an algorithm to unlock the high throughput for low number of users and for tight delay constraints as emphasized in Fig. 4.4.

The algorithmic capabilities are investigated in an ideal MAC model. However, in a real system like the mobile networks some MAC parameters such as the resource bandwidth and duration. In this case the algorithms have to evaluated against these limitations as we

have done in our work in [Gür+19c]. The 5G integration of successive interference based contention access protocols is not further investigated in this thesis.

The capabilities of SIC based algorithms are based on a certain set of assumptions such as: no external interference, white noise and perfect cancellation. For a practical evaluation some of these constraints have to be integrated to analysis and their effects have to be characterized. On the next chapter we explain practical limitations with real implementations.

Chapter 6

IEEE 802.15.4 based prototype of Delay-Constrained Reliable Access

In the previous chapter we have investigated the capabilities of access algorithms. We have demonstrated that an improved receiver capability, namely successive interference cancellation (SIC), can double the resource efficiency if combined with an access algorithm design that is SIC-aware. In this chapter we investigate the practicality of such capabilities and test further assumptions in practical settings. For instance, independent frequency channel in an intra-aircraft communication system and hardware effects for SIC receiver is evaluated via prototyping sensors used for IEEE 802.15.4.

Layering based research demands that lower layers can be treated as a black-box. The black-box has input and output parameters. The function connecting the inputs and the outputs can be assumed. However, these assumptions can be specific to certain scenarios, too complex or unrealistic. Thus, the assumptions have to be tested. One way is theoretical evaluation of the assumptions that is doing an evaluation via adding each assumption one at a time. Another way is the measurement based characterization of the black box. As measurements are based on real data, it sets a limit on what is currently available. However, measurements are limited to underlying hardware and to the system under test. Due to this, measurements provide limited insight for forecasting future capabilities of black boxes. But depiction of current limitations is exact and leads to improved algorithm design. In this chapter we follow the practical approach that enables the deployment of previously discussed algorithms.

The parameters that can affect the reliability from a physical layer perspective are: frequency band, co-existence, coding, modulation scheme, mobility, wireless channel quality, transmission power, hardware quality and many more. First, the use of frequency bands is regulated and rules set the inter-technology interference. In case multiple technologies use the same band, co-existence rules have to be set to minimize this interference. Second;

coding, modulation scheme and transmission power are system design parameters that have to be selected carefully with respect to the wireless channel quality. These parameters can be pre-set or adjusted dynamically with respect to the channel quality information. The aim here is to operate the system close to capacity and under the required error probability. This can be achieved with characterization of the scenario for expected wireless channel quality variations and setting the design parameters accordingly. Lastly, with respect to the hardware cost and quality of the device, the device characteristics may deviate from the specifications. These hardware effects are also to be taken into account for selecting the right design parameters to guarantee that in practice the devices achieve a reliable communication.

We have introduced the intra-aircraft communications as a use-case and motivation for reliable uplink radio resource management. So, we also use it to break down the motivation for this chapter that is the reliability on the physical layer. Currently, the wireless communication in the aircraft is using the ISM band. ISM band is an unlicensed band such that it is free for everyone to use for Industrial, Scientific or Medical purposes. It is also used in process and factory automation. An aircraft and a factory is a private place and can be regulated by the owner such that no ISM band can be used without appropriate allowance. The deployment for ISM band can vary for factories but currently it is used for information and entertainment, infotainment, systems with IEEE 802.11 technology. The 2.4 GHz ISM band has a 80 MHz bandwidth that can be partitioned for different technologies such that no inter-technology interference occurs. This requires measurements and evaluations of different frequency planning settings under different traffic scenarios.

One can argue that, deploying safety critical applications with ISM band is not necessary, as there is Industrial 5G for factories or Wireless Aircraft Intra-Communication (WAIC) standardization upcoming that is making use of a licensed band. However, even for those licenced bands, it can be expected that multiple applications that deploy different technologies may co-exist. A frequency planning based solution is necessary due to scarcity of any band.

Even though all kinds of interference is avoided, the wireless channel has its own challenges to be solved. Each wireless signal is affected randomly by the wireless channel. These stochastic effects stem from; fading, shadowing, reflections and other natural phenomena such as lightning and have to be limited or overcome such that the signal is received reliably. Transmission power adjustment is used to overcome damping of the signal, coding is used to recover bursty effects on the signal, correct modulation is used to limit the effect of phase shifts or amplitude shifts. These techniques are of lesser importance if the stochastic behavior of the channel is well known. As the effects of the channel on the signal is characterized, it is possible to recreate the original signal undoing the wireless channel effects. These effects can well be characterized in a static indoor environment like factories and inside of an aircraft, as

the walls provide a shielding for external effects and if there is no mobility. Even though the wireless channel can be characterized to undo its effects, it is not the only reason for stochastic changes on the received signal. There is one mostly neglected aspect that is the hardware effects.

As the wireless sensors are to be deployed in thousands at a factory or in an aircraft, it is expected that each sensor costs less than two digits of the currency. This has its drawbacks as the waveform or the wireless signal may behave imperfectly that induces different effects. This can stem from many imperfections such as bandpass filtering, low noise amplifier and oscillator circuitry and many more. This creates another channel, hardware channel, that has other stochastic effects on the signal than the wireless channel. Most importantly, this channel, as the wireless channel, is unique among a transmitter receiver pair and is asymmetric as different RC components are used on each sensor on a receive and transmit operation. Some of the effects of this channel is undo-able such as increased noise floor, so characterization of the channel is important to use MAC layer recovery methods as failures are inevitably foreseen.

This chapter is organized as follows: The Sec. 6.1 introduces required background on IEEE 802.15.4 and successive interference cancellation. In the Sec. 6.2 we introduce our measurements in an intra-aircraft-like testbed with IEEE 802.15.4 sensors, this section is based in our work in [Gür+16]. We especially investigate the wireless channel behavior with and without interference. In the Sec. 6.3 we demonstrate with measurements the practicality of the Uncoordinated Uplink algorithm, SICQTA, introduced in Sec. 5.5, that is based on in our work in [GVK19]. Following, the insights obtained from the measurements resulted in a practical resource efficiency model reflecting the hardware effects.

6.1 Background and Related Work

As our prototyping efforts focus on IEEE 802.15.4, details of the standard is summarized here.

The IEEE 802.15.4 is deployed with the MAC layer of TSCH (Time-Slotted Channel Hopping) as this enables use of set of channels are used for hopping that provides frequency diversity. The set can be selected dynamically or statically depending on the scenario. Here we consider static setting of this set of channels.

6.1.1 802.15.4e TSCH

The basic idea behind the TSCH amendment to the 802.15.4 is to introduce a MAC scheme, which (a) allows a deterministic behavior and (b) provides higher reliability by frequency diversity. 802.15.4e TSCH supports multi-hop topologies, however, in the scope of this work, only single-hop is considered. The scheduling algorithm, as well as the length of the time slots, are design choices and not defined by the standard.

For network wide synchronization Absolute Slot Number (ASN), is propagated via a beacon to nodes, which is then used for the channel selection:

$$\mathsf{CH} = \mathsf{HS}[\mathsf{ASN}\%16], \tag{6.1}$$

where HS is the hopping sequence of channels, and CH is the resulting channel for a given ASN. For the measurements, we consider two different hopping sequences. Firstly, the full sequence, as defined in the minimal implementation of a 6TiscH Network (with a slight modification of using the advertising channel as second on the list for ease of implementation)[VP16] of OpenWSN [Wat+12]: $\mathsf{HS}_{\mathsf{full}}$: [16, 20, 23, 18, 26, 15, 25, 22, 19, 11, 12, 13, 24, 14, 17, 21], and reduced sequence: $\mathsf{HS}_{\mathsf{wl}}$: [15, 20, 25, 26]. Reduced sequence takes into account only the channels free from Wi-Fi interference, thus the non-overlapping frequency planning solution, and is referred to as whitelist hopping throughout this paper.

6.1.2 Related Work

6.1.2.1 Co-existence measurements for IEEE 802.15.4

In previous work from Blanckenstein *et al.* [Bla+15] TSCH is implemented in an aircraft in order to provide the bit error rate and loss characteristics in the presence and absence of passengers to model the effect of a realistic flight environment. The interference effect of WiFi entertainment system on WSN is not included, but rather multiple access points are offered as a solution for reliability increase. In other works from Gonga [Gon+12] and Du [DR12], WiFi coexistence of TSCH based WSN is investigated in an office environment in order to extract the effect of interference and frequency diversity. The first conclusion was that packet drop rate was not directly correlated to WiFi interference due to uncontrolled environment. Several correlations between consequent packet losses on different channels is extracted, but no conclusion is deducted.

6.1.2.2 Influence of Hardware effects on reliability for Interference Cancellation

The evaluation of hardware effect is crucial to output practical limits, and this is investigated in terms of radio irregularity In [Zho+06] the main causes in radio irregularity is summarized

as irregularities with antenna angle and battery power. A recent work [TBG11] demonstrates that sensors cannot reach the documented maximum transmission powers in their data-sheets. In [ZK07], the authors analyze hardware variance and outputted packet reception rate with no interference cancellation. These work neglected the impact on SNR of other hardware effects such as that of the oscillator.

In [HAW08], the authors develop a receiver employing Successive interference cancellation for IEEE 802.15.4 for intra-slot interference cancellation. The packet of the strongest user in a collision is decoded and canceled from the same time slot. Authors in [Lv+11] implemented intra-slot SIC for scheduling in IEEE 802.15.4. There have been many implementations of successive interference cancellation that perform modification at the physical layer such as [GK08] and[KL15] which is out of scope for us, as we intend to use the improvement to work with standard compliant IEEE 802.15.4 PHY.

A SIC-SINR model based on residual interference power for imperfect successive interference cancellation is introduced in [Web+07]. The authors provide bounds on the transmission capacity for imperfect successive interference cancellation wherein a fraction of the interference power is left behind after cancellation. In our work, we decompose this model into specific errors related to channel estimation and phase estimation to provide insights on the limitation of the CoTS hardware, validated through measurements.

6.2 Packet Level Measurements for Inter-Technology Co-existence

Co-existence of multiple technologies is a challenge of the ISM band. The main challenge is the technologies using the overlapping bandwidths should not affect each other's performance. Different technologies can be deployed by different entities, and certain techniques are incorporated in standards to guarantee fair usage of the band. Listen before talk and carrier sense techniques are two among many techniques. However, these techniques provides only stochastic performance guarantees. If multiple technologies are deployed by the same entity, the behavior of the technologies can be limited in terms of frequency or time to minimize interference with each other. Such an approach is also used for the cellular networks to be able to re-use parts of the spectrum over geographically separated areas.

The main challenge of frequency planning is allocating frequencies for geographically separated cells. The deciding factor in such scenarios are the setting the distance with respect to repeating the use of same frequency as no side-band interference is assumed idealistically [Arn80]. However, adjacent frequency interference is a practical issue. This stems from imperfect bandpass filters and harmonics of other frequencies. As the interference effect of

adjacent frequencies may differ with respect to used hardware the best solution is to avoid use of adjacent channels. However, due to capacity requirements this is not possible and at that case the most practical solution is to use measurement based evaluation [Ang+11].

Frequency planning is a well investigated topic for cellular networks, however the importance of it only comes recently for ISM band using standards such as IEEE 802.15.4 and IEEE 802.15.1. We am focusing on IEEE 802.15.4 as a proof of concept for evaluating the co-existence with the IEEE 802.11. The frequency planning for a controlled industrial environment covering IEEE 802.15.4 is an open issue that We want to investigate through selection of different frequency planning scenarios.

In this section, we examine the allocation of frequencies in a controlled interference environment, while assessing the reliability. The contributions is two-fold: (1) mitigation of WiFi interference in a controlled environment through a frequency planning called whitelisting, is validated for TSCH based WSN with real implementation; (2) We demonstrate that the assumption of uncorrelated loss for subsequent packet drops against interference is a good approximation to calculate the application failure rate from individual packet drops with limited transmissions for application.

The measurement setup and details on WiFi and WSN network are provided in the next section. Measurement results in terms of reliability are introduced in subsection 6.2.4.

6.2.1 Measurements Setup

The setup entails a realistic aircraft environment, where the simultaneous communication of a Wi-Fi network and a Wireless Sensor Network occurs over the unlicensed ISM 2.4 GHz bandwidth. All the measurements are performed in an isolate environment where no external interference is present. The final setup of the scenario is depicted in Fig. 6.1.

6.2.1.1 Wi-Fi network

The purpose of the Wi-Fi network is to provide an entertainment service to the passengers of the flight. The communication occurs between 59 Raspberry Pi clients and three different Base Stations (BS) occupying non overlapping portions of the spectrum, in particular, the channels 1, 6 and 11. Every BS serves 20 or 19 clients simultaneously.

The network deploys the 802.11g standard at the MAC layer with CSMA-CA and back-off mechanism, and it is unaware of the underlying WSN. Every client emulates the behavior of a passenger's device streaming video during flight. Every client is mimicked via a Raspberry Pi streaming 1296 kbit/s data continuously which results in $24 - 25$ Mbit/s average traffic on each channel.

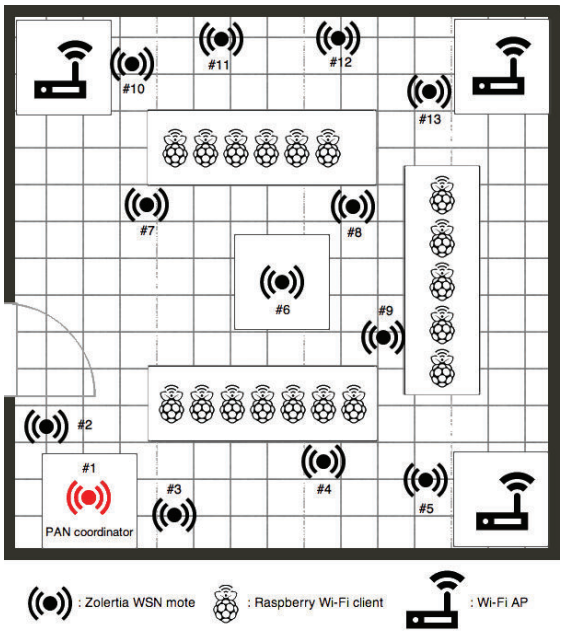


Figure 6.1: Scenario of Wi-Fi network and WSN co-existence in an aircraft environment. Room dimensions: $5.5m \times 6.35m$

6.2.1.2 WSN

The WSN is deployed to replace the fixed wired aircraft communication network with a reliable and flexible infrastructure. For this reason, the network consists in 13 randomly deployed nodes, and it has a star topology configuration. Every node is programmed through OpenWSN open stack implementation of IEEE 802.15.4 for physical layer and IEEE 802.15.4e TSCH protocol, which is followed by the 6TisCH and IPv6 on network layer [Wat+12].

Each mote is running a 15 time slot schedule. The PAN coordinator is entitled to transmit the Enhanced beacon (EB) during the first time slot containing the scheduling of the entire network. Once synchronized, every mote is allowed to transmit on a single reserved time slot with transmission power equal to -3 dBm, while during the other time slots it sleeps. Reliability is achieved by means of channel hopping and two additional MAC re-transmissions. A visual representation of the scheduling of the network is shown in Fig. 6.2. Due to this 15 slotted slotframe structure with 15 ms slot size T_s, each slotframe lasts 225 ms. This gives a transmission opportunity (TXOP) to each mote every 225 ms.

Every mote is running an uplink application that generates a packet of 53 Bytes every 700 ms. The packet generation rate from the application allows three TXOP for each packet before a new one is generated. When a packet is generated it is transmitted on the next available

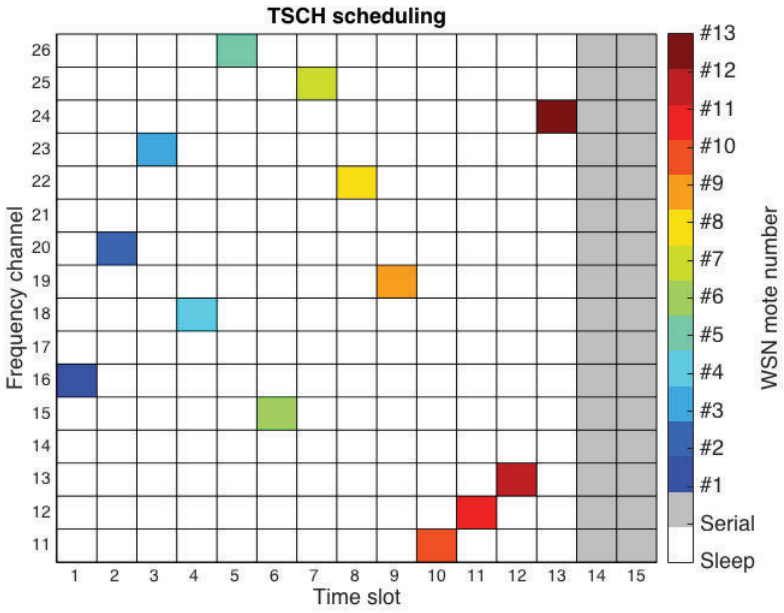


Figure 6.2: Scheduling grid in time/frequency domain deployed in the setup.

TXOP, and BS answers back with an ACK after the TXOP. If there is no ACK received back from the BS, it is re-transmitted on the next available TXOP.

We limit the maximum number of transmissions N_p including retransmissions and initial transmission to 3. Within each 700 ms a packet is generated and either successfully sent or dropped due to maximum number of retransmissions.

6.2.2 Scenarios

In order to evaluate the effects of different hopping strategies on the co-existence with Wi-Fi, we have performed the measurements with and without WLAN interference for three hopping options, resulting, in total, in six scenarios:

- **NINH**: **N**o Wi-Fi **I**nterference, **N**o **H**opping: all motes use one assigned frequency channel 20.
- **NIFH**: **N**o Wi-Fi **I**nterference, **F**ull **h**opping: the motes hop over the full list of 16 available channels.
- **NIWH**: **N**o Wi-Fi **I**nterference, **W**hitelist **H**opping: since an aircraft is a controlled environment, we can determine in advance which channels are free from or less prone to the Wi-Fi interference. Thus, we have whitelisted these channels and defined a hopping sequence consisting of channels: $\{15, 20, 25, 26\}$.

Table 6.1: Parameters and Notation Summary

Parameter	Explanation	Value
T_s	Time Slot Duration	15 ms
SFS	Slotframe size in timeslots	15
N_p	Maximum number of allowed transmissions	3
N_{cc}	Number of clear channels	4
N_{tc}	Number of available channels	16
ASN	Absolute Slot Number	
$HS[]$	Hopping Sequence	
HSS	Hoping Sequence Size	16
$CH_{j,i}$	i^{th} TXOP Channel after $CH_{j,1} = CH_j$	$1 - 16$
P_{comm_j}	Single packet loss probability on Channel j	
P_{app_j}	Application failure probability on Channel j	

- **WINH**: **W**ith Wi-Fi **I**nterference, **N**o **H**opping.
- **WIFH**: **W**ith Wi-Fi **I**nterference, **F**ull **H**opping.
- **WIWH**: **W**ith Wi-Fi **I**nterference, **W**hitelist **H**opping.

The duration of the measurements have been varied from 70 minutes to 900 minutes depending on the level of precision that is aimed for that measurement. A summary of the number of packets generated on mote and channel basis with the total number of packets can be seen in Tab. 6.2.

6.2.3 Reliability Analysis

Different measurements for application failure and the packet drop rate enable to assess the real measurements against the reliability assessment provided in [Gür+15]. The assumption in the paper was that we can neglect the correlation between subsequent packet drops in order to provide the application failure tolerance. I want to test this assumption against the real measurements to see if it holds and, furthermore, to determine the possible extension for more accurate modeling.

6.2.3.1 Top-Down Limits

The assessment of reliability with the top bottom approach using medium access parameters from our application can be evaluated. It is common practice to assume five nines for the

target reliability [Hel04], and we use this assumption for the loss rate above the link layer P_{app} of the star network as $P_{app} = 10^{-5}$. The assumption at this point is that higher layers have a reliability of one, since it is a one hop network. N_p is set to 3 with the slotframe length and the deadline of the application. Then, this information is converted to the packet loss on physical layer P_{comm} as in

$$P_{comm} = (P_{app})^{\frac{1}{N_p}}, \tag{6.2}$$

we can calculate the maximum tolerable packet drop rate as $P_{comm}^{\max} = 0.021$.

6.2.3.2 Reliability Calculations with Frequency Hopping

The channel hopping enables such that each retransmission is done on a different channel than the previous one. However, due to static slotframe and static hopping sequence, each of the retransmission channels can be calculated beforehand when the initial transmission channel $CH_{j,1}$ is known. In order to use the reliability assessment with frequency hopping, we have to modify the way we approach the calculations. The modeling of drops with constant P_{comm} is not feasible since every channel has its own characteristics due to the interference or frequency selective multipath.

To overcome this problem, after $CH_{j,1}$ is selected for the first transmission with the hopping sequence, we calculate the next frequencies used for the retransmissions with

$$\text{CH}_{j,i} = \text{HS}[(\textbf{ASN}_{\text{CH}_{j,1}} + \text{SFS} \cdot (i-1))\%\text{HSS}] \tag{6.3}$$

where $\text{CH}_{j,i}$ is the channel for i^{th} transmission after the initial channel $\text{CH}_{j,1}$, HS[] is the Hopping Sequence, HSS is the Hopping Sequence Size, ASN is the absolute slot number used and SFS is the slotframe size. N_p is 3 in our application. Let's assume a sensor have selected the channel 13. For the first transmission, $i = 1$ the SFS part is zero and this means that the ASN is 12 inside modulo 16. The SFS is 15, and the **ASN** increases by 15 with each slotframe, as i is incremented. For $i = 2, 3$, this results in ASNs of $27, 42$ which outputs $12, 11$ for the second and third transmission respectively. In short, the selected channels for 3 transmissions are $13, 12, 11$. The combined error rates are

$$P_{app_j} = \prod_{i=1}^{N_p} P_{comm_{\text{CH}_{j,i}}}. \tag{6.4}$$

This structure allows the use of drop rate for each of the channels in order to calculate the application failure probability with all possible combinations. In Eq. ((6.4)) the drop rate

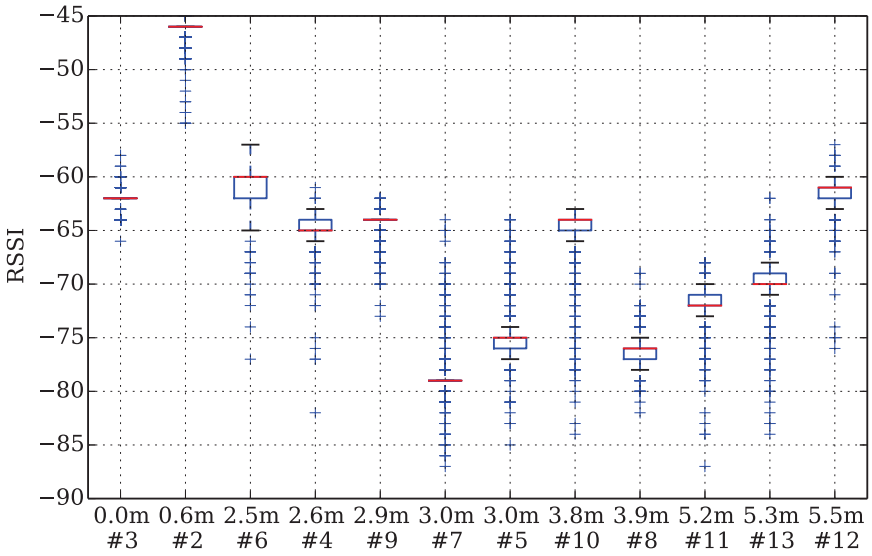


Figure 6.3: RSSI values of received packet vs. distance to PAN coordinator (# moteId) for NINH scenario.

probability for initial channel j is introduced. Since all of the j initial channels are selected randomly for a transmission, we calculate the average application failure probability with

$$P_{app} = \frac{\sum_{j=1}^{\text{HSS}} P_{app_j}}{\text{HSS}}. \tag{6.5}$$

For worst case analysis, the drop rate should be limited to the worst P_{app_j}.

6.2.4 Measurements Results

For every scenario, received packets are recorded together with the corresponding channel, Received Signal Strength Indicator (RSSI), sequence number, and MAC transmission count. In Fig. 6.3, the values of RSSI for every mote are presented. Scenario NINH is set as the base case. I observe that the distance alone is not sufficient to determine the channel quality. The RSSI fluctuations are high, and there is no correlation to the distance within our environment. Also, RSSI values are recorded only for the successfully received packets, thus, a large portion of packets (not received due to the RSSI smaller than a threshold, or discarded after the checksum check) is not included in the figure.

6.2.4.1 Reliability

For assessing the reliability, two metrics are used and compared, as defined in section 6.2.3: P_{comm} as packet drop rate and P_{app} as application failure probability. Figs. 6.4 and 6.5

illustrate these values for all six scenarios. Every box presents the distribution of the values for the set of motes, and the green line represents the averages weighted by the number of samples for every mote. We observe that, as expected, both application and communication failure rates are significantly lower for an interference-free scenario. It is also observed that the whitelist hopping reduces the average packet drop rate by 25%, whereas the application reliability is decreased by 5%. Since the channel choice for no hopping scenario is 20 (among whitelisted channels), difference between WINH and WIWH is minimal.

As the effects of hopping sequence selection is evident to investigate further, we proceed with evaluating the drop rate on a per-channel basis. The statistics for per-channel measurements are presented in Fig. 6.7 for drop rate per single channel. Also the total drop rate per "channel combination" including the re-transmission are given in Fig. 6.6.

The compliance of the measurements with the analytical values obtained with the framework in section 6.2.3 using per-channel measurements is evaluated. First, we use the packet drop for NIFH scenario 0.1783 and plug it in Eq. (6.6):

$$P_{app} = (P_{comm})^{N_p}, \tag{6.6}$$

with $N_p = 3$ transmissions, to obtain $P_{app}(ana) = 0.0057$. Then, we compare this result with the overall application failure measured for the scenario $P_{app}(meas) = 0.0062$. Following, for the interference scenario WIFH we get an application failure rate of $P_{app}(ana) = 0.0813$. When this result is compared with the overall application failure measured $P_{app}(meas) =$

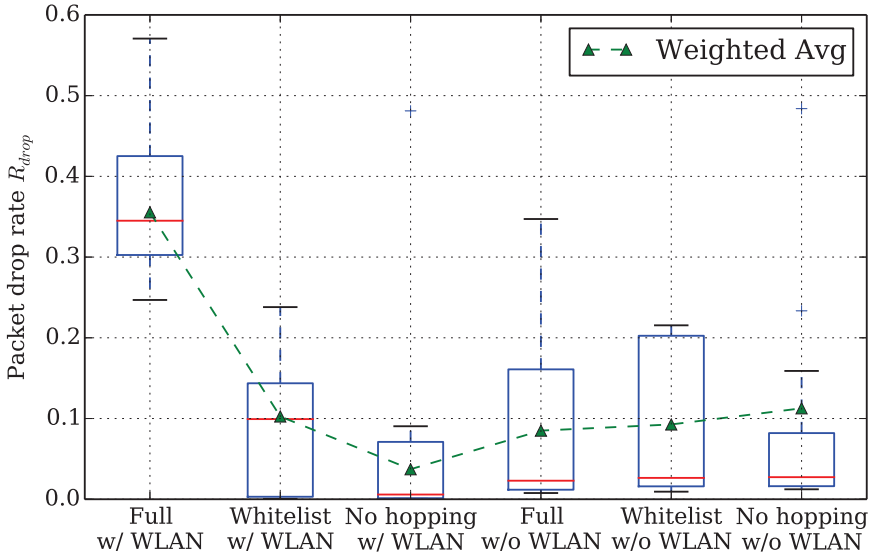


Figure 6.4: Packet drop rate for every scenario; weighted average considers number of packets every mote has generated.

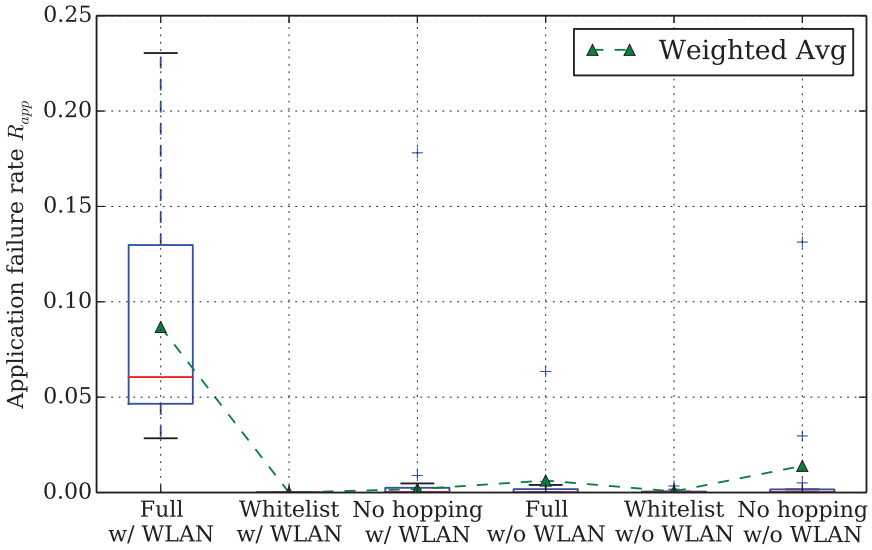


Figure 6.5: Application failure rate for all scenarios. Boxplot captures the distribution of the rate among all motes.

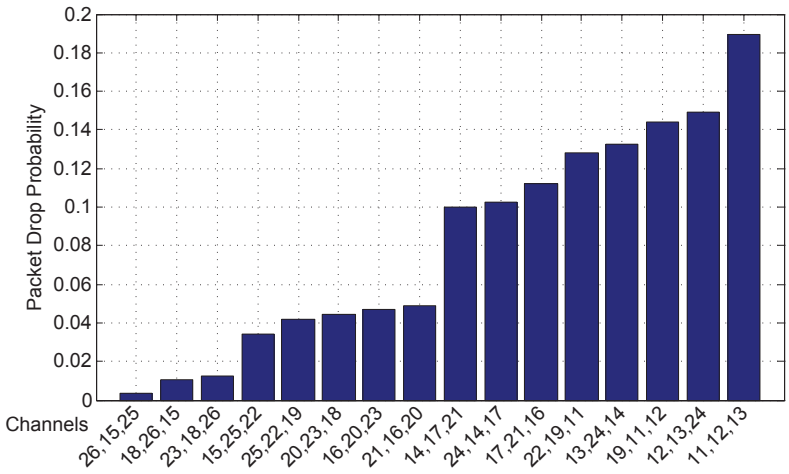


Figure 6.6: Combined Packet Drop Rate P_{app} including retransmission in IEEE 802.15.4 channels against Wifi interference in channels 1,6 and 11

0.0940, the difference is low. An important observation here is that in both of the cases the error is around 10%. This error, as explained in section 6.2.3, is, in fact, showing that effect of correlation is small when full channel list is used.

The whitelist scenarios are also evaluated with the calculations as shown in section 6.2.3. For NIWH and WIWH, the measured average packet drop rates are 0.08 and 0.09 respectively.

The application failure rate with the analysis and the measurements are summarized in Table 6.2. The analytical calculation assumes that there is no correlation between retrans-

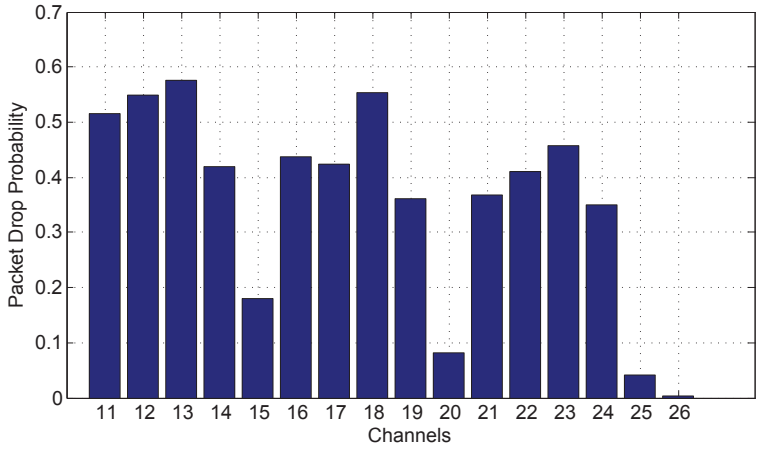


Figure 6.7: Single Packet Drop Rate P_{comm} in IEEE 802.15.4 channels against Wifi interference in channels 1,6 and 11

Table 6.2: Analysis and Measurement Comparison

	NINH	NIWH	NIFH	WINH	WIWH	WIFH
Dur.(min)	89	73	900	96	135	136
P. p. mote	7628	6257	77143	8228	11657	11571
P. p. chan.	91543	18771	57857	98743	34971	8678
P. total	91543	75086	925710	98743	139890	138860
P_{app}(ana)	$0.74 \cdot 10^{-3}$	$0.14 \cdot 10^{-3}$	$0.57 \cdot 10^{-2}$	$0.63 \cdot 10^{-3}$	$0.74 \cdot 10^{-3}$	$0.81 \cdot 10^{-1}$
P_{app}(meas)	$0.14 \cdot 10^{-1}$	$0.23 \cdot 10^{-3}$	$0.62 \cdot 10^{-2}$	$0.20 \cdot 10^{-2}$	$0.10 \cdot 10^{-3}$	$0.94 \cdot 10^{-1}$

missions. However, in the real system, this is not the case and there are different levels of correlation. From the table it can be seen that highest order of error is present for single channel measurements, which represents high correlation. We can notice that this effect decreases in whitelisted hopping where four channels are used and the assessment and measurement results are almost the same for full hopping.

As we can see from Fig. 6.6, only for channel 26 the measured packet drop rate is close to the tolerated value of 0.021. However, the channel 26 is not a solution for intra-aircraft communications since it is not part of ISM band in all of the countries. An alternative solution can be through use of re-transmissions with the whitelist. We have seen that with whitelist, the maximum tolerated application failure rate is not surpassed. For higher reliability requirements or lower latency requirements further techniques such as channel estimation, coding and/or transmission power settings is required.

6.2.5 Summary

In this section we present a reliability analysis of the co-existence of IEEE 802.15.4e TSCH Wireless Sensor Network with Wi-Fi in closed and controlled environments, such as an aircraft. We have conducted measurements on a real testbed for interference-free and interference scenarios, and evaluated packet drop rates and application failure rates. The measurements are compared to the analytical results.

The results emphasize the importance of controlling interference in case shared bands such as ISM band is used. Following, it is also shown that even without interference the reliability of some channels are not perfect. A dynamic algorithm to adjust the allocation of frequencies is investigated in our work in [Zop+17]. In the next section we investigate these imperfections for interference cancellation.

6.3 Bit level Measurements for Hardware Effects

In the previous section the effect of interference is investigated. In this section we assume the interference is controlled in a non-public area. Following, we prototype the SICQTA algorithm proposed in Sec. 5.5 with IEEE 802.15.4 commercial of the shelf sensors.

IEEE 802.15.4 [Sta] is one of the frequently available standards in factories thanks to its energy efficiency. The PHY layer is fixed in most of the devices as this is embedded in the radio chip but the MAC layer behavior is controlled through the network protocol. Thus, a lot of work in IEEE 802.15.4 [Wan+10] [DZG16] is focusing on optimizing the MAC protocol that can be updated easily for the deployed sensors. A large group of researchers focus on the coordination of the periodical monitoring sensors to optimize the use of the wireless medium. However, periodic monitoring is not fit for IIoT, as those devices are reacting to sporadic events. Random access, investigated as grant-free in 5G networks [3GP18], is a flexible solution for this problem. However, previous works on random access that consider low-latency and reliability constraints [Gür+19a] [Abb+17] [SLP17] assume a perfect physical layer performance that does not represent the practical limitations. In this section we evaluate a practical implementation to highlight the limitations with off-the-shelf hardware.

In this section, the interest is on a cell-based star topology using a time-division multiple access system. The time is divided into time-slots and only a single channel is used. Multiple users are attached to a central station. The users are not transmitting regularly. They are only reacting to events. The users can hear the broadcast downlink channel, but have to contend for the resources in the uplink. Thus, the resource is allocated in a distributed manner, using a random access algorithm, guided by the downlink feedback of the central station. As the load increases, the efficiency decreases quickly, making the solution unsuitable for traffic

bursts. It is shown that successive interference cancellation based random access algorithms can achieve efficiency matching that of the scheduled access without giving up the flexibility of random access [NP12].

In this section, we deploy an inter-slot successive interference cancellation, that fulfills low-latency and reliability constraints through inter-slot interference cancellations with an IEEE 802.15.4 network. We measure the practical performance in terms of resource efficiency, defined as packets received per slot. We extend the signal-to-interference-noise ratio (SINR) with hardware specific parameters to characterize the effect of commercial off the shelf (CotS) hardware. We use the extended SINR model to output a MAC model that shows a good match with the measurements and validate our model. The model provides insights for not only the state of the current devices but also the development of future hardware that uses successive interference cancellation (SIC).

The hardware we deploy are Z1 motes from Zolertia. Z1 motes are boards composed of off-the-shelf available chips such as CC2420 radio chip and a MSP430F2617 micro-processor both from Texas Instruments that can be bought for less than 10$. Such hardware fit perfectly for the vision of cheap hardware that will enable IoT for many use-cases including industrial communications.

Section 6.3.1 introduces the SINR model and extends it with the hardware effects. Section 6.3.2 introduces the MAC algorithm and the model used to calculate the practical throughput. The measurement setup and description of the data-set are introduced in Sec. 6.3.5 and the model is compared with measurements.

6.3.1 PHY Model

The SIC implementation employs an inter-slot interference cancellation scheme, where a replica of the packet involved in the collision is available in a different time slot without interference. This interference free signal is used to create a noise-free version of the packet. However, the channel effects have to be added over the recreated signal to properly cancel it from the collision slot. The distortions to the signal due to the channel as well as the hardware have to be estimated.

Mobility is not considered and the channel is assumed to be static. The wireless channel introduces a signal attenuation and multipath propagation in which, the copies of the signal traveling in different paths interfere at the receiver causing inter-symbol interference ISI. The channel effects are estimated using the symbol averaging method introduced in [HAW08], where the signal of the selected sequence of bits is found and averaged over the whole frame. This imitates the ISI and cancels out the noise representing the effects on that sequence of bits.

We investigate hardware effects such as phase noise by introducing a new parameter, v. The transmitter hardware noise effect is represented by a Gaussian random variable with variance v fitting a distribution to hardware effect based measurements. Let us assume the interference-free received signal is

$$r_i = x_i v_i h_i + n_i,$$

where, x_i is the modulated signal, v_i is the coefficient representing the effect of uncertainty coming from the transmitter hardware noise, h_i is the channel gain and n_i is the Additive White Gaussian Noise (AWGN) in the channel.

The channel estimation includes constant phase drift stemming from the hardware effects[1].However, this estimation depends on the hardware and is much worse when estimated in a collision slot. This error is defined inter-slot estimation and is represented as a constant ϵ, as introduced in [CH02].

The channel of the user 1 is estimated to be h_1 but it is h_1^o, with a constant empirical estimated mean ϵ,

$$h_1^o = h_1(1+\epsilon).$$

Hence, the residual signal, when the re-created signal with the estimated channel is removed from the received signal, r_1 is,

$$r_1 - x_1 h_1 = \underbrace{x_1 v_1 h_1 - x_1 h_1}_{\text{Hardware Noise Error}} + \underbrace{x_1 h_1 v_1 \epsilon}_{\text{Channel Est. Error}} + \underbrace{n_1}_{Noise}. \tag{6.7}$$

The SNR in an interference slot is the signal power γ_i divided by the residual power caused by hardware noise that is γ_i^p, the residual power caused by channel estimation error γ_i^c and the noise power γ_i^n

$$SINR_i = \frac{\gamma_i}{\gamma_i^p + \gamma_i^c + \gamma_i^n}. \tag{6.8}$$

For the case with successive interference cancellation, the SIC-SINR (SSNIR) of the i^{th} user is given by,

$$SSINR(i, \mathcal{S}, \mathcal{C}) = \Theta_i^{\mathcal{S},\mathcal{C}} = \frac{\gamma_i}{\sum_k^{\mathcal{S}} (\gamma_k^p + \gamma_k^c + \gamma_k^n) + \sum_j^{\mathcal{S}/\mathcal{C}} \gamma_j} \tag{6.9}$$

where $\mathcal{S}$ is the set of all packets transmitted at the same time and $\mathcal{C}$ represents the set of canceled packets including i.

[1]The phase noise is considered separately and not included in the channel

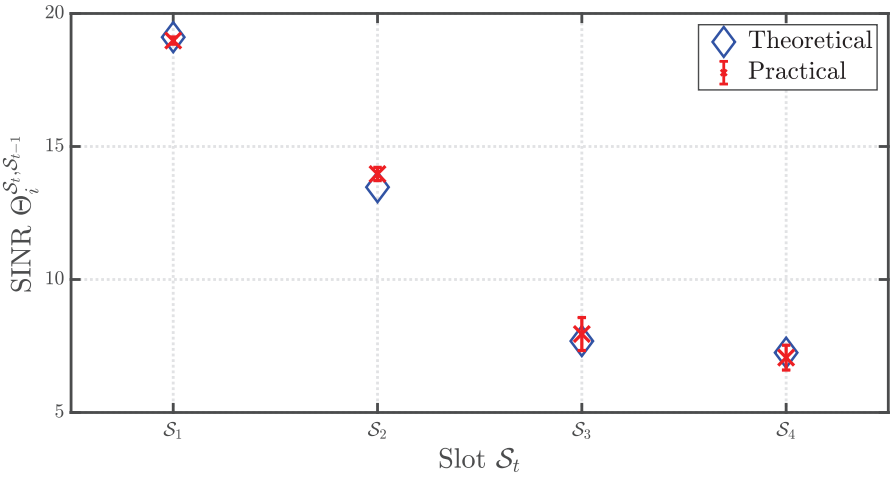


Figure 6.8: SIC-SNR model versus measurements showing validity of the selected parameters.

6.3.1.1 Measurements for SSINR model

The model parameters are obtained by performing measurements with Zolertia Z1 motes transmitting to a USRP B200-mini receiver. Channel gain h is estimated at the receiver. The statistics of noise, n, are collected from the portion of the received signal where no other signal is transmitted. The statistics of signal affected by the transmitter hardware noise and channel gain xvh for different users are estimated from the slots where the users' packets do not face any interference. Since the noise n is additive, knowing its mean and variance, the statistics of v are calculated treating them as the combination of 2 AWGN distributions with different mean and variance.

In our work, we adopt most of the PHY receiver chain of [HAW08] and adapt it for inter-slot interference cancellation adding cross slot channel estimation with phase drift estimation. The phase offset for a particular time slot can be determined by calculating the mean of the difference in phase between the ideal signal and the received signal over the length of the packet. The ideal signal is known from the interference-free replica of the packet.

The validity of the SIC-SNR model and the selected parameters is illustrated in Fig. 6.8 where measurements are given with 95% confidence intervals and compared to Eq. (6.9). The v parameter is set for each device through measurements. The typical values for the mean and variance of the v parameter as observed in the measurements for a Zolertia Z1 mote are around 1 and 0.01 respectively. The ϵ value is set as 0.2 representing channel estimation error when $i \neq k$ and set as 0.001 when $i = k$.

6.3.1.2 SNR to BER mapping

It has been reported that theoretical mapping from Signal to Noise Ratio (SNR) to Bit Error Rate (BER) is not too reliable for wireless sensor networks [Wu+12]. Channel models such as [Zhe+11], that represents the measurement setup, are discussed in state of the art. We will use the Rician SNR to BER mapping function $f_B(\gamma)$ for QPSK as given in [SA98],

$$f_B(\gamma) = \frac{1}{\pi}\int_0^{3\pi/4} \mathcal{M}_\gamma\left(-\frac{sin^2(\pi/4)}{sin^2(\theta)}\right) d\theta$$
$$\text{where, } \mathcal{M}_\gamma(s) = \frac{1+K}{1+K-s\gamma} e^{\frac{K\gamma}{(1+K)-s\gamma}}. \tag{6.10}$$

We set $K = 4$, fitting the measurements, that represents the relative strength of line of sight signal compared to non line of sight signal. The function outputs the bit error rate that can be converted to packet error rate or the decoding probability P_d using the packet length P_l in bits,

$$P_d(\Theta_i^{\mathcal{S},\mathcal{C}}) = \left(1 - f_B\left(\Theta_i^{\mathcal{S},\mathcal{C}}\right)\right)^{P_l}. \tag{6.11}$$

6.3.2 MAC Model

This section introduces the analytical modeling for the effect of user activity. We first introduce the typical MAC evaluation with interference cancellation channel model. Then, we extend this model with imperfect reception and cancellation, to show the MAC layer can provide versus the PHY layer errors.

6.3.2.1 Query Tree Algorithm with SIC (SICQTA)

In this work, we focus on an instant-feedback based RA algorithm that is the tree algorithm that is discussed in detail in chapter 5. SIC for tree algorithms is introduced in [YG05]. However, maximum latency in this algorithm is not bounded and the feedback is related to the addresses of the users to achieve latency bounds as introduced in [GGK19]. SICQTA algorithm is one of two algorithms that have deterministic latency bounds along with the access codes introduced in [BVT18b].

In SICQTA, every device has a unique ID composed of u bits. This limits the total number of devices attached to the central station to $N = 2^u$. In SICQTA, queries are used as feedback but the overhead is the same. Queries include a part of the address bits. The initial query is an empty query and all active users M answer to the query. A single bit is appended to the list of queries after each collision, starting from the left-most bit. Each new collision appends a new bit. As each device has a unique id, this guarantees that two devices have a unique access

$\mathcal{S}_6$	$\mathcal{S}_3$	$\mathcal{S}_2$	$\mathcal{S}_1$	$\mathcal{S}_5$	$\mathcal{S}_4$
A,B,C,D	A,B	A,B	A	C,D	C

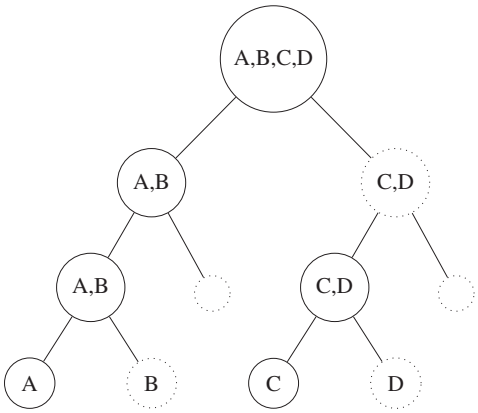


Figure 6.9: SICQTA worst case with $M = 4$

decision in worst-case after u transmissions (if all previous $u - 1$ bits are the same for two devices).

A detailed example for the worst-case behavior of SICQTA is given in Fig. 5.12b for $M = 4$ and $u = 3$, users A,B,C,D have addresses $\{000, 001, 100, 101\}$ respectively. In the first slot, all the devices are queried and it is a collision. On the second and third slot, $0xx$ and $00x$ are queried, respectively. Both are collisions. In the following slot, 000 is queried and it is a success. 001 is not queried, as the gateway recovered the packet from slot 2 and 3. This results in $k = 3$ as 2 slots are successfully recovered and this slot is a success. Ids in query list Q: 001 and $01x$ are not queried and skipped. Thus, $10x$ is queried, that results in a collision. Consequently, 100 is queried and is a success. The gateway recovered D from slot 5 and the algorithm is terminated.

6.3.3 User Activity Analysis

In this section the MAC layer model for performance is detailed.

We are interested in the actual performance of a SIC algorithm but initially, we introduce the performance a specific setting of the SICQTA algorithm with perfect cancellation and perfect reception. In our setting, the users have a $u = 3$ bit address. Hence, there are $N = 8$ users.

Consider the scenario that $M = 2$ users have collided while querying the first bit. Two possible ways in which this tree could split after this stage is as shown in Fig. 6.10. If a collision of two users happens multiple times, as in Fig. 6.10a, two slots are wasted and throughput is only 0.5 while in Fig. 6.10b throughput is 1. Thus, not all loss comes from the physical layer but due to lack of coordination in MAC.

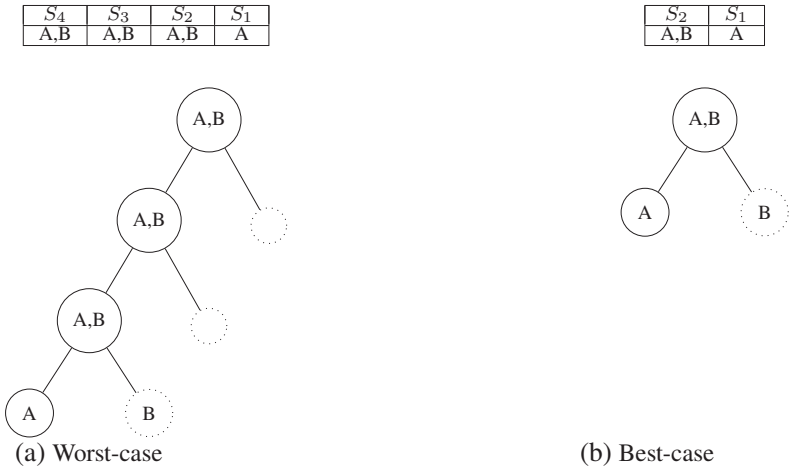


Figure 6.10: Possible tree structures for $M = 2$ active user and $u = 3$ maximum depth of the tree.

The average MAC throughput ρ is calculated by the formula,

$$\rho = \textstyle\sum_i P_{\text{occ}_i} \cdot \rho_i$$

where P_{occ_i} is the probability of occurrence of the i^{th} scenario and ρ_i is the throughput of that scenario. A scenario is the way the tree is formed.

Consider the scenario shown in Figure 6.10. The users are selecting slots uniformly. The probability of occurrence of the worst-case configuration given $M = 2$ is $\frac{1}{2^2} \cdot \frac{1}{2^2}$, as two users have done the same selection twice and $\frac{1}{2^1}$ for the best-case. The throughput is $\frac{2}{4}$ for worst-case and 1 for best-case scenario. In this way the throughput of all scenarios can be calculated.

6.3.4 Scenario Resolution Probability

The previous scenario neglected decoding errors. In order to calculate the actual throughput, we have to take into decoding errors into account. We define a scenario with errors if a decoding error occurs for any of the packets. If no decoding errors occur, scenario o can be fully decoded with probability P_{res_o}.

Thus, the throughput can be calculated as,

$$\rho = \sum_o P_{\text{occ}_o} \cdot P_{\text{res}_o} \cdot \rho_o. \tag{6.12}$$

The resolution probability P_{res_o} is assumed 1 in MAC throughput. Practically, the accumulation of noise and instantaneous variations in fading results in the SNR falling below the decoding threshold.

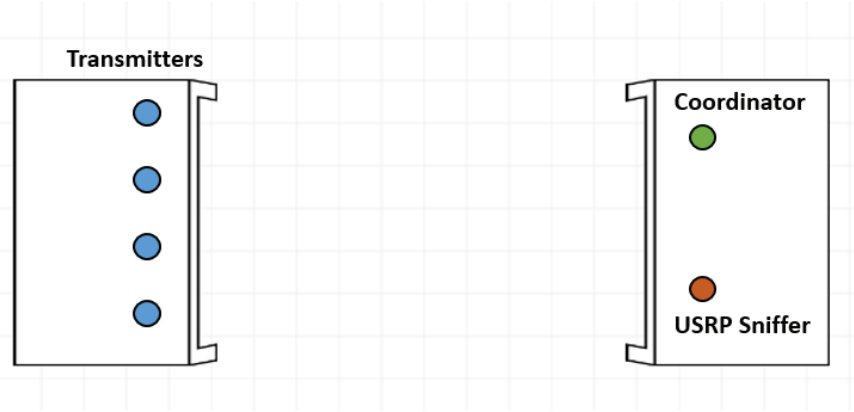


Figure 6.11: Experimental setup

The probability of resolution can be calculated as the joint probability distribution for successfully decoding all slots as in,

$$P_{\text{res}_o} = \prod_{t=1}^{n} P_d(\Theta_i^{\mathcal{S}_o^t, \mathcal{S}_o^{t-1}}), \tag{6.13}$$

where $\mathcal{S}_o^t$ is the list of all users that transmitted in slot t with scenario o. Slots in the frame can be depicted as $[\mathcal{S}_o^1, \mathcal{S}_o^2, \cdots, \mathcal{S}_o^n]$ where n is the number of slots in a frame and $\mathcal{S}_o^0$ is defined as an empty set. i is a packet that can be decoded from set $\mathcal{S}_o^t$ given set $\mathcal{S}_o^{t-1}$ is cancelled. The $\Theta_i^{\mathcal{S}_o^t, \mathcal{S}_o^{t-1}}$ depicts the SINR for packet i of a packet in slot t after the set $\mathcal{S}_o^{t-1}$ is cancelled. And example of how sets are defined can be seen in figures 6.9 and 6.10, in both examples $\mathcal{S}_1$ contains only A, while $\mathcal{S}_2$ contains A and B. As the tree structure is traceable thanks to the feedback, we assume that the slots are always ordered to make cancellation possible.

Idle slots are left out of the list as they have no effect on resolution. However, if two sets are equal, it means that the same packets are transmitted on two different slots. Thus, the success is decoding of either one of these slots. That can be calculated by checking the failure of either of values. For instance, given a scenario where the initial slot is repeated k times, Eq. (6.13) can be re-organized as,

$$P_{\text{res}_o} = \left(1 - \left(1 - P_d(\Theta_i^{\mathcal{S}_o^1})\right)^k\right) \cdots P_d(\Theta_i^{\mathcal{S}_o^n, \mathcal{S}_o^{n-1}}). \tag{6.14}$$

The hidden assumption here is, the realization on two different time-slots is not exactly the same in terms of noise and if cancellation is not possible with one, it can be with the other slot. This modification demonstrates that repetitions for successive interference cancellation increase the reliability exponentially. This will be further evaluated in the results section.

6.3.5 Measurements

In this section we explain our experimental setup, the code deployed on sensors and the measurements collected. Furthermore, the results are post-processed using the models in section 6.3.2.

6.3.5.1 Experimental setup

A testbed of 5 Zolertia Z1 nodes and a USRP B200-mini receiver is set up as shown in Figure 6.11, to collect measurements.

Four Z1 nodes act as the transmitters and each has a direct line of sight path to a fifth Z1 node acting as the network coordinator and central station. All of the Z1 used are 5 years old representing a mid-life of an industrial sensor. A USRP node is set up to act as a sniffer. It has a direct line of sight path to the transmitters and receives all the packets. All nodes in the network communicate on the IEEE 802.15.4 channel 26 which is centered at 2480 MHz. A single slot is of duration 4 ms and is implemented through OpenWSN.

We are measuring a SICQTA scenario of $M = 4$, $N = 32$. The users, as illustrated with upper nodes, are set to transmit 128 byte packets. We repeated this scenario for 100 consecutive frames. The sniffed data[2] captured with the SDR is processed in MATLAB.

6.3.5.2 Processing measurements

The raw data file sampled at 4 MHz is read into MATLAB and the beginning of a frame is detected by comparing the signal strengths to a noise and interference threshold. The threshold is manually set by observing the received data as 0.1 which is above the interference level and below the signal level.

Firstly, all time-slots with a single packet are decoded. The location of the replicas can be traced back thanks to the tree structure. Then the replicas are canceled from time-slots with collisions. After each SIC iteration, a canceled slot is decoded and the decoded packets are again canceled from other slots. A scenario is considered successful if all packets involved in the scenario are canceled. Following this, the measurement success probability is calculated.

6.3.6 Evaluation

In Tab. 6.3 we have listed the measurements and the model. The resolution probabilities in Tab. 6.3 is re-arranged using Eq. 6.12 and summarized in Tab. 6.4 among the MAC layer throughput.

[2]We plan to make the data and the processing scripts available in the camera-ready version of the paper

Scenario	Meas.	Model
2221	1	0.9997
221	0.98	0.9954
21	0.9	0.9324
3321	0.9	0.8757
3221	0.82	0.7498
321	0.74	0.7024
3311	0.98	0.9391
3121	0.9	0.9324
311	0.87	0.7532

Scenario	Meas.	Model
44211	0.84	0.8643
4321	0.59	0.5125
4311	0.56	0.5496
422121	0.88	0.9282
42121	0.81	0.8695
42211	0.7	0.7263
4211	0.64	0.6804
4111	0.66	0.7297

Table 6.3: Resolution probability P_{res_o} for the different scenarios of 2, 3 and 4 user collisions

Scenario	Perfect PHY	Measured	SSINR Model
4 users	0.875	0.5837	0.6026
3 users	0.8344	0.6926	0.6465
2 users	0.7917	0.7273	0.7495

Table 6.4: Comparison of theoretical and Practical throughputs

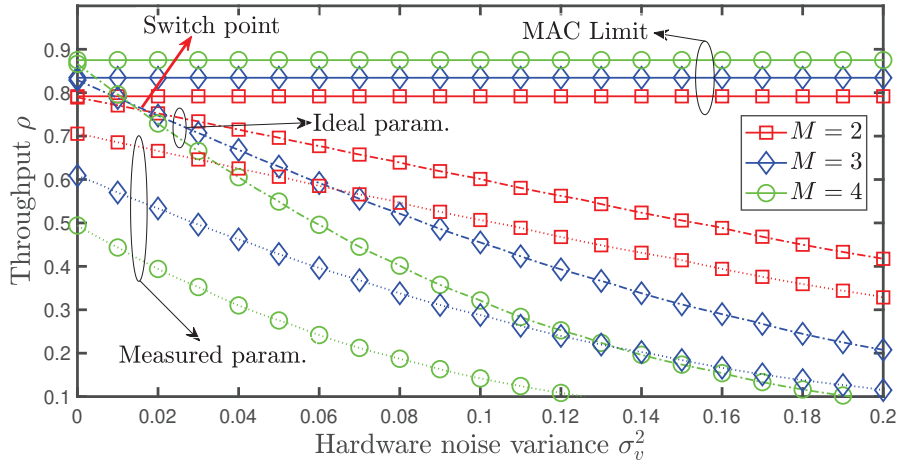


Figure 6.12: Sweep analysis for the hardware noise variance σ_v^2 using Eq. 6.12. Two different set of parameters are evaluated with ideal and measured variables. In the ideal parameter setting every parameter except the hardware noise variance is set to best possible values.

Firstly, the measurement results confirm that inter-slot SIC can be used to improve throughput for contention based access using the CotS IEEE 802.15.4 hardware with throughput reaching at least 0.58 and up to 0.72. When $M = 4$ the radio latency is at maximum 72 ms representing 3 frames with the worst-case scenario and on average 24 ms. In terms of data-rate using contention based access and a single channel this translates to a reliable 170 kbit/s. In case all 16 channels are used, this can be further boosted approximately to 2.7 Mbit/s. The results are promising for industrial scenarios with sporadic activity and a delay constraint of 100 ms.

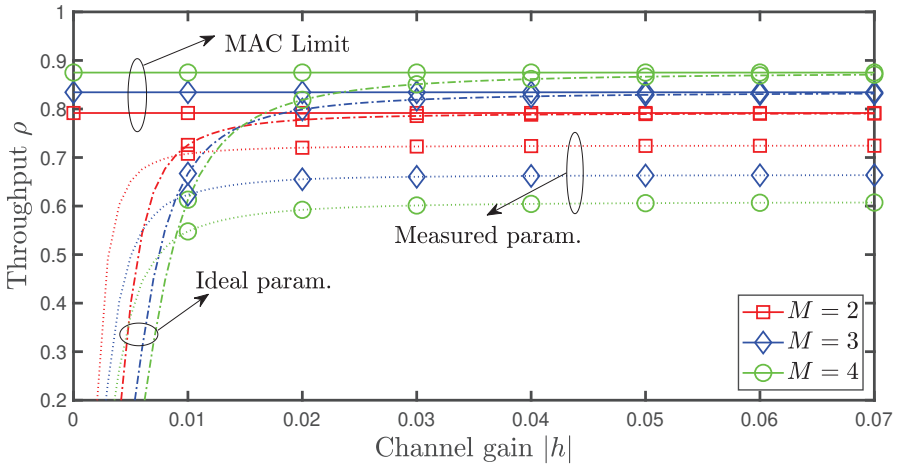


Figure 6.13: Throughput Eq. 6.12 vs. channel gain. Two different set of parameters are evaluated with ideal and measured variables. In the ideal parameter setting every parameter except the channel gain is set to best possible values.

Taking a deeper look in the results, the MAC layer throughput loss is 20% compared to 7% due to PHY problems with 2 users. As SICQTA performs better with more users e.g., 4, the MAC layer throughput loss in this case is 12.5% while the PHY loss is 29%. Thus, quickly for increasing number of users the perfect physical layer assumption is far from reality and the SSINR model becomes significantly important. The model shows a good match overall. SIC for 4 users is quite common in most inter-slot SIC algorithms and the perfect PHY assumption can be replaced with the model we provide here for realistic evaluations. Another approach is clearly use of better hardware that can decrease the worst-case latency if no re-transmissions are needed.

Encouraged by the match of the model and the measurements, we do a sweep analysis of hardware variance using the model. In Fig. 6.12 we have plotted MAC throughput on y-axis versus the hardware noise variance in the x-axis. As expected increasing hardware noise variance decreases the SIC capability. As the un-cancelled portion of the signal increases, the SIC success probability decreases. The results show that even with a really high channel gain and no estimation error, after $\sigma_v^2 = 0.03$, having $M = 2, 3, 4$ outputs the same throughput as the collisions with more users have an increased decoding error probability. The results show that hardware effects impose a big limitation to implement inter-slot SIC. In case new IEEE 802.15.4 hardware will be bought the phase noise can be evaluated to decide to consider the SIC capabilities using the switch point depicted in Fig. 6.12.

The effect of the channel gain is illustrated in Fig. 6.13. The measured parameters reflect that after a channel gain of 0.01 the maximum reachable throughput is reached due to estimation error and hardware noise. With ideal parameters the channel gain required to

reach the MAC limit is around 0.05. This evaluation reflects the required transmission power, antenna selection and placement of the sensor to reach the required MAC throughput.

6.4 Summary

In this chapter, we have presented a reliability analysis of the co-existence of IEEE 802.15.4e TSCH Wireless Sensor Network with Wi-Fi in closed and controlled environments, such as an aircraft. We have evaluated the expected application and communication reliability levels for an exemplary application. Then, we have conducted measurements on a real testbed for interference-free and interference scenarios, and concluded the effects of frequency planning. A critical observation during this measurement setup is the packet success rate varies heavily for each hardware.

Motivated by the observed behavior, we have analyzed the effect of hardware in depth. We introduced a practical SINR model including the hardware effects that we have validated through experimental results in our testbed. We have demonstrated that CotS IEEE 802.15.4 chips, in a contention based access, can achieve a similar throughput compared to scheduled access. This is achieved through using a successive interference cancellation MAC algorithm in the receiver that does inter-slot interference cancellation. To the best of our knowledge, it is the first experimental evaluation of inter-slot SIC for IEEE 802.15.4, which is particularly challenging due to variations in the signal from slot to slot due to cheap hardware.

This chapter serves as a proof of concept for the discussed algorithms in chapter 4 and 5. Even though the exact same performance is not reached, the measured performance is better than the state of the art algorithms. Thus, this chapter validates the previous assumptions partially showing with measurements that the suggested algorithms indeed outperform state of the art algorithms as implemented with off-the-shelf chips.

Chapter 7

Conclusion

Enabling the safety critical applications over wireless networks for industrial internet of things requires involvement of, from bottom to the top, all communication layers. In this thesis a similar approach is followed, starting from the definition of safety for aircraft applications until inclusion of hardware variance of sensors to make it a part of bottom to top reliability figure. This allowed us to pinpoint the most critical parts in the big picture for scaling and the physical layer reliability.

We have analyzed the scaling capability by introducing an adaptive step to the access algorithm. This step is an estimation algorithm for adapting the resources in the system. The resources are adapted with respect to the required quality of services. While guaranteeing resource efficiency, also a mechanism which reacts to unexpected events without sacrificing quality, is created.

As the physical layer is another aspect susceptible to unexpected events we also characterized such events for inter-technology interference and hardware effects. The inter-technology interference effects are characterized for an aircraft scenario where a controlled network is assumed. The effect of hardware variance on the MAC layer performance is modeled and validated with measurements that need to be considered for wireless network planning.

We believe both aspects solves a practical yet crucial problem using analytical insights, taking us one step closer to the enabling of migration and enabling of safety critical tasks over wireless networks to unleash the full potential of industrial IoT.

7.1 Future Work

In this work we have provided crucial steps towards enabling safety critical applications through wireless communications for industrial internet of things. However, there are many more steps required to achieve a complete infrastructure for industrial internet of things.

Some of these aspects are not directly related to wireless communication, such as safety and dependability for artificial intelligence and control systems; reliable and reactive code testing algorithms; digital twin systems that enrich the outcome space to foresee unlikely events and prepare the system accordingly.

The use of channel estimation pilots to improve wireless decoding quality is clear to everyone in wireless communications society. The pilots use precious network resources but their provided gain for decoding, makes this trade-off logical. This thesis demonstrates the usefulness of estimation for the user activity. Further evaluation has to be done with real wireless network traces. More concrete results have to be shown to prove that user activity estimation pilots can contribute to wireless decoding. In future we can see user activity detection signals implemented in mobile networks standard.

In Chapter 5 we have introduced a new type of random access algorithm that has deterministic delay bounds. However, this requires that the number of address, i.e., number of users is limited. For different delay guarantees, users may have to be separated to different address spaces. Feasibility and the overhead of this separation is important to analyze to prove the practicality of the proposed algorithms. Otherwise, they may lose the competition to the Combinatorial Access Codes.

Security is another one of the further challenges in wireless communications. Security can be decomposed in vulnerability to jamming; man in the middle attacks and data privacy. Such solutions can lie in the space of using multiple radio access technologies in combination that is investigated under the name MultiRAT.

Another problem is enabling diverse quality of services in the same network that is investigated under radio access network slicing. The effects among different slices to each other is still an open topic.

Going deeper in to physical layer, a missing piece is precise models for wireless channels that is appropriate for high reliability constraints. Also, channel estimation methods that are valid for high reliability is another aspect that needs deeper inspection.

Appendices

Appendix A

Reliability of FSA with K-MPR

We start with the basic definition given in Eq. (4.20)

$$r(g, \mathrm{L}, n_{\mathrm{a}}) = \frac{\mathrm{E}[N_{\mathrm{r}}]}{n_{\mathrm{a}}}. \tag{A.1}$$

For FSA with K-MPR, we can calculate $\mathrm{E}[N_{\mathrm{r}}]$ as the product of the expected number of resolved users in a K-superslot and the number of K-superslots

$$\mathrm{E}[N_{\mathrm{r}}] = \left\lfloor \frac{g\mathrm{L}}{K} \right\rfloor \mathrm{E}[N_{\mathrm{s}}] \tag{A.2}$$

where N_{s} is the random variable denoting the number of resolved users in a K-superslot.

The expected number of resolved users in a K-superslot can be calculated in the following way

$$\mathrm{E}[N_{\mathrm{s}}] = \begin{cases} \sum_{i=1}^{K} i \binom{n_{\mathrm{a}}}{i} \pi^i (1-\pi)^{n_{\mathrm{a}}-i} & \text{if } n_{\mathrm{a}} \geq K \\ \pi\, n_{\mathrm{a}} & \text{else,} \end{cases} \tag{A.3}$$

where π is the K-superslot selection probability by an arrived user, given by $\pi = \frac{1}{\lfloor \frac{g\mathrm{L}}{K} \rfloor}$. We simplify the expectation further for $n_{\mathrm{a}} \geq K$ as

$$\mathrm{E}[N_{\mathrm{s}}] = \pi\, n_{\mathrm{a}} (1-\pi)^{n_{\mathrm{a}}-1} \sum_{i=0}^{K-1} \binom{n_{\mathrm{a}}-1}{i} \left(\frac{\pi}{1-\pi}\right)^i. \tag{A.4}$$

Finally, by plugging (A.4) into (A.2) and then into (A.1), we get

$$r(g, \mathrm{L}, n_{\mathrm{a}}) = (1-\pi)^{n_{\mathrm{a}}-1} \sum_{i=0}^{K-1} \binom{n_{\mathrm{a}}-1}{i} \left(\frac{\pi}{1-\pi}\right)^i. \tag{A.5}$$

Appendix B

Proofs related to sporadic user activity evaluated in Chapter 4

In this appendix we share proofs for some of the closed form equations used in Chapter 4.

B.1 Proof for expectation calculation of Coupon Collector's Problem with unequal probabilities

We start by repeating the probability p_i that any user accesses a channel i. If we have z users in the system, we have a mean arrival of $\lambda_i = p_i \cdot z$ on the i^{th} resource. Thus the idle probability on that resource is $e^{-\lambda_i} = e^{-p_i \cdot z}$. Non idle probability on that resource is $1 - e^{-p_i \cdot z}$. If we multiply this probability for all resource that had a busy signal we get, $\prod_{i \in \mathcal{M}_{s+c}} (1 - e^{-p_i z})$, which is the probability to have non-idle on all the busy resource. This probability can be used with a maximum likelihood and the resulting z will be maximum allowed z in the system. However, if we take the probability of observing at least one non-idle in the busy resources $1 - \prod_{i \in \mathcal{M}_{s+c}} (1 - e^{-p_i z})$, we have a decreasing equation. We can take then the expected value of the effect of each user added to the system over the probability of observing at least one non-idle in the busy resources. Thus, it gives us

$$E[Z|\mathcal{M}_{s+c}] = \sum_{z=0}^{\infty} \left(1 - \prod_{i \in \mathcal{M}_{s+c}} (1 - e^{-p_i z}) \right). \tag{B.1}$$

B.2 Proof for collision probability with u users accessing M preambles with unequal probabilities

Collision probability p_c is a sub problem of probability of observing u balls in any bins, with N balls into M bins with unequal probabilities. We start this with re-defining the set of possible frequencies as $\mathcal{M} = \{1, ..., M\}$. Then we define $\mathbf{S}^J$ that denotes the sequence for all possible J-ary combination sequences of the elements of set $\mathcal{M}$. An example would be $\mathbb{S}^2 = (\{1,2\},\{1,3\},\{2,3\})$, with $\mathcal{M} = \{1,2,3\}$ and $J = 2$. We will also use the term $\mathbb{S}^J_{x,y}$ where $x \in (1, ..., \binom{M}{J})$ denotes different sets in the sequence and $y \in (1, ..., J)$ denotes different elements of each combination sequence. From the example we have $\mathbb{S}^2_{1,2} = 2$ and $\mathbb{S}^2_{3,2} = 3$. We will also use $\mathbb{S}^J_x$ when we want to refer to just the set.

Now we denote $W^L_{x,y}[u]$ as the probability function for selecting u users out of N users for the L^{th} time as given in $W^L_{x,y}[u] = \binom{N-u(y-1)}{u}\left(p_{\mathbb{S}^L_{x,y}}\right)^u$. Now we denote $Z^J_{x,y}[u]$ as the probability function for $N - J \cdot u$ users out of N users selecting all the other frequencies except the ones denoted by the set x as given in $Z^J_{x,y}[u] = \left(1 - \prod_{z=1}^{y} p_{\mathbb{S}^J_{x,z}}\right)^{N-J\cdot u}$. Using these we define the probability function to obtain J occurrence of u users out of N users with a recursive calculation

$$P^x_J[u] = \left(\prod_{y=1}^{J} W^J_{x,y}[u]\right) Z^J_{x,y}[u] - \left(\sum_{j=J+1}^{max(J)} \sum_{x \in \{\mathbb{S}^J_x \subset \mathbb{S}^j_x\}} P^x_j[u]\right) \tag{B.2}$$

where $max(J)$ is maximum number of occurrence of u given N users which is given with $min(\lfloor \frac{N}{u} \rfloor, M)$. In Eq. (B.2) the upper part calculates the joint probability of having J occurrence of u users, while the lower part is subtracting the probabilities for $j > J$ occurrences. After we have non-overlapping probabilities for all occurrences of u, i.e., probability to have **just** J occurrence of u users, we can sum them up to have the probability to obtain u users,

$$p_c[u] = \sum_{j=1}^{J} \sum_{x \in \mathbb{S}^J_x} (P^x_J[u] \cdot l) \tag{B.3}$$

where we multiply with the occurrence of u users since we treat each outcome independently and we have to weigh accordingly.

B.3 Proof for admission rejection probability p_r

The analysis of loss system is used for call blocking probabilities in [Gim65]. We can use the same analysis for loss probabilities in our system. The Erlang-B formula is given via, $p_B = \frac{\frac{A^N}{N!}}{\sum_{o=1}^{N} \frac{A^o}{o!}}$ where A denotes the traffic, N the number of servers and B is the blocking

probability. The A mean number of resources needed as in $A = h_{RAQ} \cdot \lambda_{RAQ} = L_j \cdot M_{AC_j}$. The number of servers are taking into account the parallelization factor. Each parallel tree can solve a different collision. Finally we have,

$$p_r = \frac{\frac{\left(L_j \cdot M_{AC_j}\right)^{M_G}}{M_G!}}{\sum_{o=1}^{M_G} \frac{\left(L_j \cdot M_{AC_j}\right)^{o}}{o!}}. \tag{B.4}$$

Appendix C

Parameter justification for Multichannel Tree Algorithm

In this appendix, we justify certain parameter selections for multichannel tree algorithm in Chap. 5.

C.1 Selection of M for limiting the infinite sum

M will be the maximum level that we will consider in the analysis of the pmf of $\mathcal{T}^N$.

Since any node trespassing level M will not be taken into account in the computation of the pmf, we want to set M as high as possible. On the other hand, the greater M the bulkier the operations will be, as more terms will be considered in them. In order to choose an optimum M, we need to compute the probability of a tree reaching the level M. With that objective in mind, let us define $\mathcal{M}^N$ as the random variable modeling the last level reached by a tree of N contenders. In [JJ+00], the authors provide the pmf of this random variable:

$$p_{\mathcal{M}^N}(m) = \mu(2^m, N) - \mu(2^{m-1}, N), \tag{C.1}$$

where

$$\mu(\alpha, \beta) = \begin{cases} 0 & \text{if } \alpha < \beta, \\ \frac{\alpha!}{(\alpha-\beta)!\,\alpha^\beta} & \text{if } \alpha \geq \beta. \end{cases} \tag{C.2}$$

Provided that we have chosen an accuracy ϵ, we need to select M such that:

$$\epsilon \geq \sum_{m=1}^{M} \left(\mu(2^m, N) - \mu(2^{m-1}, N)\right), \tag{C.3}$$

which can be easily accomplished by numerical search. Then, we can choose the required M for a desired accuracy. Fortunately, M grows slowly as we increase either the required

accuracy or the number of contenders, since the maximum number of nodes at each level grows exponentially, and so do the number of opportunities to successfully transmit. For instance, only $M = 36$ is required to guarantee that at least $\epsilon = 99.9\%$ of trees will be finished even if $N = 10000$, and $M = 30$ for $\epsilon = 99.99999\%$ with $N = 60$.

C.2 Derivation of the probability of a given partition

In (5.33), the probability of obtaining a partition π_i with k balls and x bins was presented as:

$$p_{\mathcal{P}^{k,x}}(\pi_i^{k,x}) = \frac{k!}{\Psi_{x,x}^{k}} \prod_{j=1}^{x} \frac{1}{\eta_{i,j}^{k,x}! \cdot \#_{i,a}^{k,x}!}. \tag{5.33}$$

In this Appendix, the derivation of this expression will be tackled, using a slightly simplified notation for clearness. Let us start by computing the number of ways $Z_k^{\eta_1}$ to choose η_1 balls out of a total of k balls:

$$Z_k^{\eta_1} = \binom{k}{\eta_1}. \tag{C.4}$$

Moreover, the number of ways $Z_{k-\eta_1}^{\eta_2}$ to choose η_2 balls out of a total of $k - \eta_1$ balls is:

$$Z_{k-\eta_1}^{\eta_2} = \binom{k-\eta_1}{\eta_2}. \tag{C.5}$$

In general, the number of ways $Z_k^{\eta_n}$ to choose η_n balls out of a total of $k - \sum_{i=1}^{n-1} \eta_i$ balls is:

$$Z_k^{\eta_n} = \binom{k - \sum_{i=1}^{n-1} \eta_i}{\eta_n}. \tag{C.6}$$

If every part η_n in the partition π with x parts is different, the total number of ways A^π to generate such partition is simply:

$$A^\pi = \prod_{j=1}^{x} Z_k^{\eta_j} = \prod_{j=1}^{x} \binom{k - \sum_{i=1}^{j-1} \eta_i}{\eta_j}. \tag{C.7}$$

After some basic manipulation based on the definition of the binomial coefficient, we can rewrite (C.7) as:

$$A^\pi = k! \prod_{j=1}^{x} \frac{1}{\eta_j!}. \tag{C.8}$$

Nevertheless, if some parts have the same value, e.g. $10 = 4 + 4 + 2$, the number of ways to select those parts would be counted multiple times, yielding an incorrect result. In order to solve this issue, we have to correct by the number of ways to arrange those repeated values:

$$A^\pi = k! \prod_{j=1}^{x} \frac{1}{\eta_j! \#_j!}. \tag{C.9}$$

Finally, the probability of partition π is obtained by dividing the number of ways A^{π} to generate that specific partition by the total number of ways to generate any partition, which is given by $\Psi^{k}_{x,x}$, i.e. the number of ways to arrange k balls in x groups, x of which have more than one ball. Therefore, we have reached our final result:

$$p_{\mathcal{P}^{k,x}}(\pi) = \frac{k!}{\Psi^{k}_{x,x}} \prod_{j=1}^{x} \frac{1}{\eta_j! \#_j!} \tag{C.10}$$

We just need to use the full notation in (C.10) to obtain (5.33).

Appendix D

Bounds for QTA latency

In this appendix we will investigate the proofs for latency bounds of Query Tree Algorithm.

D.1 Proof for upper-bound for latency of QTA

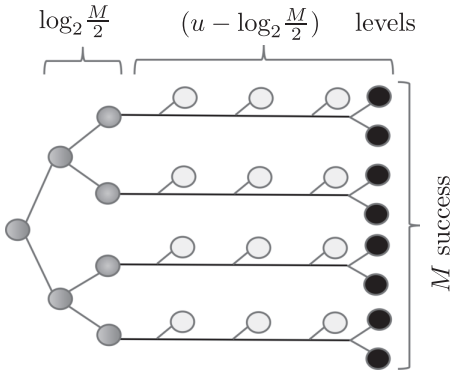


Figure D.1: The worst-case tree structure for Query Tree Algorithm.

The worst-case for QTA is illustrated in Fig. D.1. An intuitive explanation is as follows: A device can re-transmit at maximum u times in the worst-case as that is the size of IDs and every device has a unique ID. In this case the device is successful with the u^{th} transmission and it has experienced $u-1$ collisions. In order to have a collision we need at least 2 devices, and at the worst-case all devices are grouped into two, thus $\lfloor \frac{M}{2} \rfloor$ groups. Each group collides separately $u-1$ times, where there will be idles on the unexplored slots so $2 \cdot (u-1)$ slots, followed with 2 transmissions for success of each device, we get

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot (u - 1 + 1) \tag{D.1}$$

slot uses in total. We take into account, the activity of the groups of two only after the level $\left\lfloor \log_2 \frac{M}{2} \right\rfloor$. As the initial levels have a lot of overlap, we can remove these levels and consider

them separately as

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot \left(u - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + \tau, \tag{D.2}$$

where τ represents the overlapping slots. The number of overlapping slots can be calculated by summing the total number of slots up to level $\log_2 \frac{M}{2} \geq \lfloor \log_2 \frac{M}{2} \rfloor$ of the tree. We can calculate the total number of nodes in this upper part of the tree as,

$$\tau \leq 2^{m+1} - 1 = 2^{\log_2 \frac{M}{2} + 1} - 1 = M - 1. \tag{D.3}$$

Plugging this in Eq. (D.2) we get,

$$y \leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot \left(u - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + M - 1, \tag{D.4}$$

$$\leq \left\lfloor \frac{M}{2} \right\rfloor 2 \cdot \left(u + 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) - 1. \tag{D.5}$$

D.2 Proof for lower bound for latency of QTA

The best-case in the tree with M devices, is that they are organized as a triangle, guaranteeing they are as close as possible to the root. So the level of the successes are almost the same. However, the level of the devices can be the same only if $\log_2 M$ is an integer. If it is not an integer, the best-case would be some of the devices are successful at level $l_1 = \lceil \log_2 M \rceil$ and the others are at $l_2 = \lfloor \log_2 M \rfloor$. In order to have a complete triangle we would need that devices at level l_2 would each have 2 children at l_1. So the number of slots at l_1 is equal to the sum of number of devices at l_1 plus twice the number of devices at l_2. The number of slots at a level can also be written as 2^l so we can write,

$$2^{l_1} = M_{l_1} + 2 \cdot M_{l_2} \tag{D.6}$$

where M_{l_1} and M_{l_2} is the number of devices successful in l_1 and l_2 respectively. We know that the total number of devices is $M = M_{l_1} + M_{l_2}$. So we can re-write Eq. (D.6) as

$$M_{l_1} = 2 \cdot M - 2^{l_1}. \tag{D.7}$$

If we do not consider the level l_1, the tree is a full triangle up to level l_2. We can calculate the the lower bound y_{LB} for the total number of slots in the tree for through calculating the number of slots for the full tree up to l_2 and adding M_{l_1}

$$y_{LB} = 2^{l_2+1} - 1 + M_{l_1}. \tag{D.8}$$

By definition of flooring and ceiling operation $l_2 + 1 = l_1$ if $\log_2 M$ is not an integer. And we can plug Eq. (D.7) in to get,

$$y_{LB} = 2^{l_1} - 1 + 2 \cdot M - 2^{l_1} \tag{D.9}$$

$$= 2 \cdot M - 1. \tag{D.10}$$

When $\log_2 M$ is an integer the lower-bound is directly given with $2^{\log_2 M+1} - 1$, which is equal to the result so we do not mention it separately.

D.3 Proof for number of skipped slots

The skipping in SICQTA consists of two different parts. First part is skipping the idles S_I and second part is skipping the cancelled slots S_C. So we can write the total amount of skipped slots $\xi = S_I + S_C$.

The upper-bound for latency of QTA is derived using groups of 2 devices sticking together until the last level of the tree. At the last level they transmit separately, each as a success. The idles occur after separation from the top triangle until the end of the tree. We have $\left\lfloor \frac{M}{2} \right\rfloor$ collisions and the number of levels until the end of the tree gives us the number of skipped idle slots as

$$S_I = \left\lfloor \frac{M}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right). \tag{D.11}$$

Thanks to SIC, after one success the other device does not have to transmit anymore, as after one success the other device can be recovered from the previous collision. Thus, at least $\frac{M}{2}$ slots are skipped for the last level of the tree.

This skipping can be applied to also formation of groups of 2. Groups of 2 are formed from groups of 4. Thus, for the first group formed out of 4 devices, the other group can be recovered from the collision, so one slot can be saved for each separation. In this step we can save $\frac{M}{4}$ slots. This logic can be extended up to $\lceil \log_2 M \rceil$ separations as we have a binary splitting process. This gives us,

$$S_C = \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor. \tag{D.12}$$

Finally, we can write,

$$\xi = \left\lfloor \frac{M}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{M}{2} \right\rfloor \right) + \sum_{i=1}^{\lfloor \log_2 M \rfloor} \left\lfloor \frac{M}{2^i} \right\rfloor. \tag{D.13}$$

D.4 Proof for number of skipped slots with high number of active users

We plug in $M = 2^u$ to Eq. (D.12)

$$\xi = \left\lfloor \frac{2^u}{2} \right\rfloor \left(u - 1 - \left\lfloor \log_2 \frac{2^u}{2} \right\rfloor \right) + \sum_{i=1}^{\lfloor \log_2 2^u \rfloor} \left\lfloor \frac{2^u}{2^i} \right\rfloor = \sum_{i=1}^{u} 2^{u-i}. \tag{D.14}$$

As u is the number of maximum levels and is an integer we can remove the floor operation getting,

$$\begin{aligned} \xi = \sum_{i=1}^{u} 2^{u-i} = 2^u \left(\sum_{i=0}^{u-1} 2^{-i} - 1 + 2^{-u} \right) \\ = 2^u \left(\frac{1 - 2^{-u}}{1 - 2^{-1}} - 1 + 2^{-u} \right) = 2^{u+1} - 2^u - 1. \end{aligned} \tag{D.15}$$

So we can plug it in Eq. (5.44) to get,

$$y \leq \left\lfloor \frac{2^u}{2} \right\rfloor (u + 4 - \lfloor \log_2 2^u \rfloor) - 1 - 2^{u+1} + 2^u + 1. \tag{D.16}$$

Bibliography

Publications by the author

Journal publications

[Vil+17] M. Vilgelm, H. M. Gürsu, W. Kellerer, et al. "LATMAPA: Load-Adaptive Throughput-MAximizing Preamble Allocation for Prioritization in 5G Random Access." In: *IEEE Access* 5 (2017), pp. 1103–1116.

[Gür+17b] H. M. Gürsu, M. Vilgelm, W. Kellerer, et al. "Hybrid collision avoidance-tree resolution for m2m random access." In: *IEEE Transactions on Aerospace and Electronic Systems* 53.4 (2017), pp. 1974–1987.

[Gür+19a] H. M. Gürsu, M. Vilgelm, A. M. Alba, et al. "Admission Control Based Traffic-Agnostic Delay-Constrained Random Access (AC/DC-RA) for M2M Communication." In: *IEEE Transactions on Wireless Communications* 18.5 (2019), pp. 2858–2871.

[Zop+18] S. Zoppi, A. Van Bemten, H. M. Gürsu, et al. "Achieving Hybrid Wired/Wireless Industrial Networks With WDetServ: Reliability-Based Scheduling for Delay Guarantees." In: *IEEE Transactions on Industrial Informatics* 14.5 (2018), pp. 2307–2319.

Conference publications

[Gür+15] H. M. Gürsu, M. Vilgelm, W. Kellerer, et al. "A wireless technology assessment for reliable communication in aircraft." In: *2015 IEEE International Conference on Wireless for Space and Extreme Environments (WiSEE)*. IEEE. 2015, pp. 1–6.

[GAK17a] H. M. Gürsu, A. M. Alba, and W. Kellerer. "Slotted ALOHA Filtered Tree (SAFT) for a Reliable LTE RACH." In: *European Wireless 2017; 23th European Wireless Conference*. VDE. 2017, pp. 1–7.

[Gür+18] H. M. Gürsu, B. Köprü, S. C. Ergen, et al. “Multiplicity Estimating Random Access Protocol for Resource Efficiency in Contention based NOMA.” In: *2018 IEEE 29th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE. 2018, pp. 817–823.

[GKS19] H. M. Gursu, W. Kellerer, and C. Stefanovic. “On throughput maximization of grant-free access with reliability-latency constraints.” In: *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE. 2019, pp. 1–7.

[GGK19] H. M. Gursu, F. Guan, and W. Kellerer. “Hard Latency-Constraints for High-Throughput Random Access: SICQTA.” In: *ICC 2019-2019 IEEE International Conference on Communications (ICC)*. IEEE. 2019, pp. 1–7.

[Gür+16] H. M. Gürsu, M. Vilgelm, S. Zoppi, et al. “Reliable co-existence of 802.15. 4e TSCH-based WSN and Wi-Fi in an Aircraft Cabin.” In: *2016 IEEE International Conference on Communications Workshops (ICC)*. IEEE. 2016, pp. 663–668.

[GK17a] H. M. Gürsu and W. Kellerer. “Deadline-aware wireless sensor network routing: The JLAT metric.” In: *2017 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE. 2017, pp. 612–617.

[Vil+16] M. Vilgelm, H. M. Gürsu, S. Zoppi, et al. “Time Slotted Channel Hopping for smart metering: Measurements and analysis of medium access.” In: *2016 IEEE International Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2016, pp. 109–115.

[GK17b] H. M. Gürsu and W. Kellerer. “Gains of Deadline based Discarding (DbD) over Lossy Wireless Sensor Networks.” In: *2017 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE. 2017, pp. 1–6.

[Gür+19c] H. M. Gürsu, M. Ç. Moroğlu, M. Vilgelm, et al. “System Level Integration of Irregular Repetition Slotted ALOHA for Industrial IoT in 5G New Radio.” In: *2019 IEEE 30th Annual International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. IEEE. 2019, pp. 1–7.

[GVK19] H. M. Gürsu, H. Vijayaraghavan, and W. Kellerer. “Boost Your CotS IEEE 802.15.4 Network with Inter-Slot Interference Cancellation for Industrial IoT.” In: *CCNC 2020* (2019).

[Zop+17] S. Zoppi, H. M. Gürsu, M. Vilgelm, et al. “Reliable hopping sequence design for highly interfered wireless sensor networks.” In: *2017 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*. IEEE. 2017, pp. 1–7.

Others

[Gür+17a] H. M. Gürsu, M. Vilgelm, E. Fazli, et al. “A Medium-access Approach to Wireless Technologies for Reliable Communication in Aircraft.” In: *Wireless Sensor Systems for Extreme Environments: Space, Underwater, Underground, and Industrial* (2017).

[GAK17b] M. Gürsu, A. M. Alba, and W. Kellerer. *Delay Analysis of Multichannel Parallel Contention Tree Algorithms (MP-CTA)*. 2017.

[Gür+19b] H. M. Gürsu, S. Zoppi, E. Fazli, et al. “MAC-aware Delay Optimization for Wireless Sensor Network Gateway with a Microcontroller.” In: (2019).

[GZK19] H. M. Gürsu, S. Zoppi, and W. Kellerer. “Wireless Sensor to Cloud Delay: A Network Hardware and Processing Perspective for Industrial Internet of Things.” In: (2019).

General publications

[QHR15] S. Qiu, J. Hu, and M. Rauterberg. “Nonverbal signals for face-to-face communication between the blind and the sighted.” In: *Proceedings of International Conference on Enabling Access for Persons with Visual Impairment*. 2015, pp. 157–165.

[PB09] V. Petukhova and H. Bunt. “Who’s next? Speaker-selection mechanisms in multiparty dialogue.” In: *Workshop on the Semantics and Pragmatics of Dialogue*. 2009.

[Bia00] G. Bianchi. “Performance analysis of the IEEE 802.11 distributed coordination function.” In: *IEEE Journal on selected areas in communications* 18.3 (2000), pp. 535–547.

[Sha48] C. E. Shannon. “A mathematical theory of communication.” In: *Bell system technical journal* 27.3 (1948), pp. 379–423.

[Bro03] E. Brown. “Fundamentals of terrestrial millimeter-wave and THz remote sensing.” In: *International journal of high speed electronics and systems* 13.04 (2003), pp. 995–1097.

[TV05] D. Tse and P. Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[Zan97] J. Zander. "Radio resource management in future wireless networks: Requirements and limitations." In: *IEEE Communications magazine* 35.8 (1997), pp. 30–36.

[Mar05] J. W. Mark. "Downlink resource management for packet transmission in OFDM wireless communication systems." In: *IEEE Transactions on Wireless Communications* 4.4 (July 2005), pp. 1688–1703. ISSN: 1536-1276. DOI: 10.1109/TWC.2005.850272.

[Saq+12] N. Saquib, E. Hossain, L. B. Le, et al. "Interference management in OFDMA femtocell networks: issues and approaches." In: *IEEE Wireless Communications* 19.3 (June 2012), pp. 86–95. ISSN: 1536-1284. DOI: 10.1109/MWC.2012.6231163.

[MM85] J. Massey and P. Mathys. "The collision channel without feedback." In: *IEEE Transactions on Information Theory* 31.2 (1985), pp. 192–204.

[GS87] D. J. Goodman and A. A. M. Saleh. "The near/far effect in local ALOHA radio communications." In: *IEEE Transactions on Vehicular Technology* 36.1 (Feb. 1987), pp. 19–27. ISSN: 0018-9545. DOI: 10.1109/T-VT.1987.24093.

[Str+94] S. Striglis, A. Kaul, N. Yang, et al. "A multistage RAKE receiver for improved capacity of CDMA systems." In: *Proceedings of IEEE Vehicular Technology Conference (VTC)*. IEEE. 1994, pp. 789–793.

[NP12] K. R. Narayanan and H. D. Pfister. "Iterative collision resolution for slotted ALOHA: An optimal uncoordinated transmission policy." In: *2012 7th International Symposium on Turbo Codes and Iterative Information Processing (ISTC)*. Aug. 2012, pp. 136–139. DOI: 10.1109/ISTC.2012.6325214.

[YG05] Y. Yu and G. B. Giannakis. "SICTA: a 0.693 contention tree algorithm using successive interference cancellation." In: *Proc. IEEE Infocom*. Vol. 3. 2005, pp. 1908–1916.

[Aur08] M. Aurelio Locsin Demand. *Is Air Travel Safer Than Car Travel?* [Online; posted at 2008]USA TODAY. 2008. URL: http://traveltips.usatoday.com/air-travel-safer-car-travel-1581.html/.

[TLS04] P. Traverse, I. Lacaze, and J. Souyris. "Airbus fly-by-wire: A total approach to dependability." In: *WCC Building the Information Society* (Aug. 2004), pp. 191–212. ISSN: 18684238. URL: http://link.springer.com/chapter/10.1007/978-1-4020-8157-6%5C_18.

[Sch+07] G. Scheible, D. Dzung, J. Endresen, et al. “Unplugged but connected [Design and implementation of a truly wireless real-time sensor/actuator interface].” In: *IEEE Industrial Electronics Magazine* 1.2 (Summer 2007), pp. 25–34. ISSN: 1932-4529. DOI: 10.1109/MIE.2007.901481.

[Com+10] I. Commission et al. “Industrial Communication Networks-Wireless Communication Network and Communication Profiles-WirelessHART™.” In: *Switzerland: IEC* (2010), p. 944.

[Erg04] S. C. Ergen. “ZigBee/IEEE 802.15. 4 Summary.” In: *UC Berkeley, September* 10 (2004), p. 17.

[36313] 3. T. 36.300. “Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN); Overall Description; Stage 2.” In: (2013).

[All08] W. Alliance. “Ecma-368 high rate ultra wideband phy and mac standard.” In: *ECMA Std* 20 (2008).

[Com+99] I. C. S. L. M. S. Committee et al. “Wireless LAN medium access control (MAC) and physical layer (PHY) specifications.” In: *ANSI/IEEE Std. 802.11-1999* (1999).

[AE10] H. Alemdar and C. Ersoy. “Wireless sensor networks for healthcare: A survey.” In: *Computer networks* 54.15 (2010), pp. 2688–2710.

[Cav+14] R. Cavallari, F. Martelli, R. Rosini, et al. “A survey on wireless body area networks: Technologies and design challenges.” In: *IEEE Communications Surveys & Tutorials* 16.3 (2014), pp. 1635–1657.

[Har02] R. M. Harman. “Wireless solutions for aircraft condition based maintenance systems.” In: *Aerospace Conference Proceedings, 2002. IEEE*. Vol. 6. IEEE. 2002, pp. 6–6.

[ISW12] K. Islam, W. Shen, and X. Wang. “Wireless sensor network reliability and security in factory automation: A survey.” In: *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 42.6 (2012), pp. 1243–1256.

[GH+09] V. C. Gungor, G. P. Hancke, et al. “Industrial wireless sensor networks: Challenges, design principles, and technical approaches.” In: *IEEE Trans. Industrial Electronics* 56.10 (2009), pp. 4258–4265.

[LSS07] J. Lee, Y. Su, and C. Shen. "A Comparative Study of Wireless Protocols: Bluetooth, UWB, ZigBee, and Wi-Fi." In: *IECON 2007 - 33rd Annual Conference of the IEEE Industrial Electronics Society*. Nov. 2007, pp. 46–51. DOI: 10.1109/IECON.2007.4460126.

[WMW05] A. Willig, K. Matheus, and A. Wolisz. "Wireless technology in industrial networks." In: *Proceedings of the IEEE* 93.6 (2005), pp. 1130–1151.

[US 11] F. A. A. U.S. Department of Transportation. *Advisory Circular on System Safety and Assessment for Part 23 Airplanes*. Tech. rep. Nov. 2011.

[BK06] F. Bai and H. Krishnan. "Reliability analysis of DSRC wireless communication for vehicle safety applications." In: *Intelligent Transportation Systems Conference, 2006. ITSC'06. IEEE*. IEEE. 2006, pp. 355–362.

[Fro+14] A. Frotzscher, U. Wetzker, M. Bauer, et al. "Requirements and current solutions of wireless communication in industrial automation." In: *Communications Workshops (ICC), 2014 IEEE International Conference on* (2014), pp. 67–72. DOI: 10.1109/ICCW.2014.6881174.

[Val12] A. E. Vallestad. *WISA becomes WSAN - from proprietary technology to industry standard Outline Factory Automation vs . Process Automation*. Tech. rep. Apr. 2012.

[Sav+13] S. Savazzi, U. Spagnolini, L. Goratti, et al. "Ultra-wide band sensor networks in oil and gas explorations." In: *IEEE Communications Magazine* 51.4 (2013), pp. 150–160. ISSN: 01636804. DOI: 10.1109/MCOM.2013.6495774.

[Fan09] Z. Fan. "Bandwidth allocation in UWB WPANs with ECMA-368 MAC." In: *Computer Communications* 32.5 (2009), pp. 954–960. ISSN: 01403664. DOI: 10.1016/j.comcom.2008.12.024. URL: http://dx.doi.org/10.1016/j.comcom.2008.12.024.

[Lei11] D. F. M. Leipold. "Wireless UWB Aircraft Cabin Communication System." PhD thesis. Technische Universität München, 2011. URL: http://mediatum.ub.tum.de/doc/1079692/1079692.pdf.

[Vie+13] R. Viegas, L. A. Guedes, F. Vasques, et al. "A new MAC scheme specifically suited for real-time industrial communication based on IEEE 802.11e." In: *Computers and Electrical Engineering* 39.6 (2013), pp. 1684–1704. ISSN: 00457906. DOI: 10.1016/j.compeleceng.2012.10.008. URL: http://dx.doi.org/10.1016/j.compeleceng.2012.10.008.

[Che+09] F. Chen, T. Talanis, R. German, et al. “Real-time enabled IEEE 802.15. 4 sensor networks in industrial automation.” In: *Industrial Embedded Systems, 2009. SIES’09. IEEE International Symposium on*. IEEE. 2009, pp. 136–139.

[Dan+13] K. Dang, J. Z. Shen, L. D. Dong, et al. “A graph route-based superframe scheduling scheme in WirelessHART mesh networks for high robustness.” In: *Wireless Personal Communications* 71.4 (2013), pp. 2431–2444. ISSN: 09296212. DOI: 10.1007/s11277-012-0946-2.

[PC09] S. Petersen and S. Carlsen. “Performance evaluation of WirelessHART for factory automation.” In: *ETFA 2009 - 2009 IEEE Conference on Emerging Technologies and Factory Automation* (2009). ISSN: 1946-0759. DOI: 10.1109/ETFA.2009.5346996.

[LA11] A. Lioumpas and A. Alexiou. “Uplink scheduling for Machine-to-Machine communications in LTE-based cellular systems.” In: *Proc. IEEE Globecom Workshops (GC Wkshps)*. Dec. 2011, pp. 353–357. DOI: 10.1109/GLOCOMW.2011.6162470.

[BK12] J. Brown and J. Y. Khan. “Performance comparison of LTE FDD and TDD based Smart Grid communications networks for uplink biased traffic.” In: *2012 IEEE 3rd International Conference on Smart Grid Communications, SmartGridComm 2012* (2012), pp. 276–281. DOI: 10.1109/SmartGridComm.2012.6485996.

[DJ10] O. Delgado and B. Jaumard. “Scheduling and resource allocation in LTE Uplink with a delay requirement.” In: *CNSR 2010 - Proceedings of the 8th Annual Conference on Communication Networks and Services Research* (2010), pp. 268–275. DOI: 10.1109/CNSR.2010.33.

[Kim+14] J. Kim, J. Lee, J. Kim, et al. “M2M service platforms: survey, issues, and enabling technologies.” In: *IEEE Communications Surveys & Tutorials* 16.1 (First Quarter 2014), pp. 61–76.

[LAA14a] A. Laya, L. Alonso, and J. Alonso-Zarate. “Is the random access channel of LTE and LTE-A suitable for M2M communications? A survey of alternatives.” In: *IEEE Communications Surveys & Tutorials* 16.1 (First Quarter 2014), pp. 4–16.

[11] *3GPP TR 37.868: Technical Specification Group Radio Access Network; Study on RAN Improvements for Machine-type Communications*. 2011.

[3rd00] 3rd Generation Partnership Project (3GPP). *RAN WG2 #71 R2-104663: MTC LTE simulations*. Tech. rep. Madrid, Aug. 2010, 2000. URL: http://www.3gpp.org/DynaReport/TDocExMtg--R2-71--28035.htm.

[Lan+13a] M. Laner, P. Svoboda, N. Nikaein, et al. "Traffic Models for Machine Type Communications." In: *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*. Aug. 2013, pp. 1–5.

[MSP14a] G. Madueno, C. Stefanovic, and P. Popovski. "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling." In: *IEEE Wireless Communications Letters* 3.4 (Aug. 2014), pp. 429–432. ISSN: 2162-2337. DOI: 10.1109/LWC.2014.2326674.

[Che+12] R.-G. Cheng, C.-H. Wei, S.-L. Tsao, et al. "RACH collision probability for machine-type communications." In: *Proc. IEEE Vehicular Technology Conference (VTC Spring)*. 2012, pp. 1–5.

[16] *Evolved Univ. Ter. Radio Access (E-UTRA): Medium Access Control (MAC) Protocol Spec., v.13.2.0, Cedex, France, 3GPP TS36.321*. 2016.

[Pop92] B. M. Popovic. "Generalized chirp-like polyphase sequences with optimum correlation properties." In: *IEEE Transactions on Information Theory* 38.4 (1992), pp. 1406–1409.

[Sch83a] F. Schoute. "Dynamic frame length ALOHA." In: *IEEE Transactions on communications* 31.4 (1983), pp. 565–568.

[Mas81] J. L. Massey. "Collision-resolution algorithms and random-access communications." In: *Multi-User Communication Systems*. Springer, 1981, pp. 73–137.

[MF85] P. Mathys and P. Flajolet. "Q-ary collision resolution algorithms in random-access systems with free or blocked channel access." In: *IEEE Transactions on Information Theory* 31.2 (1985), pp. 217–243.

[Dua+16] S. Duan, V. Shah-Mansouri, Z. Wang, et al. "D-ACB: Adaptive Congestion Control Algorithm for Bursty M2M Traffic in LTE Networks." In: *IEEE Trans. Vehicular Techn.* 65.12 (Dec. 2016), pp. 9847–9861.

[WCT12] C.-H. Wei, R.-G. Cheng, and S.-L. Tsao. "Modeling and Estimation of One-Shot Random Access for Finite-User Multichannel Slotted ALOHA Systems." In: *IEEE Commun. Letters* 16.8 (Aug. 2012), pp. 1196–1199.

[WCT13] C.-H. Wei, R.-G. Cheng, and S.-L. Tsao. "Performance analysis of group paging for machine-type communications in LTE networks." In: *IEEE Trans. on Vehicular Technology* 62.7 (Sept. 2013), pp. 3371–3382.

[Che+15] R.-G. Cheng, F. Al-Taee, J. Chen, et al. "A Dynamic Resource Allocation Scheme for Group Paging in LTE-Advanced Networks." In: *IEEE Internet of Things Journal* 2.5 (Oct. 2015), pp. 427–434.

[WBC15a] C.-H. Wei, G. Bianchi, and R.-G. Cheng. “Modeling and Analysis of Random Access Channels With Bursty Arrivals in OFDMA Wireless Networks.” In: *IEEE Transactions on Wireless Communications* 14.4 (Apr. 2015), pp. 1940–1953.

[CLL11] J.-P. Cheng, C.-H. Lee, and T.-M. Lin. “Prioritized Random Access with dynamic access barring for RAN overload in 3GPP LTE-A networks.” In: *Proc. IEEE Globecom Workshops (GC Wkshps)*. Dec. 2011, pp. 368–372. DOI: 10.1109/GLOCOMW.2011.6162473.

[IW08] H. Inaltekin and S. Wicker. “The Analysis of Nash Equilibria of the One-Shot Random-Access Game for Wireless Networks and the Behavior of Selfish Nodes.” In: *IEEE/ACM Transactions on Networking* 16.5 (Oct. 2008), pp. 1094–1107. ISSN: 1063–6692. DOI: 10.1109/TNET.2007.909668.

[ST12] S. Y. Shin and D. Triwicaksono. “Radio resource control scheme for machine-to-machine communication in LTE infrastructure.” In: *Proc. ICT Convergence (ICTC)*. Oct. 2012, pp. 1–6. DOI: 10.1109/ICTC.2012.6386765.

[Pan+14] Y.-C. Pang, S.-L. Chao, G.-Y. Lin, et al. “Network access for M2M/H2H hybrid systems: a game theoretic approach.” In: *IEEE Communications Letters* 18.5 (May 2014), pp. 845–848.

[Jia+13] X. Jian, Y. Jia, X. Zeng, et al. “A novel class-dependent back-off scheme for Machine Type Communication in LTE systems.” In: *Proc. Wireless and Opt. Commun. Conf. (WOCC)*. May 2013, pp. 135–140. DOI: 10.1109/WOCC.2013.6676356.

[OHL15] C.-Y. Oh, D. Hwang, and T.-J. Lee. “Joint Access Control and Resource Allocation for Concurrent and Massive Access of M2M Devices.” In: *IEEE Trans. Wireless Commun.* 14.8 (Aug. 2015), pp. 4182–4192. ISSN: 1536-1276. DOI: 10.1109/TWC.2015.2417873.

[PLW16] Y.-C. Pang, G.-Y. Lin, and H.-Y. Wei. “Context-aware Dynamic Resource Allocation for Cellular M2M Communications.” In: *IEEE Internet of Things Journal* 3.3 (June 2016), pp. 318–326. ISSN: 2327-4662. DOI: 10.1109/JIOT.2015.2496626.

[And+13] S. Andreev, A. Larmo, M. Gerasimenko, et al. “Efficient small data access for machine-type communications in LTE.” In: *Proc. IEEE ICC*. June 2013, pp. 3569–3574. DOI: 10.1109/ICC.2013.6655105.

[Dhi+14] H. S. Dhillon, H. Huang, H. Viswanathan, et al. "Fundamentals of throughput maximization with random arrivals for M2M communications." In: *IEEE Trans. on Communications* 62.11 (Nov. 2014), pp. 4094–4109.

[She+11] S.-T. Sheu, C.-H. Chiu, S. Lu, et al. "Efficient data transmission scheme for MTC communications in LTE system." In: *Proc. ITS Telecommunications (ITST)*. Aug. 2011, pp. 727–731. DOI: 10.1109/ITST.2011.6060150.

[LCL11] S.-Y. Lien, K.-C. Chen, and Y. Lin. "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications." In: *IEEE Communications Magazine* 49.4 (Apr. 2011), pp. 66–74. ISSN: 0163-6804. DOI: 10.1109/MCOM.2011.5741148.

[CZB14] Y. Chang, C. Zhou, and O. Bulakci. "Coordinated Random Access Management for Network Overload Avoidance in Cellular Machine-to-Machine Communications." In: *Proc. Eu. Wireless*. May 2014, pp. 1–6.

[Lin+14] T.-M. Lin, C.-H. Lee, J.-P. Cheng, et al. "PRADA: Prioritized Random Access With Dynamic Access Barring for MTC in 3GPP LTE-A Networks." In: *IEEE Transactions on Vehicular Technology* 63.5 (June 2014), pp. 2467–2472. ISSN: 0018-9545. DOI: 10.1109/TVT.2013.2290128.

[BMG14] L. M. Bello, P. Mitchell, and D. Grace. "Application of Q-Learning for RACH Access to Support M2M Traffic over a Cellular Network." In: *Proc. European Wireless*. May 2014, pp. 1–6.

[Pra+12] N. Pratas, H. Thomsen, C. Stefanovic, et al. "Code-expanded random access for machine-type communications." In: *Proc. IEEE Globecom Workshops (GC Wkshps)*. Dec. 2012, pp. 1681–1686. DOI: 10.1109/GLOCOMW.2012.6477838.

[DSW13] S. Duan, V. Shah-Mansouri, and V. W. S. Wong. "Dynamic Access Class Barring for M2M Communications in LTE Networks." In: *Proc. IEEE Globecom*. Dec. 2013, pp. 4876–4881.

[Ko+12] K. S. Ko, M. J. Kim, K. Y. Bae, et al. "A novel random access for fixed-location machine-to-machine communications in OFDMA based systems." In: *IEEE Communications Letters* 16.9 (Sept. 2012), pp. 1428–1431.

[WW15a] Z. Wang and V. W. Wong. "Optimal Access Class Barring for Stationary Machine Type Communication Devices With Timing Advance Information." In: *IEEE Transactions on Wireless Communications* 14.10 (Oct. 2015), pp. 5374–5387.

[JWD14] W. Jiang, X. Wang, and T. Deng. “Performance Analysis of a Pre-backoff Based Random Access Scheme for Machine-type Communications.” In: *Proc. IEEE Int. Conf. on Intelligent Green Building and Smart Grid (IGBSG)*. 2014, pp. 1–4.

[Har+15] R. Harwahyu, X. Wang, R. F. Sari, et al. “Analysis of group paging with pre-backoff.” In: *EURASIP J. Wireless Communications and Networking* 2015.1 (2015), pp. 1–9.

[DW15] T. Deng and X. Wang. “Performance Analysis of a Device to Device Communication Based Random Access Scheme for Machine Type Communications.” In: *Wireless Personal Communications* 83 (2015), pp. 1251–1272.

[Phu+12] U. Phuyal, A. T. Koc, M.-H. Fong, et al. “Controlling access overload and signaling congestion in M2M networks.” In: *Conference Record of Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. 2012, pp. 591–595.

[Tsa+12] A.-H. Tsai, L.-C. Wang, J.-H. Huang, et al. “Overload Control for Machine Type Communications with Femtocells.” In: *Proc. IEEE Vehicular Technology Conference (VTC Fall)*. Sept. 2012, pp. 1–5. DOI: 10.1109/VTCFall.2012.6399236.

[KHT12] A. Ksentini, Y. Hadjadj-Aoul, and T. Taleb. “Cellular-based machine-to-machine overload control.” In: *IEEE Network* 26.6 (Nov. 2012), pp. 54–60. ISSN: 0890-8044. DOI: 10.1109/MNET.2012.6375894.

[Wu+13] H. Wu, C. Zhu, R. La, et al. “FASA: Accelerated S-ALOHA Using Access History for Event-Driven M2M Communications.” In: *IEEE/ACM Trans. Netw.* 21.6 (Dec. 2013), pp. 1904–1917. ISSN: 1063-6692. DOI: 10.1109/TNET.2013.2241076.

[He+15] H. He, P. Ren, Q. Du, et al. “Estimation Based Adaptive ACB Scheme for M2M Communications.” In: *Wireless Algorithms, Systems, and Applications, LNCS, Vol. 9204*. Springer, 2015, pp. 165–174.

[Lie+12] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, et al. “Cooperative Access Class Barring for Machine-to-Machine Communications.” In: *IEEE Transactions on Wireless Communications* 11.1 (Jan. 2012), pp. 27–32. ISSN: 1536-1276. DOI: 10.1109/TWC.2011.111611.110350.

[WD14] P. Wali and D. Das. “A novel access scheme for IoT communications in LTE-Advanced network.” In: *Proc. IEEE Int. Conf. on Advanced Networks and Telecommuncations Systems (ANTS)*. Dec. 2014, pp. 1–6. DOI: 10.1109/ANTS.2014.7057255.

[VB04] B. Van Houdt and C. Blondia. "Robustness of Q-ary collision resolution algorithms in random access systems." In: *Performance Evaluation* 57.3 (July 2004), pp. 357–377.

[CS88] I. Cidon and M. Sidi. "Conflict multiplicity estimation and batch resolution algorithms." In: *IEEE Trans. Information Theory* 34.1 (Jan. 1988), pp. 101–110.

[PFP05] P. Popovski, F. Fitzek, and R. Prasad. "A class of algorithms for batch conflict resolution algorithms with multiplicity estimation." In: *Algorithmica, Springer-Verlag* 49.4 (2005), pp. 286–317.

[Zan12a] A. Zanella. "Adaptive batch resolution algorithm with deferred feedback for wireless systems." In: *IEEE Trans. Wireless Communications* 11.10 (Oct. 2012), pp. 3528–3539.

[MSP14b] G. C. Madueno, S. Stefanovic, and P. Popovski. "Efficient LTE Access with Collision Resolution for Massive M2M Communications." In: *Proc. IEEE Globecom Workshops (GC Wkshps)*. Dec. 2014, pp. 1433–1438.

[VAA16] F. Vázquez-Gallego, L. Alonso, and J. Alonso-Zarate. "Energy harvesting-aware contention tree-based access for wireless Machine-to-Machine networks." In: *Proc. IEEE ICC*. June 2016, pp. 1–6.

[Cap79a] J. Capetanakis. "Tree algorithms for packet broadcast channels." In: *IEEE Transactions on Information Theory* 25.5 (1979), pp. 505–515. DOI: 10.1109/TIT.1979.1056093.

[Gal78] R. G. Gallager. "Conflict resolution in random access broadcast networks." In: *Proc. of the AFOSR Workshop in Communication Theory and Applications*. 1978, pp. 74–76.

[PVB07] G. T. Peeters, B. Van Houdt, and C. Blondia. "A multiaccess tree algorithm with free access, interference cancellation and single signal memory requirements." In: *Performance Evaluation* 64.9 (Oct. 2007), pp. 1041–1052.

[YG07] Y. Yu and G. B. Giannakis. "High-throughput random access using successive interference cancellation in a tree algorithm." In: *IEEE Trans. Information Theory* 53.12 (Dec. 2007), pp. 4628–4639.

[WC15] D. T. Wiriaatmadja and K. W. Choi. "Hybrid Random Access and Data Transmission Protocol for Machine-to-Machine Communications in Cellular Networks." In: *IEEE Transactions on Wireless Communications* 14.1 (Jan. 2015), pp. 33–46.

[Liu+14] Y. Liu, C. Yuen, X. Cao, et al. “Design of a scalable hybrid MAC protocol for heterogeneous M2M networks.” In: *IEEE Internet of Things Journal* 1.1 (Feb. 2014), pp. 99–111.

[Ver+16] P. K. Verma, R. Verma, A. Prakash, et al. “Throughput-Delay Evaluation of a Hybrid-MAC Protocol for M2M Communications.” In: *Int. Journal of Mobile Computing and Multimedia Commun. (IJMCMC)* 7.1 (2016), pp. 41–60.

[Yu+15] X. Yu, P. Navaratnam, K. Moessner, et al. “Distributed Admission Control with Soft Resource Allocation for Hybrid MAC in Home M2M Networks.” In: *Proc. IEEE Vehicular Techn. Conf. (VTC Spring)*. 2015, pp. 1–5.

[Ste+13] C. Stefanovic, K. F. Trilingsgaard, N. K. Pratas, et al. “Joint estimation and contention-resolution protocol for wireless random access.” In: *Communications (ICC), 2013 IEEE International Conference on*. IEEE. 2013, pp. 3382–3387.

[SLP17] C. Stefanovic, F. Lazaro, and P. Popovski. “Frameless ALOHA with Reliability-Latency Guarantees.” In: *arXiv preprint arXiv:1709.02177* (2017).

[Vil+18a] M. Vilgelm, S. Schiessl, H. Al-Zubaidy, et al. “On the Reliability of LTE Random Access: Performance Bounds for Machine-to-Machine Burst Resolution Time.” In: *IEEE International Conference on Communications*. May 2018.

[Jia+17a] X. Jian, Y. Liu, Y. Wei, et al. “Random access delay distribution of multichannel Slotted ALOHA with its applications for machine type communications.” In: *IEEE Internet of Things Journal* 4.1 (2017), pp. 21–28.

[Jac+17] T. Jacobsen, R. Abreu, G. Berardinelli, et al. “System level analysis of uplink grant-free transmission for URLLC.” In: *Globecom Workshops (GC Wkshps), 2017 IEEE*. IEEE. 2017, pp. 1–6.

[Kot+18] R. Kotaba, C. N. Manchón, T. Balercia, et al. “Uplink Transmissions in URLLC Systems with Shared Diversity Resources.” In: *IEEE Wireless Communications Letters* (2018).

[SC17] S. Saur and M. Centenaro. “Radio access protocols with multi-user detection for urllc in 5g.” In: *European Wireless 2017; 23th European Wireless Conference; Proceedings of*. VDE. 2017, pp. 1–6.

[Sin+18] B. Singh, O. Tirkkonen, Z. Li, et al. “Contention-based access for ultra-reliable low latency uplink transmissions.” In: *IEEE Wireless Communications Letters* 7.2 (2018), pp. 182–185.

[Abb+17] R. Abbas, M. Shirvanimoghaddam, Y. Li, et al. "Random Access for M2M Communications With QoS Guarantees." In: *IEEE Transactions on Communications* 65.7 (July 2017), pp. 2889–2903. ISSN: 0090-6778. DOI: 10.1109/TCOMM.2017.2690900.

[KLE95] M. J. Karol, Z. Liu, and K. Y. Eng. "Distributed-queueing request update multiple access (DQRUMA) for wireless packet (ATM) networks." In: *Communications, 1995. ICC'95 Seattle,'Gateway to Globalization', 1995 IEEE International Conference on*. Vol. 2. IEEE. 1995, pp. 1224–1231.

[Lay+16] A. Laya, C. Kalalas, F. Vazquez-Gallego, et al. "Goodbye, aloha!" In: *IEEE access* 4 (2016), pp. 2029–2044.

[LTE15] LTE-U Forum. "Coexistence Study for LTE-U SDL." In: (2015).

[Lan+13b] M. Laner, P. Svoboda, N. Nikaein, et al. "Traffic Models for Machine Type Communications." In: *Proc. Int. Symp. Wireless Commun. Systems (ISWCS)*. Aug. 2013, pp. 1–5.

[Bil12] P. Billingsley. *Probability and Measure*. 3rd Edition. John Wiley & Sons, 2012.

[MP93a] M. L. Molle and G. C. Polyzos. *Conflict resolution algorithms and their performance analysis*. Tech. rep. Tec. Rep. Dep. of Computer Science and Engineering, Univ. of California at San Diego, LaJolla, 1993.

[Tya+17] R. R. Tyagi, F. Aurzada, K.-D. Lee, et al. "Connection establishment in LTE-A networks: Justification of Poisson process modeling." In: *IEEE Systems Journal* 11.4 (2017), pp. 2383–2394.

[LAA14b] A. Laya, L. Alonso, and J. Alonso-Zarate. "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives." In: *IEEE Communications Surveys and Tutorials* 16.1 (2014), pp. 4–16.

[Mad+16] G. C. Madueno, J. J. Nielsen, D. M. Kim, et al. "Assessment of LTE Wireless Access for Monitoring of Energy Distribution in the Smart Grid." In: *IEEE Journal on Selected Areas in Communications* 34.3 (Mar. 2016), pp. 675–688.

[Eri+18] M. Ericson, P. Spapis, M. Säily, et al. "Initial Access, RRC and Mobility." In: *5G System Design: Architectural and Functional Considerations and Long Term Research* (2018), pp. 367–407.

[Pop+19] P. Popovski, Č. Stefanović, J. J. Nielsen, et al. "Wireless Access in Ultra-Reliable Low-Latency Communication (URLLC)." In: *IEEE Transactions on Communications* (2019).

[18] *3GPP RP-181477: SID on Physical Layer Enhancements for NR URLLC; NR eURLLC L1*. 2018.

[GSP18] J. Goseling, C. Stefanovic, and P. Popovski. "Sign-Compute-Resolve for Tree Splitting Random Access." In: *IEEE Trans. Info. Theory* 64.7 (July 2018), pp. 5261–5276.

[MDA17] A. Mengali, R. De Gaudenzi, and P.-D. Arapoglou. "Enhancing the physical layer of contention resolution diversity slotted ALOHA." In: *IEEE Transactions on Communications* 65.10 (2017), pp. 4295–4308.

[Mag+18] D. Magrin, C. Pielli, C. Stefanovic, et al. "Enabling LTE RACH collision multiplicity detection via machine learning." In: *arXiv preprint arXiv:1805.11482* (2018).

[KN06] M. Kodialam and T. Nandagopal. "Fast and reliable estimation schemes in RFID systems." In: *Proceedings of the 12th annual international conference on Mobile computing and networking*. ACM. 2006, pp. 322–333.

[AOR03] I. Adler, S. Oren, and S. M. Ross. "The coupon-collector's problem revisited." In: *Journal of Applied Probability* 40.2 (2003), pp. 513–518.

[Zan12b] A. Zanella. "Estimating Collision Set Size in Framed Slotted Aloha Wireless Networks and RFID Systems." In: *IEEE Communications Letters* 16.3 (Mar. 2012), pp. 300–303. ISSN: 1089-7798. DOI: 10.1109/LCOMM.2012.011312.112067.

[Dem+09] I. Demirkol, C. Ersoy, F. Alagöz, et al. "The impact of a realistic packet traffic model on the performance of surveillance wireless sensor networks." In: *Computer Networks* 53.3 (2009), pp. 382–399.

[3GP12] 3GPP. *Tech. Rep. 36.888: Study on provision of low-cost MTC UEs based on LTE*. Tech. rep. Valbonne, France, 2012.

[Gim65] L. Gimpelson. "Analysis of mixtures of wide-and narrow-band traffic." In: *IEEE Transactions on Communication Technology* 13.3 (1965), pp. 258–266.

[WW15b] Z. Wang and V. W. Wong. "Optimal access class barring for stationary machine type communication devices with timing advance information." In: *IEEE Transactions on Wireless communications* 14.10 (2015), pp. 5374–5387.

[Kra69] A. Kramer. "Improving communication reliability by use of an intermittent feedback channel." In: *IEEE Transactions on Information Theory* 15.1 (Jan. 1969), pp. 52–60. ISSN: 0018-9448. DOI: 10.1109/TIT.1969.1054272.

[GN10] R. G. Gallager and B. Nakiboğlu. "Variations on a Theme by Schalkwijk and Kailath." In: *IEEE Transactions on Information Theory* 56.1 (Jan. 2010), pp. 6–17. ISSN: 0018-9448. DOI: 10.1109/TIT.2009.2034896.

[SDG05] A. Sahai, S. C. Draper, and M. Gastpar. "Boosting reliability over AWGN networks with average power constraints and noiseless feedback." In: *Proceedings. International Symposium on Information Theory, 2005. ISIT 2005.* Sept. 2005, pp. 402–406. DOI: 10.1109/ISIT.2005.1523364.

[SG10] A. Sarwate and M. Gastpar. "A little feedback can simplify sensor network cooperation." In: *IEEE Journal on Selected Areas in Communications* 28.7 (Sept. 2010), pp. 1159–1168. ISSN: 0733-8716. DOI: 10.1109/JSAC.2010.100920.

[Abr70] N. Abramson. "THE ALOHA SYSTEM: another alternative for computer communications." In: *Proceedings of the November 17-19, 1970, fall joint computer conference*. ACM. 1970, pp. 281–285.

[KL75] L. Kleinrock and S. Lam. "Packet Switching in a Multiaccess Broadcast Channel: Performance Evaluation." In: *IEEE Transactions on Communications* 23.4 (Apr. 1975), pp. 410–423. ISSN: 0090-6778. DOI: 10.1109/TCOM.1975.1092814.

[Sch83b] F. Schoute. "Dynamic Frame Length ALOHA." In: *IEEE Transactions on Communications* 31.4 (Apr. 1983), pp. 565–568. ISSN: 0090-6778. DOI: 10.1109/TCOM.1983.1095854.

[Bin00] B. Bing. "Stabilization of the randomized slotted ALOHA protocol without the use of channel feedback information." In: *IEEE Communications Letters* 4.8 (Aug. 2000), pp. 249–251. ISSN: 1089-7798. DOI: 10.1109/4234.864184.

[Kos16] M. Koseoglu. "Lower Bounds on the LTE-A Average Random Access Delay Under Massive M2M Arrivals." In: *IEEE Transactions on Communications* 64.5 (May 2016), pp. 2104–2115. ISSN: 0090-6778. DOI: 10.1109/TCOMM.2016.2550526.

[Cap79b] J. Capetanakis. "Tree algorithms for packet broadcast channels." In: *IEEE transactions on information theory* 25.5 (1979), pp. 505–515.

[HB85] J.-C. Huang and T. Berger. "Delay analysis of interval-searching contention resolution algorithms." In: *IEEE Transactions on Information Theory* 31.2 (1985), pp. 264–273.

[KG85] M. Kaplan and E. Gulko. "Analytic properties of multiple-access trees." In: *IEEE Transactions on Information Theory* 31.2 (1985), pp. 255–263.

[JJ00] A. J. Janssen and M. de Jong. “Analysis of contention tree algorithms.” In: *IEEE Transactions on Information Theory* 46.6 (2000), pp. 2163–2172.

[GVS88] S. Ghez, S. Verdu, and S. C. Schwartz. “Stability properties of slotted Aloha with multipacket reception capability.” In: *IEEE transactions on automatic control* 33.7 (1988), pp. 640–649.

[BM82] T. Berger and N. Mehravari. “Conflict resolution protocols for random multiple-access channels with binary feedback.” In: *IN: Annual Allerton Conference on Communication, Control, and Computing, 19th, Monticello, IL, September 30-October 2, 1981, Proceedings (A83-35101 15-63). Urbana, IL, University of Illinois, 1982, p. 404-411.* 1982, pp. 404–411.

[Jia+17b] X. Jian, Y. Liu, Y. Wei, et al. “Random access delay distribution of multichannel Slotted ALOHA with its applications for machine type communications.” In: *IEEE Internet of Things Journal* 4.1 (2017), pp. 21–28.

[Vil+18b] M. Vilgelm, S. Schiessl, H. Al-Zubaidy, et al. “On the Reliability of LTE Random Access: Performance Bounds for Machine-to-Machine Burst Resolution Time.” In: *2018 IEEE International Conference on Communications (ICC)*. IEEE. 2018, pp. 1–7.

[Cho85] J. H. Cho. *Multichannel Collision Resolution Algorithm.* Tech. rep. NAVAL POSTGRADUATE SCHOOL MONTEREY CA, 1985.

[JJ+00] A. Janssen, M. J. de Jong, et al. “Analysis of contention tree algorithms.” In: *IEEE Transactions on Information Theory* 46.6 (2000), pp. 2163–2172.

[MP93b] M. L. Molle and G. C. Polyzos. “Conflict resolution algorithms and their performance analysis.” In: *University of Toronto, CS93-300, Tech. Rep* (1993).

[LLS00] C. Law, K. Lee, and K.-Y. Siu. “Efficient memoryless protocol for tag identification.” In: *Proceedings of the 4th international workshop on Discrete algorithms and methods for mobile computing and communications*. ACM. 2000, pp. 75–84.

[CLL07] J. H. Choi, D. Lee, and H. Lee. “Query tree-based reservation for efficient RFID tag anti-collision.” In: *IEEE Communications Letters* 11.1 (2007).

[WBC15b] C.-H. Wei, G. Bianchi, and R.-G. Cheng. “Modeling and Analysis of Random Access Channels With Bursty Arrivals in OFDMA Wireless Networks.” In: *IEEE Transactions on Wireless Communications* 14.4 (2015), pp. 1940–1953.

[ElS+18] H. El-Sayed, S. Sankar, M. Prasad, et al. "Edge of things: the big picture on the integration of edge, IoT and the cloud in a distributed computing environment." In: *IEEE Access* 6 (2018), pp. 1706–1717.

[Liv11] G. Liva. "Graph-based analysis and optimization of contention resolution diversity slotted ALOHA." In: *IEEE Transactions on Communications* 59.2 (2011), pp. 477–487.

[SPV12] C. Stefanovic, P. Popovski, and D. Vukobratovic. "Frameless ALOHA protocol for wireless networks." In: *IEEE Communications Letters* 16.12 (2012), pp. 2087–2090.

[BVT18a] C. Boyd, R. Vehkalahti, and O. Tirkkonen. "Interference Cancelling Codes for Ultra-Reliable Random Access." In: *International Journal of Wireless Information Networks* 25.4 (Dec. 2018), pp. 422–433. DOI: 10.1007/s10776-018-0411-6. URL: https://doi.org/10.1007/s10776-018-0411-6.

[Kum+11] R. Kumar, T. F. La Porta, G. Maselli, et al. "Interference cancellation-based RFID tags identification." In: *Proceedings of the 14th ACM international conference on Modeling, analysis and simulation of wireless and mobile systems*. ACM. 2011, pp. 111–118.

[BVT17] C. Boyd, R. Vehkalahti, and O. Tirkkonen. "Combinatorial code designs for ultra-reliable IoT random access." In: *Personal, Indoor, and Mobile Radio Communications (PIMRC), 2017 IEEE 28th Annual International Symposium on*. IEEE. 2017, pp. 1–5.

[VP16] X. Vilajosana and K. Pister. *Minimal 6TiSCH Configuration*. Internet-Draft. Work in Progress. Internet Engineering Task Force, Mar. 2016. 28 pp. URL: https://tools.ietf.org/html/draft-ietf-6tisch-minimal-15.

[Wat+12] T. Watteyne, X. Vilajosana, B. Kerkez, et al. "OpenWSN: a standards-based low-power wireless development environment." In: *Transactions on Emerging Telecommunications Technologies* 23.5 (2012), pp. 480–493.

[Bla+15] J. Blanckenstein, C. Nardin, J. Klaue, et al. "Error characterization of multi-access point WSNs in an aircraft cabin." In: *Communication Workshop (ICCW), 2015 IEEE International Conference on*. IEEE. 2015, pp. 2363–2368.

[Gon+12] A. Gonga, O. Landsiedel, P. Soldati, et al. "Revisiting multi-channel communication to mitigate interference and link dynamics in wireless sensor networks." In: *Distributed Computing in Sensor Systems (DCOSS), 2012 IEEE 8th International Conference on*. IEEE. 2012, pp. 186–193.

[DR12] P. Du and G. Roussos. “Adaptive time slotted channel hopping for wireless sensor networks.” In: *Computer Science and Electronic Engineering Conference (CEEC), 2012 4th*. IEEE. 2012, pp. 29–34.

[Zho+06] G. Zhou, T. He, S. Krishnamurthy, et al. “Models and solutions for radio irregularity in wireless sensor networks.” In: *ACM Transactions on Sensor Networks (TOSN)* 2.2 (2006), pp. 221–262.

[TBG11] P. Trenkamp, M. Becker, and C. Goerg. “Wireless sensor network platforms Datasheets versus measurements.” In: *Local Computer Networks (LCN), 2011 IEEE 36th Conference on*. IEEE. 2011, pp. 966–973.

[ZK07] M. Z. Zamalloa and B. Krishnamachari. “An analysis of unreliability and asymmetry in low-power wireless links.” In: *ACM Transactions on Sensor Networks (TOSN)* 3.2 (2007), p. 7.

[HAW08] D. Halperin, T. Anderson, and D. Wetherall. “Taking the sting out of carrier sense: interference cancellation for wireless LANs.” In: *Proceedings of the 14th ACM international conference on Mobile computing and networking*. ACM. 2008, pp. 339–350.

[Lv+11] S. Lv, W. Zhuang, X. Wang, et al. “Scheduling in wireless ad hoc networks with successive interference cancellation.” In: *2011 Proceedings IEEE INFOCOM*. IEEE. 2011, pp. 1287–1295.

[GK08] S. Gollakota and D. Katabi. “Zigzag Decoding: Combating Hidden Terminals in Wireless Networks.” In: *SIGCOMM Comput. Commun. Rev.* 38.4 (Aug. 2008), pp. 159–170. ISSN: 0146-4833. DOI: 10.1145/1402946.1402977. URL: http://doi.acm.org/10.1145/1402946.1402977.

[KL15] L. Kong and X. Liu. “mZig: Enabling Multi-Packet Reception in ZigBee.” In: *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*. MobiCom ’15. Paris, France: ACM, 2015, pp. 552–565. ISBN: 978-1-4503-3619-2. DOI: 10.1145/2789168.2790104. URL: http://doi.acm.org/10.1145/2789168.2790104.

[Web+07] S. P. Weber, J. G. Andrews, X. Yang, et al. “Transmission Capacity of Wireless Ad Hoc Networks With Successive Interference Cancellation.” In: *IEEE Transactions on Information Theory* 53.8 (Aug. 2007), pp. 2799–2814. ISSN: 0018-9448. DOI: 10.1109/TIT.2007.901153.

[Arn80] J.-F. Arnaud. “Frequency planning for broadcast services in Europe.” In: *Proceedings of the IEEE* 68.12 (1980), pp. 1515–1522.

[Ang+11] V. Angelakis, S. Papadakis, V. A. Siris, et al. “Adjacent channel interference in 802.11 a is harmful: Testbed validation of a simple quantification model.” In: *IEEE Communications Magazine* 49.3 (2011).

[Hel04] B. E. Helvik. “Perspectives on the Dependability of Networks and Services.” In: *Telektronikk (100th Anniversary Issue: Perspectives in telecommunications)* 3 (2004), pp. 27–44.

[Sta] I. Standards. “IEEE Standard for Local and metropolitan area networks–Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 1: MAC sublayer.” In: *IEEE Std 802.15.4e* (). DOI: 10.1109/IEEESTD.2012.6185525.

[Wan+10] X. Wang, X. Wang, G. Xing, et al. “Dynamic duty cycle control for end-to-end delay guarantees in wireless sensor networks.” In: *Quality of Service (IWQoS), 2010 18th International Workshop on*. IEEE. 2010, pp. 1–9.

[DZG16] F. Dobslaw, T. Zhang, and M. Gidlund. “End-to-End Reliability-Aware Scheduling for Wireless Sensor Networks.” In: *IEEE Transactions on Industrial Informatics* 12.2 (Apr. 2016), pp. 758–767. ISSN: 1551-3203. DOI: 10.1109/TII.2014.2382335.

[3GP18] 3GPP. *Discussion on the reliability enhancement for grant-free transmission*. Tech. rep. Aug. 2018.

[CH02] H. Cheon and D. Hong. “Effect of channel estimation error in OFDM-based WLAN.” In: *IEEE Communications Letters* 6.5 (2002), pp. 190–192.

[Wu+12] K. Wu, H. Tan, H. Ngan, et al. “Chip error pattern analysis in IEEE 802.15. 4.” In: *IEEE Transactions on Mobile Computing* 11.4 (2012), pp. 543–552.

[Zhe+11] G. Zheng, D. Han, R. Zheng, et al. “A link quality inference model for ieee 802.15. 4 low-rate wpans.” In: *IEEE Global Telecommunications Conference*. IEEE. 2011, pp. 1–6.

[SA98] M. K. Simon and M.-S. Alouini. “A unified approach to the performance analysis of digital communication over generalized fading channels.” In: *Proceedings of the IEEE* 86.9 (1998), pp. 1860–1877.

[BVT18b] C. Boyd, R. Vehkalahti, and O. Tirkkonen. “Interference Cancelling Codes for Ultra-Reliable Random Access.” In: *International Journal of Wireless Information Networks* 25.4 (2018), pp. 422–433.

List of Figures

List of Tables