Technische Universität München
Fakultät für Elektrotechnik und Informationstechnik
Lehrstuhl für Kommunikationsnetze

Random Access Protocols for Massive and Reliable Machine-to-Machine Communication

Mikhail Vilgelm, 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:		Prof. Dr. rer. nat. Thomas Hamacher
Prüfer der Dissertation:	1.	Prof. Dr.-Ing. Wolfgang Kellerer
	2.	Prof. Petar Popovski, Ph.D.

Die Dissertation wurde am 30.08.2018 bei der Technischen Universität München eingereicht und durch die Fakultät für Elektrotechnik und Informationstechnik am 10.01.2019 angenommen.

Abstract

The technological landscape for the firth generation (5G) wireless networks is evolving towards three directions: to increase mobile broadband capacity, to accommodate massive Machine-to-Machine (M2M) devices, and to enable ultra reliable Machine-to-Machine (uM2M). The first direction has been a conventional way of evolution from 2G to 3G and further to 4G networks. In contrast to that, the latter two directions are truly novel for wireless communication. They are inspired by emerging applications such as smart grids, industrial automation, or vehicular-to-X communication. M2M applications significantly differ from conventional human-to-human applications in their communication patterns and service requirements, creating a unique set of challenges for the full communication stack. Efficiency and overhead of medium access control protocols, regulating access to the shared wireless medium, are among these challenges. While the density of M2M devices is expected to be very high, individual devices communicate infrequency and with small amounts of data. This parameter constellation renders most of the traditional access reservation protocols, such as dynamic scheduling in LTE and 5G, inefficient. To this end, a group of random access protocols has re-gained attention, since they allow scalable and low-overhead coordination of the access to wireless medium.

In this thesis, we revisit random access protocols with a resource- and application-centric approach. We attempt to bring random access up-to-date by using the advanced resource management capabilities of modern cellular networks. On the same time, we introduce new application-aware performance metrics, to make random access protocols ready for M2M. We consider LTE and 5G Random Access Procedure (RAP) as a starting point. In order to comprehensively characterize its performance, we separate the study into two distinct regions: steady state, reflecting reactions to average-to-long term changes; and transient state, reflecting reactions to short term changes in the network.

First, for the steady state region, we address the question of efficient resource allocation. We devise a load-adaptive resource allocation policy, which allows service differentiation while maximizing the throughput and decreasing the request drop probability of the prioritized class. To further extend the approach to resource-limited scenarios, we propose random access with spatial aggregation and introduce an analytical model for its performance analysis.

Next, we investigate the transient performance region. We show that the resource consumption model of classical ALOHA-like random access protocols does not generalize well to modern RAP. Instead, we suggest an alternative resource consumption model, and, based on it, we develop a framework for bi-objective resource-aware optimization of the protocol. We apply the framework to devise two Pareto-optimal dynamic burst

resolution algorithms.

We further direct our attention to the requirements of ultra-reliable M2M applications, which brings our study beyond the notion of *expected performance*. Instead, we evaluate *reliability* of random access protocols by the means of a novel methodology based on stochastic network calculus. The methodology allows to derive probabilistic latency-constrained performance bounds for common random access protocols.

Finally, we investigate a detailed model of a large class of M2M applications, networked control systems: control loops with feedback indirectly coupled via shared communication medium. We model event-triggered traffic patterns of control systems, and, using this model, we evaluate the impact of random access protocols on the control performance. In a cross-layer design framework, we develop adaptive random access protocols which incorporate control systems' performance as objectives.

Kurzfassung

Die drahtlosen Mobifunknetze der fünften Generation (5G) werden in drei Richtungen entwickelt: Erhöhung der mobilen Breitbandkapazität, Unterstützung der massive Maschine-zu-Maschine (M2M) Kommunikation, und Ermöglichung der M2M Kommunikation mit ultra- niedriger Latenz und hoher Zuverlässigkeit. Die erste Richtung entspricht dem bekannten Evolutionspfad der drahtlosen Netze von 2G zu 3G und weiter zu 4G. Im Gegensatz dazu sind die beiden anderen Entwicklungsrichtungen neuartig für die drahtlose Kommunikation. Diese Evolutionsrichtungen sind getrieben von neu aufkommenden Anwendungen, wie zum Beispiel intelligenten Energienetzen, Industrieautomatisierung oder Fahrzeug-zu-X Kommunikation. M2M Anwendungen unterscheiden sich durch ihre Kommunikationsmuster und Serviceanforderderungen stark von der klassischen Mensch-zu-Mensch Kommunikation. Dadurch entstehen einzigartige Herausforderungen für alle Schichten der Kommunikationsprotokolle. Die Steigerung der Effizienz und die Reduzierung des Kommunikationsoverheads der genutzten Protokolle für das Medienzugriffsverfahren sind ein Teil dieser Herausforderungen. Es ist zu erwarten, dass die Anzahl von M2M Geräten in 5G Netzen sehr hoch sein wird, während die individuellen Geräte die Daten aber nur selten und in geringen Mengen übertragen. Solche Verkehrsmuster führen dazu, dass der Einsatz der konventionellen Medienszugriffsprotokolle, so wie die dynamische Zeitplanerstellung in LTE Netzen, ineffizient sein wird. Demzufolge haben die Verfahren des zufallsbasierten Medienzugriffs wieder an Aufmerksamkeit gewonnen, da diese eine skalierbare Koordination des Zugriffs auf ein drahtloses Medium mit geringem Overhead ermöglichen.

In dieser Doktorarbeit werden zufallsbasierte Medienzugriffsverfahren mit einem ressourcen- und anwendungsorientierten Ansatz erforscht. Unser Ziel ist die zufallsbasierten Protokolle mithilfe erweiterter Ressourcenverwaltungsfunktionen moderner Mobilfunknetze zu verbessern. Gleichzeitig führen wir neue anwendungsbezogene Leistungsmetriken ein, um diese Protokolle für M2M vorzubereiten. Wir nehmen LTE und 5G Random Access Procedure (RAP) als Ausgangspunkt der Forschung an. Um deren Leistung umfassend zu charakterisieren, unterteilen wir die Untersuchung in zwei unterschiedliche Zustände: Den Gleichgewichtszustand, der Reaktionen auf mittel- bis langfristige Veränderungen beschreibt; und den transienten Zustand, der Reaktionen auf kurzfristige Änderungen im Netzwerk beschreibt.

Als Erstes betrachten wir die Frage der effizienten Ressourcenzuweisung im Gleichgewichtszustand. Wir entwickeln ein lastadaptives Ressourcenzuweisungsverfahren, das die Unterscheidung zwischen Dienstklassen ermöglicht und gleichzeitig den Durchsatz maximiert und die Blockierwahrscheinlichkeit der priorisierten Klasse verringert. Für die Anwendung in Szenarien mit begrenzten Ressourcen erweitern wir unser Ver-

fahren um räumliche Aggregation und führen ein neues analytisches Modell für dessen Leistungbewertung ein.

Anschließend untersuchen wir den transienten Zustand. Wir zeigen, dass das Ressourcenverbrauchsmodell der klassischen ALOHA-ähnlichen Medienzugriffsprotokolle nicht für modernes RAP geeignet ist. Wir schlagen ein alternatives Ressourcenverbrauchsmodell vor und, basierend darauf, entwickeln ein Framework für eine bikriterielle ressourcenbewusste Optimierung des Protokolls. Als Anwendungsbeispiele für das neue Framework entwickeln wir zwei Pareto-optimale dynamische Verfahren für die Auflösung einer Überlast im Kommunikationsmedium.

Um auch die Anforderungen kritischer M2M Anwendungen hinsichtlich ihrer Zuverlässigkeit zu berücksichtigen, gehen wir über die Erwartungswerte hinaus. Stattdessen analysieren wir die Zuverlässigkeit von zufallsbasierten Protokollen mit Hilfe einer neuartigen Methodik, die auf der stochastischen Verkehrstheorie basiert. Die Methode ermöglicht die Ableitung von probabilistischen Leistungsgrenzen für häufig verwendete zufallsbasierter Medienzugriffsprotokolle.

Abschließend untersuchen wir ein detailliertes Modell einer umfangreichen Klasse von M2M-Anwendungen, den vernetzten Regelungssystemen: Regelkreise mit indirekter Kopplung über ein gemeinsames Kommunikationsmedium. Wir modellieren ereignisgesteuerte Verkehrsmuster von Regelungssystemen und bewerten anhand dieses Modells die Auswirkung von zufallsbasierten Protokollen auf die Regelleistung. In einem schichtübergreifenden Design-Framework entwickeln wir adaptive zufallsbasierte Protokolle, bei denen die Leistung von Regelungssystemen als Metrik berücksichtigt wird.

Contents

Chapter 1

Introduction

Machine-to-Machine (M2M) is a term referring to a broad class of applications, where the endpoints are exchanging information without or with limited intervention of humans [Boc+16; GJ15]. With the saturating revenues of the telecommunications providers, M2M applications are a promising new source of income and innovations for the industry [Wu+11]. Deploying communication infrastructure with M2M support enables a wide range of novel applications in such areas as industrial automation, autonomous driving, or smart grids. Typical application examples are process automation in industrial facilities [XHL14], teleoperation [Con+16a], and smart metering [Gun+11; KRR16; Zha+12]. From the perspective of 5G networks [1] [Boc+14], M2M communications are further classified into two categories. The first category is massive Machine-to-Machine (mM2M), characterized by a significantly larger density of connected devices of up to 10^6 per mobile cell. mM2M devices are often assumed to be delay- and packet loss-tolerant, with rare transmissions from individual devices. Examples of mM2M applications are smart metering [Ara+13], vending machines, or remote monitoring of non-critical facilities [Oss+14]. The second category are ultra reliable Machine-to-Machine (uM2M) applications [Boc+14]. They are characterized by stringent requirements on the underlying communication links with availability up to 99.99999 % and low latency down to one millisecond [Aij+17; Pop14], but on the same time with small amounts of transmitted data and sporadic activity. Examples of uM2M applications are teleoperation [Aij+17], industrial automation [XHL14], and in-cabin airplane sensing [G¨+17b].

The characteristics and the requirements of M2M are highly contrasting with Human-to-Human (H2H) applications, i.e. traditional services such as voice calls, video streaming, or web browsing. Modern communication networks are primarily designed for H2H applications, with bursty and data rate hungry user sessions but often with relaxed delay- and packet-loss application requirements. To make 5G networks ready for M2M applications, traditional data rate centric approach must be adapted to explicitly consider other metrics. Naturally, this requires novel concepts and approaches in the whole protocol stack [Boc+16; Gaz17]. The strive for low complexity and low cost hardware is driving the physical layer developments. Efficiency, scalability, and interoperability are the main drivers for the development of higher layers protocols.

[1] In 3GPP terminology, M2M communication is often referred to as Machine Type Communications (MTC). In this thesis, we used both terms interchangeably.

On the Medium Access Control (MAC) layer, large number of devices in the network, sporadic transmission patterns, and small amounts of data transmitted per user emphasize the problem of efficient coordination of access to the shared wireless medium. Instead of a typical schedule-based access coordination, common in LTE, the group of *random access protocols* is of potential interest for M2M communications, both in the context of data transmission (user plane) as well as in the context of associated signaling procedures (control plane). Random access protocols allow efficient low-overhead access coordination in many M2M scenarios, when the instantaneous amount of active devices is low compared to the overall population, and the exact identity of active devices is unknown to the network. On the other hand, the stochastic nature of random access and the lack of determinism do not cope well with the other M2M scenarios, where the latency must be guaranteed with high reliability. In high load scenarios, the random access protocols are notorious for causing high delays due to an excessive amount of collisions.

In the thesis, we revisit the random access protocols and their applications to 5G networks. We take a *resource-centered approach* to random access, exploiting the fine-granular resource management capabilities of modern cellular networks. Unlike in the early days of random access where time was the only resource dimension, modern networks allow multiplexing the devices in frequency or code domains, thereby complicating the problem of resource allocation and expanding it to multiple dimensions. At the same time, we study random access protocols from an *application perspective*, by considering specific scenarios, traffic models, and requirements of both mM2M and uM2M. The novel outcomes of the thesis are: (1) methodologies for performance analysis and optimization random access protocols; (2) novel enhanced random access protocols targeting M2M-specific use cases and performance metrics; (3) comprehensive evaluations of random access protocols in M2M scenarios. The results of this thesis target primarily 3GPP LTE and 5G NR Random Access Procedure (RAP), but can be transfered to a wide range of different random access protocols.

1.1 Research Challenges and Contributions

Here, we review research challenges addressed by contributions in Chapters 3–6. We categorize the challenges and the contributions into three groups. The first group (A) deals with random access protocols with a high steady-state load and addresses the question of resource management, Quality of Service (QoS) provisioning via prioritization, and resource re-use. The second group (B) analyzes and optimizes transient random access protocols behavior under novel traffic patterns characteristic for M2M, i.e., large spikes in the load correlated in time and space. The key difference between the groups (A) and (B) is the time scale of the operation: (A) concerns with resource management on an average-to-long time scale (in the order seconds and longer), while (B) is dealing with short-term load spikes, thus operating on a frame-by-frame basis. Finally, the third group (C) dives deeper into M2M applications by modeling them as controller-actuator feedback loops and addresses control-specific performance metrics and application-aware optimization techniques.

(A). Modeling, Performance Analysis, and Optimization of the Long-Term Steady-State Performance

Since typically data transmissions are infrequent and average packet size is small for M2M applications, radio resources are typically not maintained continuously. Instead, most M2M devices must complete the RAP to connect to the network before transmitting the data. Together with the massive number of M2M devices and their high density, this leads to high steady-state load in Random Access CHannel (RACH). Due to the waterfall-like decrease of RACH throughput [Vil+17b] once a certain load threshold is surpassed, RAP configuration in high load regimes must be carefully managed to avoid the overload. In such scenario, multiple research questions arise. First, RAP must be optimized in an adaptive way according to the load in the channel, i.e., throughput optimal amount of resources has to be determined. Second, it is foreseen that many M2M deployments will be rolled out in networks where M2M users must coexist with traditional H2H users. Hence, resource allocation must be flexible to enable their coexistence and prioritization of M2M and H2H applications, as well as of M2M users belonging to different QoS classes. While many existing technologies have mechanisms to prioritize data transmissions from different QoS classes, prioritization of control plane data is a largely missing aspect. Finally, in some scenarios, existing resources might be insufficient to support the RAP load. Thus, approaches for better utilization of the existing resources or to expand the amount of available resources must be studied.

The thesis addresses the challenges in steady-state random access performance in Chapter 3 with following **contributions**:

- Based on the steady-state performance analysis under a fixed back-off policy and infinite-source traffic model, we derive a Physical Random Access CHannel (PRACH) resource allocation policy maximizing the steady-state throughput for a given load. We study resource separation as a tool for coexistence of multiple QoS classes in a RACH system. We model the system with two generic QoS classes: non-prioritized delay-tolerant User Equipments (UEs) and prioritized UEs demanding low delay. We analytically investigate the impact of resource separation on the performance of such system in terms of delay, drop ratio, and throughput. Utilizing the analytical results, we devise Load-Adaptive Throughput-Maximizing Preambles Allocation (LATMAPA) policy, to provide adaptive prioritization of QoS classes during RAP. The policy is benchmarked with the state of the art approaches and is shown to outperform them in terms of the achieved prioritization.
- We study the performance of connection request aggregation as a technique for spatial reuse of time-frequency resources whenever the load is exceeding the capacity of RACH. We propose a Markov chain model to analyze the performance of RACH with aggregation. We demonstrate the benefits and trade-offs of connection request aggregation and its impact on the steady-state delay, drop ratio, and throughput in a finite user setting. We illustrate that there exists an optimal configuration where the sum of aggregation delay and random access delay is minimized.

(B). Modeling, Performance Analysis, and Optimization of the Short-Term Transient Performance

Apart from raising the steady-state load, M2M creates novel scenarios, atypical for conventional use of random access protocols. The primary example is a burst arrival scenario [3GP11]: An event causing a semi-simultaneous triggering of a large amount of devices to attempt a connection to Next Generation Node B (gNB). Such scenario is likely to cause a sudden overload in the RACH, and random access protocols are known to be prone to very large delays under synchronized arrival conditions. Steady-state analysis does not capture the effects of temporary overload scenarios, since they are averaged out in the long run. Therefore, transient behavior of the system must be modeled and analyzed using different methodologies.

LTE and 5G NR provide Access Class Barring (ACB) as a "toolbox" to react to the load changes in the RAP, which can be utilized to mitigate temporary overload by spreading the load in the time domain. However, purely time-domain load control has a negative impact on the delay and delay-constrained reliability. Instead, once again, we argue that a *resource-centric approach* to the optimization of transient behavior is needed. For that, the trade-off between resource consumption and the provided QoS (e.g., in terms of delay) must be systematically quantified. Additionally, existing overload control mechanisms do not consider the peculiarities of M2M traffic pattern, i.e., *correlations in time and space*. If neglected, these correlations can lead to excessing packet loss and delay in random access. Instead, we suggest to exploit the correlations to design novel overload control protocols. Finally, uM2M applications pose addition challenges for random access due to tight *reliability constraints*. Typically, access protocols are optimized with respect to their *expected* performance, however, for uM2M applications it is insufficient, and higher order statistics of the performance must be characterized. There is a need for a methodology to assess the reliability of the random access protocols, and a way of quantifying it for the existing solutions.

The thesis addresses the challenges in transient random access in Chapters 4 and 5 with following **contributions**:

- We analyze the inherent trade-off between the resource consumption and the throughput on the protocols for the burst arrival scenario. We demonstrate and compare two ways of designing random access protocols aware of the resource consumption: Taking the resource efficiency as an optimization metric or considering resource consumption and throughput as competing objective in a single multi-objective optimization problem. Taking latter approach, we derive the set of Pareto optimal solutions and devise an algorithm to adapt barring probability and the number of available preambles to the current backlog. We demonstrate that the algorithm achieves lower burst resolution time with respect to the state of the art under equal resource constraints.
- Exploiting the spatial correlation of the M2M burst arrivals, we propose to aid the existing RAP with a listen-before-talk scheme based on Binary Countdown Contention Resolution (BCCR). It improves the throughput by reducing the collisions

during the third step of the procedure. We analyze the performance of the resulting system, standalone and as a combination with the standardized ACB. Using the insights about the Pareto optimality provided in the first part of the chapter, we devise a policy to jointly determine barring probability and BCCR configuration. By the means of event-based simulations of the burst resolution scenario, we demonstrate that the proposed protocols significantly improve the throughput and burst resolution delay compared to the state of the art.

- We study the stochastic performance bounds of random access protocols with respect to latency and reliability. We propose a methodology for reliability assessment using stochastic network calculus. Furthermore, we demonstrate how the methodology can be used by analyzing reliability of RAP with static and dynamic ACB. Finally, we illustrate the effect of the estimation, i.e., when the number of devices is unknown, on the reliability-latency performance.

(C). Cross-Layer Design of Random Access Protocols and Networked Control Systems (NCSs)

The drastic difference between the requirements of M2M and H2H for communication networks comes from the *underlying control loops* of M2M applications. Jointly, control and network processes are often modeled as NCSs: Feedback control loops, consisting of a plant or a physical process, a sensor observing the plant's state, and a controller sharing a communication network in-between. A large variety of M2M scenarios can be modeled as NCS, from smart grids (e.g., controllable loads, energy markets) or vehicular communications (e.g., platooning) to industrial automation (e.g., control of a manufacturing process). Conventionally, the coupling of control and network performance is implemented via layered abstraction. The requirements and traffic profile are abstracting the control application, while protocol performance guarantees are abstracting the network. However, in many cases such abstractions are inefficient, and there might not be a clear one-to-one relationship between the network and the application performance. For example, latency requirements of a control systems might be time-varying depending on the system's current state and dynamics. It is thus beneficial to have a tighter cross-layer interaction between the control application and the network.

MAC layer protocols, managing the access to shared wireless resources, have a direct impact on NCS performance as a primary coupling point between individual control systems. The coupling is especially prominent for random access protocols, where consecutive collisions can implicitly correlate data transmissions from multiple control systems. To understand how to design MAC protocols for M2M, detailed control system models and realistic MAC protocols must used in a joint study.

The thesis addresses the challenges in cross-layer design for networked control systems in Chapter 6 with following **contributions**:

- A detailed model of a Networked Control System is considered, where multiple Linear Time-Invariant (LTI) sub-systems are coupled while sharing the wireless medium with multi-channel slotted ALOHA protocol. We analyze the behavior of an NCS under local event-triggered scheduling policy common for all sub-systems. We further introduce an adaptive scheduler to improve the event-trigger design. In the new scheduler, the network and control systems are coupled via the knowledge of the network state: Each local scheduler adapts its threshold based on the available network resources. Simulatively, we demonstrate that an adaptive choice of the transmission threshold is beneficial compared to a non-adaptive static design.
- We introduce an approach to dynamically prioritize random access among multiple sub-systems, employing the binary countdown technique. In the proposed approach, priority of every system is determined dynamically and locally based on the plant state. Numerical analysis illustrates a considerable performance improvement compared to the state of the art decentralized and centralized techniques. It is demonstrated that the proposed scheme can be deployed more efficiently by significantly lowering the collision rate in case of large number of systems utilizing the communication network.

1.2 How To Read This Thesis

The structure of the thesis is illustrated in Fig. 1.1. The remainder consists of two parts: an overview of the background in Chapter 2 and main contribution Chapters 3–6. The main contribution chapters are mostly independent from each other, therefore the thesis is written with the intention to facilitate "random access" to the individual chapters. A common system model is introduced in 2, and every chapter from 3–6 has a brief recap of it. Additionally, every chapter 3–6 provides a brief overview of closely related works. The notations are kept consistent within every chapter and, whenever convenient, between the chapters. The reader however should not count on full notation consistency between the chapters.

Chapter 2 introduces the reader to M2M application, their communication requirements, and MAC layer challenges for communication technologies. It outlines the role of random access protocols and introduces necessary background on RAP.

The main contributions of the thesis are presented in Chapters 3–6. In Chapter 3, the problem of average-to-long term resource management for steady-state performance of M2M random access is studied. Chapters 4–5 address the problem of short-term resource management in the transient state random access, with the scenario of a burst arrival of connection requests. In Chapter 4, a resource-centric minimization of the average burst resolution time is presented. In Chapter 5, higher order statistics and the reliability aspect the random access protocols are studied. In Chapter 6, a detailed model of interaction between an Networked Control Systems (NCSs) and underlying communication

protocols is considered. We develop dynamic cross-layer optimization approaches NCS with random access protocols.

Chapter 7 concludes the thesis with a summary of the results and directions for further work.

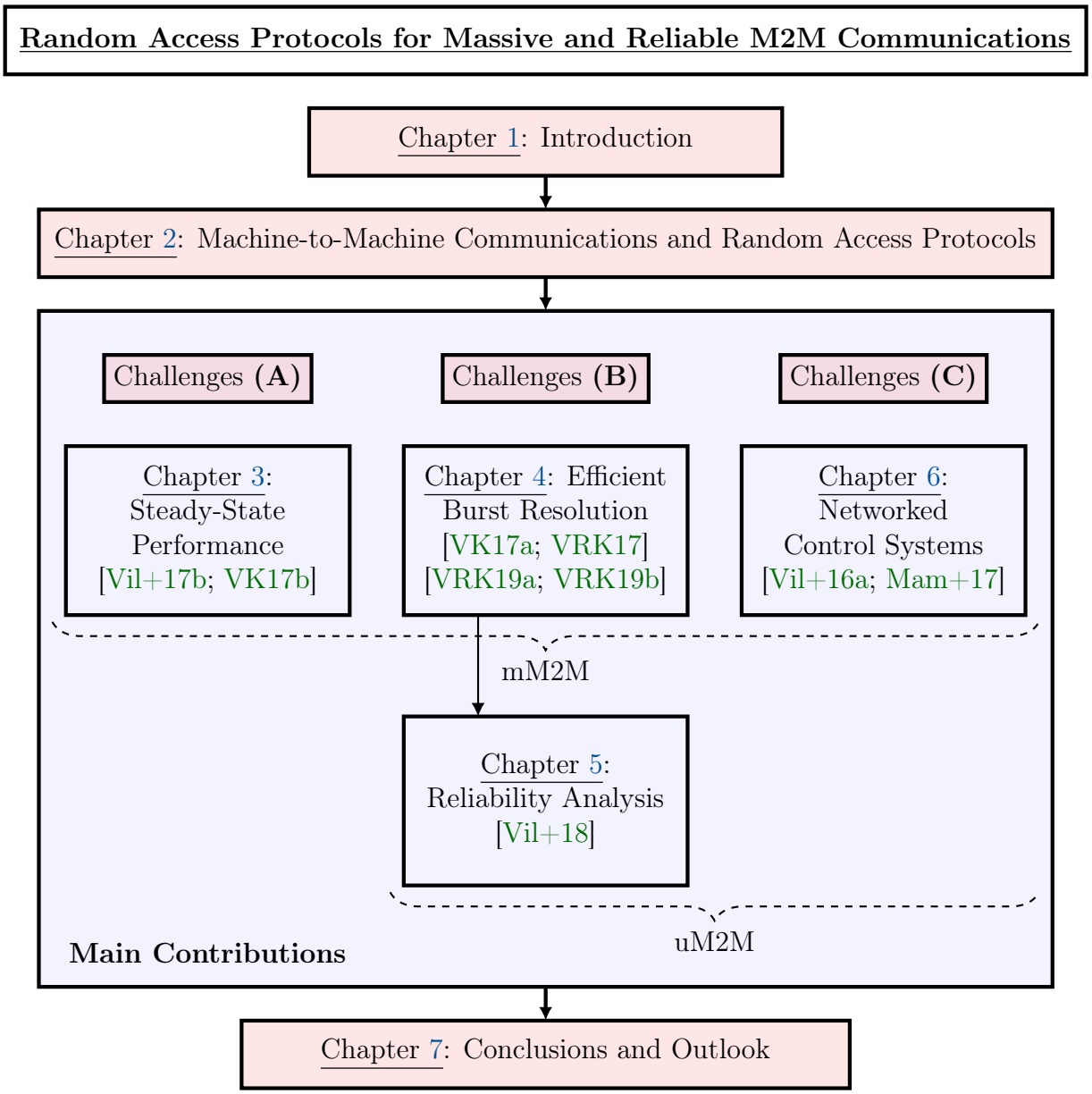

Figure 1.1: The outline of the thesis, with the chapters mapped to respective research challenges and publications.

Chapter 2

Machine-to-Machine Communications and Random Access Protocols

The aim of this chapter is to familiarize the reader with Machine-to-Machine (M2M) applications and random access protocols. We explain the need for the random access protocols and their role in M2M communications. First, in Sec. 2.1, we introduce M2M as the main type of application motivating the work in later chapters. We outline the exemplary applications and their requirements. In Sec. 2.2, we give an overview of the communication aspects of M2M including challenges and potential communication technologies. Then, in Sec. 2.3, we introduce the fundamentals of LTE and NR Random Access Procedure (RAP) and a basic approach of modeling RAP as a multi-channel slotted ALOHA. Finally, we review recent advanced on random access for M2M in Sec. 2.4.

2.1 Applications and Scenarios

There exist a large variety of M2M applications, roughly classified into two groups. Massive Machine-to-Machine (mM2M) applications are typically described as non-critical, tolerant to delay and packet loss. These applications involve a massive number of devices densely populating the areas. A typical example is smart metering, also referred to as advanced metering infrastructure: Meters deployed in the private houses or business facilities for automatic reporting of the utilities consumption. According to [Gun+13], it is sufficient for the smart meter readings to be delivered within 2 seconds, and the connection must be available 99 % of the time. Considering a densely populated city, where every household is equipped with a smart meter, the resulting number of deployed devices per large cell can reach hundreds or thousands. Alongside with other deployed applications, such as smart city or vending machines, meters create a significant strain on the network.

The other group, ultra reliable Machine-to-Machine (uM2M), is characterized by the stringent requirements for latency and availability. Third Generation Partnership Project (3GPP) provides a general requirement for a Ultra-Reliable Low Latency Communication (URLLC) application of 99.999 % reliability and 1 ms latency for a 32 byte packet [3GP18a]. A prominent motivating example here is industrial automation. This use case combines a variety of applications requiring low latency and high availability

communication in challenging environments. For example, control panels with safety functions for interaction with factory machines require latencies down to 30 ms with 99.9999 % reliability, while additionally imposing requirements on the jitter not to exceed 50 % of the latency [3GP18b]. Supporting such requirements is challenging and could consume a lot of wireless resources to provision necessary level of redundancy and diversity in frequency or time. Therefore, there cannot be many uM2M devices supported in the network on the same time.

While the performance requirements of uM2M and mM2M differ drastically, they have multiple features in common. First, applications of both types usually transmit *small amounts of data* at once. E.g., mobile control panels from the previous example use $40 - 250$ byte packets for communication, and smart meters ≈ 100 bytes [3GP18b; KPR14]. This motivates the notorious problem of short packet communication, an active research subject in multiple communication domains, including coding and information theory [DKP16]. For the Medium Access Control (MAC) layer, transmission of small amounts of data means that the signaling overhead can drive the efficiency down. Mitigation of the overhead is important for both mM2M and uM2M, but for different reasons. In mM2M, per-packet overhead creates network scalability issues and large waste of resources. In uM2M, overhead means longer latency, which is to be avoided.

Networked Control Systems

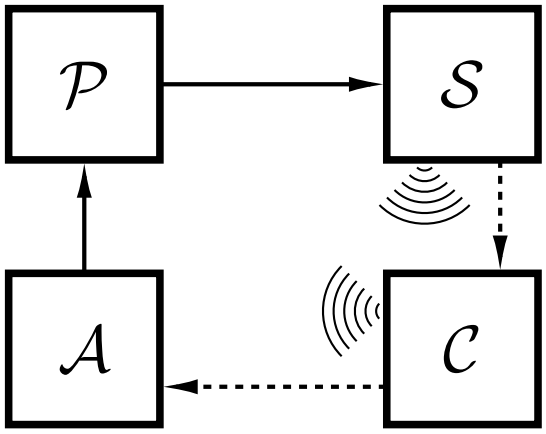

Figure 2.1: An exemplary single-loop NCS, consisting of a plant $\mathcal{P}$, sensor $\mathcal{S}$, controller $\mathcal{C}$, and actuator $\mathcal{A}$. Sensor $\rightarrow$ controller and controller $\rightarrow$ actuator pairs might be collocated or distant. In the latter case, the respective information is transferred via wired or wireless communication links.

A special class of M2M applications are Networked Control Systems (NCSs). This class is representative for a wide range of M2M applications, since many of them have underlying *control loops*, for instance, in the mircogrid control or production line in an industrial facility [Par+17; WY01].

The application layer of an NCS is modeled as a feedback control loop, as illustrated

in Fig. 2.1. In general, it consists of a *plant* $\mathcal{P}$ representing a physical process, e.g., changes in the temperature of the room for Heating Ventilation and Air Conditioning (HVAC) applications. The output of the plant is observed by a *sensor* $\mathcal{S}$. Sensor might be collocated (e.g., for HVAC) or remote (e.g., power grid monitoring). The observations of the sensor are delivered to the *controller* unit $\mathcal{C}$. If the observations are noisy, or if the $\mathcal{S} \rightarrow \mathcal{C}$ communication link is lossy, estimation is deployed to recover the true state of the plant (or its closest estimate) and calculate a control input to the system [Sin+04].

Depending on the observation and estimation results, the controller might need to adjust its actuation input to to change the plant's state to the desired, by delivering the control input towards the *actuator* $\mathcal{A}$. In many scenarios, such as power grid control, plant and controller might be collocated. In general, however, information both from the sensor to controller and from the controller to the actuator is delivered via wired or wireless communication links. The communication links might also be shared between multiple (possibly non-independent [MDH15]) feedback control loops, with the resulting system referred to as multi-loop NCS [Vil+16a]. Individual control loops are commonly modeled in theoretical works as an Linear Time-Invariant (LTI) system [WYB02].

In this thesis, Chapters 3–5 abstract the application by its traffic pattern and requirements. In contrast to it, Chapter 6 is dedicated to a study of the interplay of a NCS and the underlying communication protocols.

2.2 Machine-to-Machine Communications

In this section, we overview high-level aspects of M2M communication. We introduce challenges of M2M communication (2.2.1), relevant radio resource management concepts (2.2.2), and communication technologies and standards with a potential to serve as a basic for M2M communication (2.2.3).

2.2.1 Overview

Although individual applications have been already deployed in conventional commercial cellular networks for several years [Sha+12], current networks are not designed to fully support mM2M and uM2M [Oss+14; LCL11]. 4G networks have been primarily designed to support low number of data rate hungry users with applications like video streaming or web browsing. As a consequence, they lack a number of important features [Nok15]. The challenging requirements for M2M communications are summarized in the following [3GP17b; 3GP11].

Large amount of devices. The scalability of the communication network is one of the major bottlenecks on the way to mM2M. Since M2M devices are typically communicating infrequently and with low data rate, current technologies, such as WLAN or LTE, might have sufficient bandwidth to accommodate large number of M2M devices. However, the

scalability of communication protocols and control plane procedures has been shown to be an issue [LAAZ14]. The problem arises since the conventional networks have been optimized to support frequent, high data rate transmissions from a small number of devices, where the overhead of establishing the connection and maintaining the radio resources is negligible compared to the data. The situation is reversed for M2M applications, where the connection establishment and its maintenance consume a significant and often the dominant part of the resources.

Low cost and low complexity hardware. With the number of users and their density going up in mM2M, cheaper hardware and lower modem costs are required to further increase the penetration of mobile connectivity. Typical M2M users do not require high data rates supported by the conventional networks, hence, there is a potential to trade-off device capabilities with the complexity and respective costs.

Enhanced coverage. Possible basement or rural installments of M2M devices, e.g., in the case of smart metering, make them often unreachable for cellular communication [HIW14]. Technologies with extended coverage and with high capabilities of building penetration are necessary. This challenge is both technology- and standard-related, i.e., both increased link budget or different carrier frequency can help to overcome it.

Lower power consumptions. M2M devices might be deployed in remote locations where access to power supply is limited and maintenance is costly. This leads to an increase in the number of battery powered devices, hence, hardware and communication protocols need to be optimized to keep the devices in operation as long as possible. Requirements of up to 10 years of battery life are envisioned for some applications [Nok15]. A trade-off of device capabilities and its power consumption can also be utilized to address this challenge.

Wide range of performance requirements. The diversity in M2M applications creates a diversity in the requirements for underlying communications, in particular, with respect to latency and reliability. The LTE-like approach of classifying the applications into several Quality of Service Class Indicator (QCI) classes cannot represent all possible M2M requirements. Additionally, possible coexistence with uM2M applications calls for flexible protocols and resource management approaches, so that incorporating diverse requirements is enabled.

Uplink-driven communication. In 4G networks, the communication is imbalanced towards the downlink, i.e., from Next Generation Node Bs (gNBs) towards the User Equipments (UEs), since most data rate hungry applications like video streaming or web browsing involve downloading large amounts of data. For M2M, this is likely to change, since the data is mostly collected from the field devices through the uplink, while the control signals in the downlink are expected to be less frequent.

To address the challenges of M2M application, multiple architectures have been proposed for M2M communications by 3GPP [3GP12], European Telecommunications Standards Institute (ETSI) [ETS11], and other organizations. The proposals are determining different communication models for M2M (referred to as Machine Type Communications (MTC) in 3GPP context). According to [DOK15], ETSI documents are focused on the application aspects of M2M and are largely abstracting the underlying network, while 3GPP is focused on communication aspects and is defining the network functions and their respective hosting entities. On a high level, both architectures can be summarized as illustrated in Fig. 2.2. M2M devices are distributed in an area and are connected together via an *M2M area network*. The M2M area network is providing physical and MAC layer connectivity and is typically controller by a base station (gNB in 5G terminology). The gNB is also a gateway into the operator network and public network. In an indirect model, M2M application is communicating with an M2M UE via the M2M server, but a direct communication without server as a proxy is also foreseen. Depending on the business model, the server might be part of the operator network or reside outside of it and belong to the third party.

2.2.2 Radio Resources Management

In the thesis, we only concern with the M2M area network and especially with its MAC layer, determining how the radio resources in the network are managed. To this end, we give a brief recap of the relevant MAC layer and radio resource management concepts.

A radio resource is defined as any distinct part of a frequency (bandwidth), time, or code domain, available for a data transmission. MAC protocols are regulating access to the radio resources, i.e., the logic of how the users *obtain*, *maintain*, and *utilize* the radio resources. The MAC protocols can be organized into two groups: protocols based on access reservation and contention-based protocols, also denoted as *random access*. Ac-

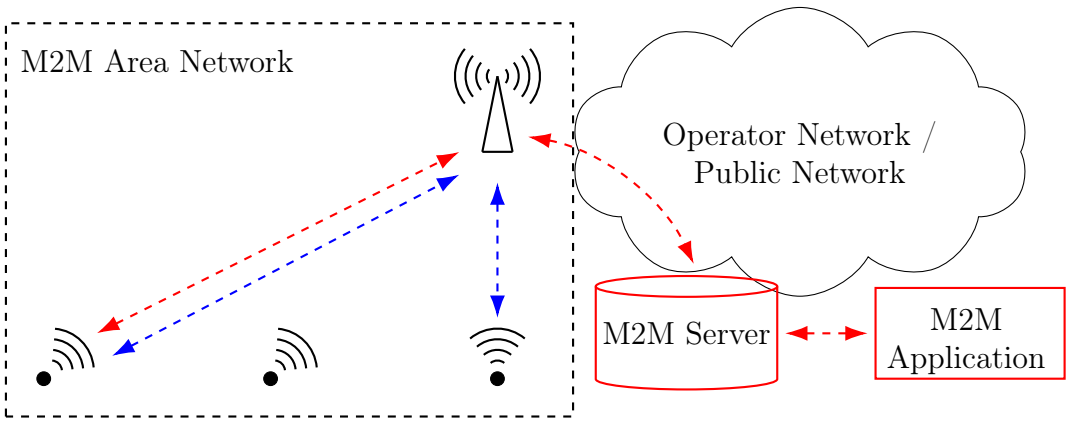

Figure 2.2: M2M communication architecture, consisting of: devices connected with M2M Area Network, Operator Network, M2M Server (may be located outside or inside the operator network), M2M Application. Devices might communicate with each other or with the server and application.

cess reservation protocols imply that the same resource must not be accessed by different users, e.g., as in Time Division Multiple Access (TDMA), Frequency Division Multiple Access (FDMA), or Orthogonal Frequency Division Multiple Access (OFDMA) based systems, thus guaranteeing that the signal from other users does not affect the user to whom the resource is assigned. A centralized coordinator is typically determining the assignment of the resources to the users (i.e., the *schedule*), although its functionality can also be implemented in a distributed fashion. In contention-based protocols, e.g., ALOHA and its derivations, multiple users might access the same resource blocks, hence, interference between the users is accepted. The interference is typically causing collisions which are resolved by the means of contention resolution procedure. Since we also consider code domain as a resource dimension, the group of Non-Orthogonal Multiple Access (NOMA) [Sai+13] protocols might be implemented both with access reservation or without access reservation (grant-free NOMA). Hybrid MAC protocols also exist, where the periods of contention-based access are followed by slots with scheduled access [Geh+14].

In the context of cellular networks, MAC procedures of obtaining and maintaining the resources are denoted as the *control plane*, and the transmission of application data itself as the *user or data plane*. The control plane operation in random access protocols is often implemented in a distributed fashion, where there is either no explicit coordination present (e.g., Distributed Coordination Function (DCF) in Wi-Fi), or it is simplified and performed by a single centralized entity, e.g., a gNB. In access reservation MAC protocols, the control plane is typically implemented in a centralized fashion and the access is fully coordinated by the central entity, e.g., a gNB.

2.2.3 M2M Technologies and Random Access

To that end, no single technology is capable of addressing all the M2M challenges. Industry and research are working towards new standards for M2M communications. In this subsection, we review standards and technologies seen as potential candidates for M2M communications, with the focus on MAC layer and on the relevance of random access protocols for these technologies. We restrict ourselves to wide area networks and intentionally leave out personal and local area network standards [Gaz17]. Moreover, only the technologies providing physical and MAC layer connectivity are reviewed, i.e., M2M area networks [DOK15]. The summary of this brief review is given in Table 2.1.

The technologies can be classified into two large groups by the spectrum they are operating in: unlicensed and licensed. Technologies operating in *unlicensed spectrum* typically offer cheaper and easier infrastructure set-up, but suffer from such drawbacks as interference and lack of coordination between co-deployed networks, both inter- and intra-technology [Cen+16]. On the other hand, setting up the infrastructure in the licensed spectrum incurs significant costs and the spectrum is typically auctioned. In return, the licensee has full control over the resources, which potentially enables higher spectrum utilization.

Spectrum	Technology	Random Access
Licensed	3GPP LTE, NR	Obtaining resources
	LTE MTC [Nok15]	Obtaining resources
	NB-IoT [Roh16]	Obtaining resources
	Grant-free NOMA [Dai+15]	Maintaining resources & Data transmission
Unlicensed	LTE Unlicensed [Muk+16; Mul17]	Obtaining resource & Data transmission
	LoRa [Lora]	Data transmission
	SigFox [Sig]	Data transmission
	IEEE 802.15.4g [IEE12]	Data transmission or Obtaining resources

Table 2.1: Summary of cellular M2M technologies. The right column specifies which MAC functionality is using random access protocols: *To obtain* the resources (e.g., RAP); *To maintain* the resources (e.g., channel estimation via pilots or buffer status reports); *To transmit* the data.

Unlicensed Spectrum

Prominent examples of unlicensed M2M are Low-Power Wide-Area Network (LPWAN) technologies: LoRa [Lora; Aug+16] and SigFox [Sig], introduced in 2008 and 2009, respectively. LoRa relies on a proprietary physical layer solution based on the chirp spread spectrum and on the open Long Range Wide-Area Network (LoRaWAN) specification of the MAC and network layers [Cen+16]. SigFox is using ultra narrowband modulation, and proprietary non-disclosed network layer protocols. Both LoRa and SigFox operate in sub-GHz bands, thus allowing wide coverage of tens of kilometers or even up to hunderds of kilometers in certain circumstances [Lorb].

On MAC layer, both SigFox and LoRaWAN use unslotted random access for user plane transmissions, thus *avoiding the overhead of obtaining and maintaining radio resources*. SigFox is creating additional time and frequency diversity using packet replicas and channel hopping to provide more reliable delivery. In LoRaWAN, uplink transmissions are unregulated and can be performed asynchronously, while downlink is constrained by the reception windows of the devices. Both SigFox and LoRaWAN limit data rates and restrict traffic patterns by imposing mandatory duty cycling for the end devices. As long as a given frequency band is not densely populated, the standards could offer satisfactory service for delay-tolerant mM2M applications. However, if many networks with large numbers of users are deployed in the same area, the performance of LoRa and SigFox significantly degrades [MPH16]. The problem is amplified when co-deployed networks belong to different operators, since it complicates the interference management. In addition, inter-technology interference can negatively affect LoRa and SigFox

performances [DP+17].

Another example of a LPWAN technology is IEEE 802.15.4g: A long-range extension of IEEE 802.15.4 standard for wireless personal area networks, targeting primarily smart utility networks [IEE12]. The standard adapts the IEEE 802.15.4 physical layer to support operation in lower frequency bands (868 MHz in EU and 915 MHz in USA) and thus extend the coverage. IEEE 802.15.4g can operate with any MAC protocol supported by 802.15.4, e.g., Carrier Sense Multipe Access (CSMA) or Time Slotted Channel Hopping (TSCH) [Vil+16b], the latter being essentially TDMA aided with channel hopping for better resistance to cross-technology interference [G¨+16].

Licensed Spectrum

Using mobile network standards in the licensed spectrum, in particular LTE and NR, provides multiple advantages: (1) existing infrastructure and good coverage; (2) mature standardization body; (3) high bandwidth; (4) mobility support. While LTE has high spectral efficiency and supports high data rates, it does not address any of the mM2M challenges we described in 2.2.1. The capabilities and complexity of the LTE protocol stack result in high cost modems with high power consumption. To address these LTE limitations, several derivate technologies have been developed and standardized from release 12 onwards: MTC [Nok15], NB-IoT [Roh16], and Extended Coverage-GSM-IoT (EC-GSM-IoT). The first two standards rely on the evolution of the existing LTE stack, and can operate in different as well as in the same band of an existing deployed LTE network. MTC has been introduced in the LTE release 12, and later developed into enhanced MTC (eMTC) in release 13. It presents new UE categories: LTE Cat. 0 and Cat. M1, trading off the achievable data rates and system bandwidth for lower complexity and costs. Both categories support up to 1 Mbps peak downlink or uplink data rate, with Cat. 0 supporting up to 20 MHz bandwidth, and Cat. M1 only 1.4 MHz. To extend the network coverage, eMTC suggests repetition-based modifications to the LTE signals, and revises several procedures and channels.

For the large part, 5G New Radio (NR) standard has evolved out of LTE in a classical direction of pushing the data rates up, especially by utilizing higher frequency range (FR2). However, it also has a number of prominent features which can be used to better support M2M. First of all, lean design minimizing always-on transmissions, introduction of bandwidth parts, and introduction of network slicing help in reducing the energy usage and potentially enable co-existence of diverse application. Second, low-latency support features, e.g., fine granular resource allocation down to 1 Orthogonal Frequency Division Multiplexing (OFDM) symbol or reduction in processing and waiting times for grant allocation [DPS18], make NR a unique standard capable to support uM2M requirements. To that end, 5G NR is a promising candidate for future M2M roll-outs.

From the MAC perspective, 3GPP LTE, NR, and all their derivatives deploy access reservation-based MAC protocols for the data transmission, with UEs accessing the resources in an OFDMA fashion. The resources are assigned to UEs by a centralized scheduler located at the gNB. The difference to many other MAC protocols is that the

scheduler is typically assigning the resources *dynamically* on-demand. This means that LTE and its derivatives have significant control plane *overhead to obtain and maintain transmission resources*. To obtain the resources after long inactivity periods or during a handover, UEs have to undergo Random Access Procedure (RAP). The MAC layer is thus tailored to bursty heavy-tailed traffic, and not well-suited for M2M. Currently, 3GPP is studying semi-persistent resource allocation (under umbrella term "grant-free access") as a lightweight M2M-friendly scheduling strategy. *Both grant-free access and RAP are contention-based, random access protocols.*

Remark 1. *For completeness, grant-free NOMA [Dai+15] and unlicensed version of LTE (Licensed Assisted Access [Muk+16] and MulteFire [Mul17]) deserve a mention as potential M2M technologies. Grant-free NOMA relies on power- or code-domain multiplexing and subsequent Successive Interference Cancellation (SIC). In contrast to its grant-based counterpart, it does not require explicit scheduling of the transmissions. From a MAC layer perspective, grant-free NOMA can be viewed as a part of the sub-group of random access protocols with SIC [YG07; Liv11]. In addition to that, NOMA requires accurate channel estimation via pilots. The technology is a trending research topic and can be a potential candidate for M2M communications, however, to the best knowledge of the author, no standardization activity exists yet. LTE variations for unlicensed spectrum are complicated in Europe by ETSI requirements for fair coexistence with WLAN [Muk+16], forcing LTE to deploy carrier sensing and contention resolution techniques similar to WLAN [Muk+16; SVK17]. This means that, LTE in unlicensed spectrum uses random access both to obtain the resources for transmission, and for the actual transmission. The interplay between these two coupled processes is studied in [SVK17].*

Summary

The takeaway message from this section is that medium access of all modern M2M technologies relies on random access. For technologies operating in unlicensed spectrum, it is common to use random access for data plane transmissions. For licensed spectrum technologies, random access is typically used for control plane transmissions.

Due to high penetration of the standard and mature standardization body, we choose to focus on the application of random access protocols for LTE and 5G NR in this thesis. The results are in large part generalizable to other licensed spectrum technologies, such as NB-IoT and LTE-M, and some insights and methodologies could well be reused in unlicensed spectrum.

2.3 Random Access Protocols

This section links the control plane procedures of LTE and NR to the theoretical problem of random access protocols. We first introduce the reader to RAP (2.3.1). Then, we present the most common queuing-theoretic MAC layer model of RAP: multi-channel

slotted ALOHA (2.3.2). Finally, we introduce the performance metrics arising from novel M2M-specific challenges for random access protocols (2.3.3).

2.3.1 Random Access Procedure

The description of RAP in this subsection is based on 3GPP MAC, radio resource control (RRC), and physical layer LTE and NR specifications [3GP18b; 3GP16; 3GP15b]. In addition to this subsection, we provide a brief recap in every chapter 3 to 5. From now on, we use NR terminology to denote users as UEs and the base station as gNB.

A UE needs to go through the procedure in order to obtain initial synchronization with the gNB and a grant for subsequent data transmission. RAP is necessary in case of a transition from RRC-IDLE or RRC-INACTIVE state to RRC-CONNECTED or in other cases of lost synchronization such as handover. The RAP can be initiated by the UE or by the gNB as a part of the paging procedure [HHN13]. An outcome of a successful RAP is the acquisition of resources for an uplink transmission on the Physical Uplink Shared CHannel (PUSCH)[1]. There are two modes of RAP: contention-free or contention-based. In the contention-free mode, the gNB can uniquely identify a UE by a received preamble sequence, for the preamble to be used has been communicated to the UE in advance. Such a scenario is possible in certain cases, e.g., during a handover between two gNBs. In the following, we consider only the contention-based RAP.

A message exchange chart[2] for RAP with Access Class Barring (ACB) is depicted in Fig. 2.3. The procedure starts with a UE listening for the System Information Blocks (SIB2 for LTE) messages advertised by the gNB on the broadcast channel. gNB broadcast provides a valuable tool for enhancements of the RAP as it can potentially carry additional information for all users, e.g., amount of available Physical Random Access CHannel (PRACH) resources, access probability, back-off parameters. The broadcast message contains the PRACH Configuration index and the frequency offset. These two parameters inform UEs about the sub-frames and Resource Blocks (RBs) that are reserved for PRACH in the next frame. Depending on the Configuration index, one or more sub-frames can be reserved for PRACH. We refer to a *PRACH slot* as the time between the beginning time instants of two consecutive PRACH sub-frames. The length of the PRACH slot depends on the PRACH Configuration index, and can vary from 1 ms up to 20 ms. PRACH slot also corresponds to a slot in the generic multi-channel slotted ALOHA model introduced later in 2.3.2.

Message 1 (MSG1): PRACH Preamble Transmission. The first step of RAP uses a dedicated PRACH. An example location of the PRACH on the resource grid is depicted in Fig. 2.4. A UE selects a preamble from the available set and sends its to the gNB as Message 1 (**MSG1**)[3]. In a typical configuration, PRACH has 64 available pream-

[1] In some cases, data transmission might already occur during RAP. Such operation mode is referred to as early data transmission [Hog+18].

[2] Only the basic version of RAP is depicted here. In NR, the procedure might be additionally complicated by beam establishment if beamforming is used.

[3] It is important to remark that prior to this point barring or back-off mechanisms may be applied.

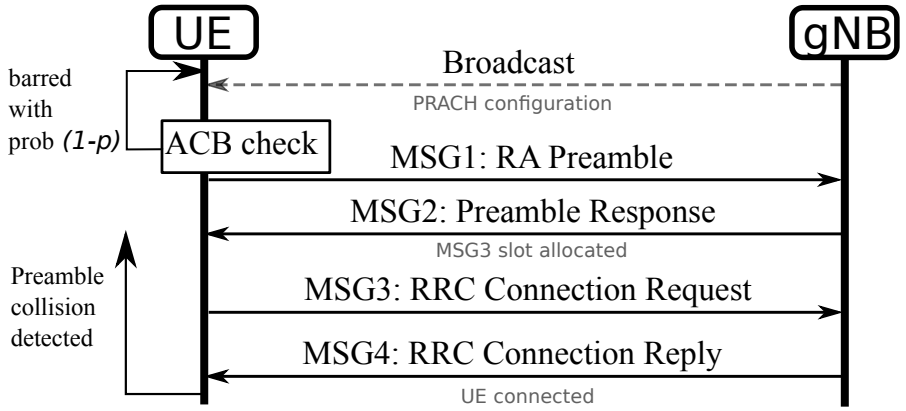

Figure 2.3: Illustration of the four-step contention-based Random Access Procedure.

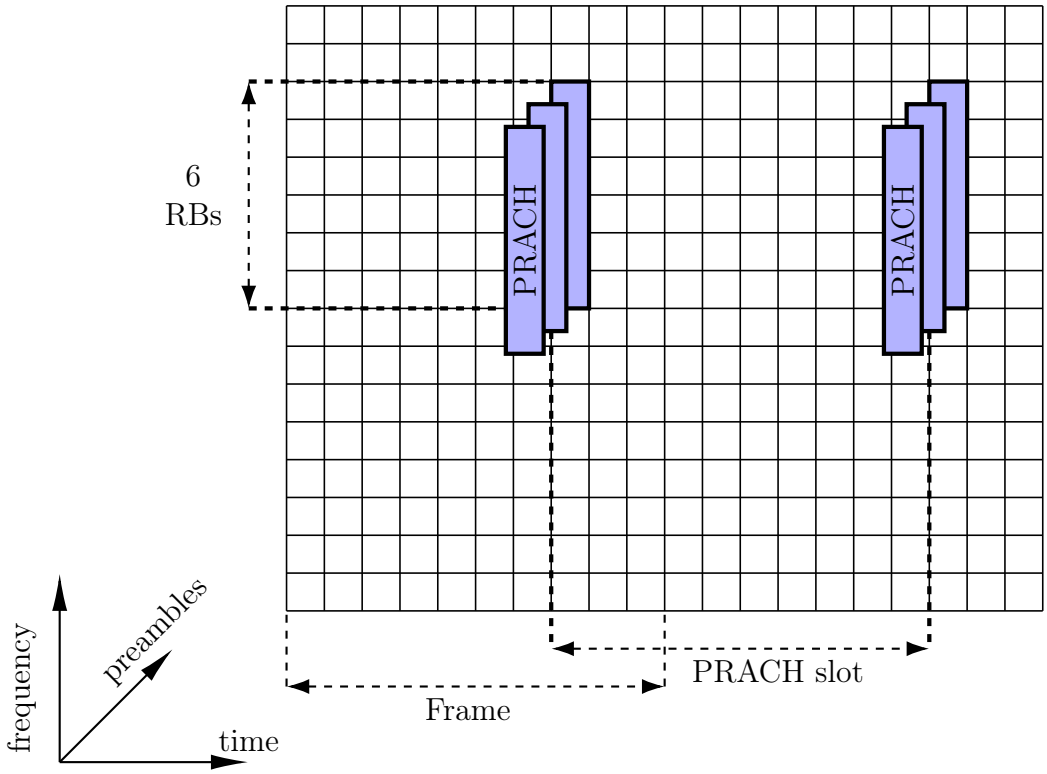

Figure 2.4: Exemplary PRACH allocation on the resource grid: 3 MHz bandwidth corresponding to 16 RBs for all channels, PRACH configuration index 5, frequency offset 7, preamble length in frequency domain 6 RBs [3GP15b].

bles, whereby 10 preambles are reserved for contention-free access and 54 preambles are available for contention-based access.

A preamble is a Zadoff-Chu sequence. UE can choose a preamble sequence from a set, obtained by the cycling shift from the root sequence advertised by the gNB in the system information. The preambles enable detection of multiple users in the same time-

Here, we omit them for clarity and only describe the relevant details in the respective chapters.

frequency resource, if the preambles users choose are different. This is achieved since the Zadoff-Chu sequences obtained from the same root sequence have a zero correlation property [3GP15b]. Note that MSG1 does not contain any information about UEs identities. Thus a possible collision can not be identified upon the reception of MSG1. The gNB can only detect whether a particular preamble has been *activated* (i.e., selected by at least one UE) or not, but it cannot detect how many UEs have selected the preamble [MSP14]. This property is at the core of modeling RAP as a variant of multi-channel slotted ALOHA protocol [Tya+15], where every preamble sequence is considered as an orthogonal channel.

Message 2 (MSG2): Random Access Response (RAR) Transmission. After the reception of PRACH preambles, gNB attempts to decode them and indicates which preambles have been activated and which not. Then, the gNB sends a random access response in a downlink channel, in which it indicates PUSCH resources for the next transmission for every activated preamble. In current LTE implementation, the limitations of the RAR message do not allow to send the responses for more 15 preambles [WBC15]. However, throughout most of this thesis, we ignore this limitation unless otherwise stated, since it does not qualitatively change the results and it is possible to accommodate this limitation into our models whenever necessary. RAR message also contains timing advance correction for uplink and a temporary identity for a UE.

Message 3 (MSG3): RRC Connection Request Transmission. Every UE which successfully received RAR, proceeds with the third step of RAP. It synchronizes its uplink timing with the network and sends an RRC Connection Request in the allocated PUSCH resources. The request contains the UE's identity information, and, if successfully decoded, it is used to authenticate the UE and grant the connection. If more than one UE have selected the same preamble for MSG1, a collision will occur and none of the collided UEs will be granted access. If, however, a UE had selected a unique preamble for MSG1, no collision occurs and the UE receives the necessary connection setup response as **MSG4**.

Remark 2. *After RRC Connection Request reception and before Connection Reply, UE identity may be communicated to Mobility Management Entity (MME) (or Access and Mobility Management Function (AMF) in NR), and the core network signaling operations corresponding to the initial attach procedure are performed. In this thesis, we focus on the Radio Access Network (RAN) aspects only, as we keep the work generic and independent on the core architecture and implementation. Hence, we ignore possible effects of the core signaling and bearer establishment, assuming that all authentication and verification procedures can be performed with negligible delay.*

Message 4 (MSG4): RRC Connection Reply. For every successfully decoded RRC Connection Request, the gNB is sending an RRC Connection Reply confirming that the connection has been successfully established. The connected UE is granted uplink resources to transmit its data. If a UE does not receive RRC Connection Reply within a specified time window, UE decides that RAP has failed and that a preamble transmission should be repeated, possibly with a random back-off. If MSG4 is successfully received,

the RAP is considered successfully completed, and uplink data transmission explicitly scheduled by the gNB can be performed.

2.3.2 Multi-Channel Slotted ALOHA

In this thesis, we exploit a well-know *multi-channel slotted ALOHA model* (MS-ALOHA) of RAP, where PRACH preambles represent channels. The basic model includes a set of users attempting to transmit data towards a common receiver. In a typical scenario, we assume that UEs are attempting to establish a connection to gNB. The set of UEs can be finite, where each UE has a certain traffic or activation pattern [3GP11], or infinite[4]. The latter is typically described by the total traffic intensity [BGH87; Tya+15]. A more detailed discussion on the validity and applicability of these two modeling approaches can be found in Chapter 3.

A *slotted time* system is assumed, where all UEs are synchronized sufficiently to be aware of the beginning and the end of every slot. Slot synchronization is achieved via periodic synchronization signals from the gNB. A slot denotes the time domain resource, and there might be multiple Random Access Opportunitys (RAOs) per slot for the systems with multiple resource dimensions (e.g., frequency or code). Multiple parallel RAOs in the same slot are referred to as *channels*.

Definition 1 (Slot)**.** *A random access slot is defined as the time between two consecutive time-domain opportunities to perform a transmission. It is assumed that a slot is of sufficient duration for data (or a connection request) of any UEs to be fully transmitted.*

Definition 2 (Random Access Opportunity)**.** *A RAO is defined as a time × frequency × code resource which is sufficient for one UE to fully transmit its payload.*

In many protocols, control information, including collision feedback or contention parameters, cannot be assumed to be communicated to the contending UEs in every slot due to time constraints. Therefore, it is common to generalize the collection of slots as a frame. To avoid confusion with the frames in LTE, we use the definition of a *contention round*. The relationship between the concepts of RAOs, slots, and contention rounds is illustrated in Fig. 2.5.

Definition 3 (Contention round)**.** *The time is divided into contention rounds. A contention round defines the smallest period within which contention parameters (access probability, number of resources, etc.) can be changed. The smallest possible duration of a contention round is one slot.*

Assume that n UEs attempt a transmission in a given RAO. A simple Signal to Interference to Noise Ratio (SINR) threshold model can be used to determine whether the

[4] In the context of traffic modeling, these two approaches are often referred to as source model and aggregated model [Gri+17].

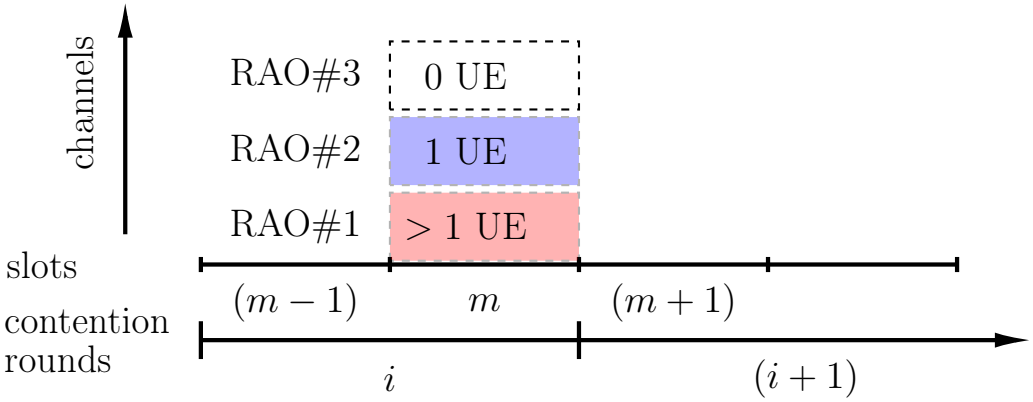

Figure 2.5: Exemplary timeline of MS-ALOHA protocol, with two slots per contention round and three RAOs per slot. According to the basic 0/1 collision channel model, RAO#1 is resulting in a collision (> 1 UE), RAO#2 is success (1 UE), RAO#3 remains idle (0 UE).

data of a particular UE j is successfully decoded:

$$\gamma_j = \frac{|h_j|^2 P_{\text{tx},j}}{\sum_{i\in\{1\ldots n\}\setminus j} |h_i|^2 P_{\text{tx},i} + \eta} \geq \gamma_{\min}, \tag{2.1}$$

where η is the noise at the receiver, $P_{\text{tx},i}, P_{\text{tx},j}$ are transmission powers of the ith and jth UE, respectively, h_i, h_j are the channel coefficients from UEs i, j to the gNB, respectively. In other words, SINR for UE j must be beyond a certain threshold $\gamma_{\min}$ for the data to be successfully decoded.

If channel coefficients of all UEs are considered, analysis of MAC protocols becomes increasingly complex and often intractable. A common approach to decrease the complexity is to reduce the threshold-based model to a *collision channel model* [BGH87]. The standard 0/1 collision channel assumes that following two conditions are satisfied:

$$\frac{|h_j|^2 P_{\text{tx},j}}{\eta} \geq \gamma_{\min}, \quad \forall j. \tag{2.2a}$$

$$\frac{|h_j|^2 P_{\text{tx},j}}{|h_i|^2 P_{\text{tx},i} + \eta} < \gamma_{\min}, \quad \forall i, j, \quad i \neq j. \tag{2.2b}$$

Condition (2.2a) assumes *high Signal to Noise Ratio (SNR) regime*, i.e., it guarantees that a single UE's SNR is always greater than the threshold, hence, any UE's packet can be decoded without interference. Condition (2.2b) tells us that if there is more than one UE using a RAO, none of the packets is decodable. It is implicitly assumed that the gNB is able to distinguish collided packets by using error detection methods, e.g., cyclic redundancy check. These conditions are justified for many scenarios, e.g., if channel coefficients of UEs are similar, or all UEs are sufficiently close to each other and the gNB. A simple 0/1 collision model without capture describing the success probability of

a packet transmission of the jth UE in a RAO based on conditions (2.2a)-(2.2b) is:

$$\mathbb{P}[\gamma_j \geq \gamma_{\min}] = \begin{cases} 1 & \text{if } n = 1, \\ 0 & \text{otherwise.} \end{cases} \tag{2.3}$$

The model provides a powerful abstraction for high SNR regime, and it is widely used in seminal works on random access protocols analysis [Bia00; Cap77; Riv87; WBC15]. In some scenarios, one or both conditions (2.2a)-(2.2b) do not hold, leading to such effects as *detection error* (condition (2.2a) violated), or *capture* (condition (2.2b) violated). In such cases, variations of collision model are often studied, e.g., channels with multi-packet reception [GVS88; JVK19]. In other cases, the time diversity can be utilized to cancel the interference of some users, which is a premise for SIC techniques and coded slotted ALOHA protocols [CGH07; Liv11]. Another possible variation of a collision channel model is unit-disk graph model [Wat+01], where the transmission power is determining a collision radius. It is often used for wireless ad-hoc networks for the analysis of multi-hop connectivity, where the geometry of a network must be considered.

2.3.3 M2M Performance Metrics in Random Access

In this section, we answer the question: What performance metrics are relevant for M2M random access? Most of them are inherited from traditional Human-to-Human (H2H) systems, however some metrics are novel and only arise from M2M use cases.

We separate the metrics into five groups: (I) Instantaneous performance, (II) Steady state performance, (III) Transient performance, (IV) Reliability-latency performance, (V) Application performance. Prior to introduction of M2M, mostly the first two groups had been used for random access protocols [BGH87]. While they remain relevant, the latter three groups become increasingly more important specifically for M2M.

I. Contention round outcome. Instantaneous performance metrics, characterizing the outcome of a single contention round, mostly correspond to performance metrics of

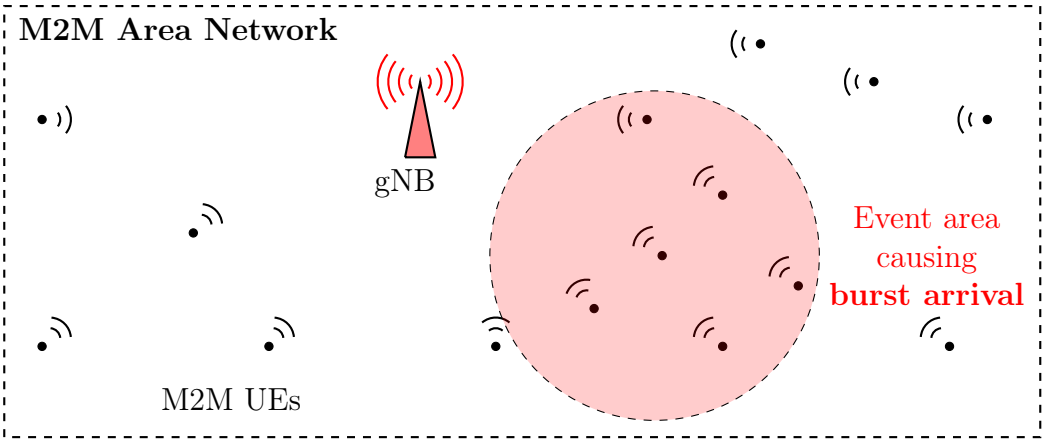

Figure 2.6: Illustration for the burst arrival triggered by an event in a specific area.

classical slotted ALOHA. These metrics include instantaneous throughput (the number of successful RAOs in a round), resource efficiency (ratio of successful RAOs in a round), number of collisions, and number of idle preambles. A less common metric related to the specifics of LTE and NR RAP, which we introduce in detail in Chapter 4, is instantaneous resource consumption.

II. Steady-state performance. Conventional random access protocols and their derivatives are often characterized in terms of their steady-state performance. This includes expected steady-state throughput, collision, and idle RAOs rate. In addition, steady-state performance is characterized by the expected delay of a packet. Delay is closely to related to the concept of stability of the protocols, i.e., if the expected delay is infinite, the protocol is considered unstable [BGH87]. If a retransmission limit is enforced, the resulting ratio of dropped packets is an important performance metric (see Chapter 3).

III. Transient performance. Classic random access protocols were initially designed to support sporadic independent transmissions from multiple sources. Naturally, this is the reason why Poisson traffic models were used. While Poisson models are still useful to characterize independent behavior of mM2M UEs, in certain cases correlated traffic patterns arise [3GP11]. For instance, consider a scenario of a power grid blackout in a neighborhood. After the power is back on, all the UEs in the area attempts to reconnect to the gNB. It creates a large burst of connection requests, potentially blocking Random Access CHannel (RACH) due to collisions for a longer time (see illustration in Fig. 2.6).

The change in the traffic pattern leads to important consequences for the random access protocols. First, as pointed out in early works [CS88], burst arrivals are significantly degrading the performance of the protocols, since correlation creates congestion which can persist for longer time. For independent arrivals, it is possible to average out the outliers and dimension the system for the averages. Here, however, long term averaging is likely to be insufficient. Hence, burst arrival scenarios call for a different analysis methodology and performance metrics. Instead of the steady-state, *transient behavior* is analyzed. The respective performance metrics account for burst resolution parameters: average throughput during the burst resolution, *burst resolution delay*, and the amount of resources spent on the burst resolution. We address transient performance aspects in Chapters 4-5.

IV. Reliability-latency perspective. As discussed in 2.1, for uM2M users, reliability is an important factor. Hence, performance metrics characterizing the expectations are not sufficient, and *reliability guarantees* must be provided. No deterministic guarantees can be obtained due to the stochastic nature of wireless channels and random access protocols. Instead, stochastic guarantees must be defined. E.g., given a per-packet delay d, and a delay requirement $\bar{d}$, we can guarantee that the probability of exceeding the delay is not larger than a certain value: $\mathbb{P}[d < \bar{d}] \leq \varepsilon$. The concept is generalized for the burst resolution delay in Chapter 5. For uM2M, traditional expectation-based optimization must be replaced by the optimization for an arbitrary stochastic Quality of Service (QoS) requirement. This requires a derivation of probability distribution of the respective metric, which is typically a considerably harder combinatorial task than just

computing expected performance. The reliability-delay perspective on RAP is addressed in Chapter 5.

V. Application Perspective. Performance metrics of groups I-IV can be used to qualitatively assess the performance of an application. It is intuitively clear that, for example, the protocols with low collision rates and low burst resolution time are at the same time providing good application performance. However, if the *cost* of providing low collision rate and burst resolution time is taken into account (e.g., resource consumption), then there is a need to define a utility function for arbitration between the cost and the protocol's performance. One approach for such arbitration is to use *application performance* as a utility function. While it is not possible to find the common metric for all M2M application, we will use metrics of a generalized class of Networked Control Systems, such as networked-induced error. The application performance is studied in detail in Chapter 6.

2.4 Overview of Recent Results

In this section, we review the general state of the art on random access protocols for M2M communications. First, we outline modeling and performance analysis studies. Then, we review the improvement proposals for RAP in LTE and NR. This section is dedicated to the main studies which are relevant throughout the whole thesis, and its content is based on the state of the art reviews conducted in our published work [Vil+17b; G¨+17b; VRK19a]. In addition to this review, every remaining chapter of the thesis gives an overview of the studies closely related to the content of the respective chapters.

2.4.1 Historic Perspective

Historically, random access protocols have been a subject of scientific study since 1970s, when ALOHA and Slotted ALOHA (s-ALOHA) have been introduced [Abr70]. Basic results on performance analysis, stability, and optimization techniques have been developed primarily in the 1970s-1980s by Rivest, Capetanakis, Gallager, Mikhailov and others [Riv87; Cap79; CS88; Mik79; Cap77]. A comprehensive summary of early ALOHA and slotted ALOHA research can be found in [BGH87, Chapter 4], where a variety of protocols on the basis of ALOHA are described. After the introduction of contention-based access in WLAN and such techniques as Carrier Sense Multipe Access with Collision Avoidance (CSMA/CA), Distributed Coordination Function (DCF), and their derivatives, the topic of random access protocols has been revisited again, resulting in a number of seminal works [Bia00] in the end of 1990s and early 2000s. Collision avoidance and contention resolution methods have also received significant attention in application to Radio Frequency IDendintification (RFID) networks [Shi+06; KCR10]. Specifics of M2M communications and the respective challenges, which we discuss in 2.2.2 and 2.3.3, have triggered a new wave of interest ot random access protocols [LAAZ14; AHK17; HHN13], in particular to multi-channel slotted ALOHA.

2.4.2 Modeling and Analysis

LTE RAP has been primarily studied as a derivative model of MS-ALOHA, with the number of preambles available is considered equivalent to the number of available channels and periodicity of PRACH is equivalent to a slot [RS90; And+15]. The performance of standardized LTE RAP has been studied for two major scenarios: steady-state, typically for Poisson arrivals, [Tya+15; Tya+17; NWK14; MM16; Gha+16], and transient state for burst arrivals [WBC15; Jia+17; Kos16; Che+15; LM+17].

Tyagi *et al.* [Tya+15] study LTE RAP with uniform back-off and without access barring and evaluate the impact of retransmission limits on the steady-state performance and equilibrium points. In the follow-up work, the authors study RAP with ACB [Tya+17], and demonstrate that ACB is responsible for "Poissonization" of the arrival process, converting any arrival pattern to a Poisson-like. Steady-state M2M random access behavior has also been analyzed in the context of heterogeneous networks with relays in [NWK14]. The effects of power ramping have been studied in [MM16; MM18]. The authors also followed up presenting a potential use of power ramping for prioritization in [MMA17]. Geometry of the network and the effects of spatial distribution of the devices have been analytically studied by means of stochastic geometry in [Gha+16].

While the above described papers focus on the average steady-state performance, Wei *et al.* [WBC15] present an exact analysis of the burst resolution process. They demonstrate that the analysis is complex and impractical for many scenarios and subsequently devise a drift approximation approach for simplified analysis. The essence of the approach is to approximate evolution of the backlog by its expected trajectory. The exact probabilistic analysis of LTE RAP is extended and elaborated in [Jia+17]. Additionally, the Extended Access Class Barring (EAB) for M2M is analyzed in detail in [Che+15]. In [Kos16], the authors zoom in on the burst resolution delay and derive a lower bound on its average. More recently, detailed performance evaluation of the standardized 3GPP ACB has been proposed in [LM+17].

2.4.3 Potential Improvements for M2M Random Access

The methods for RAP improvement can be roughly classified into two groups: MAC layer methods and non-MAC layer methods. Since RACH overload is essentially a MAC layer problem, the first group is the dominant one.

MAC layer methods have been primarily focused on the time domain. As a 3GPP approved overload mitigation solution, ACB has been introduced in LTE release 8 [3GP11], and later re-worked into EAB. ACB defines a specific barring probability parameter p_b, which is used by every UE to decide probabilistically whether or not to attempt a transmission. ACB is essentially smoothing the arrivals in time domain, avoiding high collision rate region. Building upon the standardized solution and re-using the concepts of from the seminal work on classical slotted ALOHA from Rivest [Riv87], the authors in [Jin+17; Dua+16] have proposed a dynamic ACB algorithm, adaptively modifying the access probability according to the traffic load, with the possibility to combine dynamic

allocation of preambles for M2M traffic with ACB. The algorithm relies on the estimation of the number of backlogged UEs n, and subsequent setting of the barring probability according to the rule $p_b = 1 - \min\{1, M/n\}$, where M is the number of available channels (preambles). The concept of *load estimation* is an important topic for random access protocols, since the amount of active users is typically unknown to the network.

A number of variations on ACB and EAB have been proposed in the literature. A load-aware scheme similar to [Jin+17] for dynamic ACB adjustment has been introduced in [Son+17]. A two-stage resource allocation via special MSG2 has been proposed in [MG16], where special MSG2 allows contention-based access to unscheduled PUSCH resources, thus improving total utilization. Extensions to ACB, allowing cooperation and load-balancing between neighboring gNBs have been proposed in [HWT14; Lie+12]. Going beyond pure ACB and considering other time domain contention parameters, the authors in [YFE12] have examined dynamic adjustments of the contention window and retransmission limit based on the current load, while the authors in [PL16] used the future load predictions to update the access barring parameters. In order to limit the cross-influence of M2M and H2H devices, M2M-specific back-off or variable access cycles for M2M-devices can be employed [HHN13]. The authors in [Lo+11] have suggested self-optimizing methods for PRACH resource allocation, combining different overload control strategies suggested by 3GPP. The work, however, only provides a heuristic way of determining contention parameters, and does not provide any simulation results to demonstrate its effectiveness. The authors in [YHH11] present an analytical framework of RAP optimization, where the preamble split between contention-based and contention-free RAP is adjusted according to the load. The authors in [WC15] propose a hybrid protocol which combines RAP and payload transmission, leveraging the fact that M2M devices are likely to have low-to-moderate payload volume.

Due to the similarity of RAP problem with slotted ALOHA, many methods from earlier studies on s-ALOHA have been adopted. A prominent example are Tree Resolution Algorithms (TRA), first introduced for slotted ALOHA in [TM78; Cap79]. Madueno *et al.* have analyzed tree-based collision resolution for LTE [MSP14]. In the follow-up work, they have also examined splitting the RACH cycle into a phase for estimating the number of arrivals, followed by a phase for serving the arrivals with tree algorithms [Mad+15]. [G¨+17b] proposes two hybrid collision avoidance-tree algorithms for LTE, combining TRA with pre-back-off and dynamic ACB and achieving a higher per-preamble throughput than dynamic ACB alone [Jin+17; Dua+16]. Analytical performance assessment of the multi-channel parallel TRA has been presented in [GAK17b], alongside with its applications to burst resolution in [GAK17a].

The second group of non-MAC layer solutions is very versatile. Pratas *et al.* [Pra+12] and Condoluci *et al.* [Con+16b; Con+16c] have investigated an expansion of the random access contention space through a combination of conventional preambles and codewords, thus using coding techniques. In [Ko+12], the authors explore how different UEs choosing the same preamble can be distinguished based on the timing advance, which allows to reduce the effect of collisions. Jang *et al.* [Jan+14] introduce a novel spatial group based RAP, expanding the available preamble space by exploiting the inverse relation between

worst delay profile difference among UEs and number of distinguishable preambles. Similarly, Kim *et al.* [KJS15] propose a spatial group based RAP with reusable preamble allocation, which effectively increases the preamble space if delay profile differences between spatial groups which are allocated the same preambles are assumed larger than multi-path delay spread. In [Pra+16], an approach to use frames composed of multiple successive PRACH slots is investigated. Random access with multi-user detection for multiple-antenna OFDMA has been developed in [BJR17].

2.4.4 Summary

Despite the wide variety of existing works on M2M random access, plenty of challenges remain still open. From the above overview, we can distill following observations:

- The diversity of M2M applications, and their co-existence with H2H, calls for more emphasis on priorization in RAP.
- MAC layer solutions are largely focused on time domain, often ignoring other resource dimensions as a tool for improvement.
- Both analysis and improvement techniques target *expected performance* and ignore strict reliability requirements of uM2M.
- The impact of random access on application performance is studied only implicitly via MAC layer performance metrics. An explicit study of application performance is needed.

Our contributions in the following chapters largely follow these observations from the literature review. In every chapter, we dive deeper into these topics and additionally provide reviews of related work on each of them: Chapters 3 reviews the topic of prioritization (3.2). Chapter 4 reviews the works on resource consumption and the use of resource dimension for RAP (4.2), and Chapter 5 focuses on RAP reliability aspect (5.2). Finally, Chapter 6 provides an overview of application-aware RAP (6.2).

As a final remark, we note that the topic of M2M random access is certainly not limited to RAP, and other research directions can be reviewed here. This includes *grant-free access* as an optional feature of NR [3GP17a], random access in LoRa [Lora] and other standards, M2M via satellites [DS+15], and many more. However, these topics are out of the scope of the thesis, and we refer the reader to the respective literature.

Chapter 3

Resource Allocation and Aggregation for Steady-State Random Access

Deployment of massive Machine-to-Machine (M2M) communication brings two different traffic models for Random Access CHannel (RACH): burst arrivals with large amount of correlated requests and Poisson arrivals with independent requests from uncorrelated sources. Intuitively, the first model represents an emergency situation, extraordinary event in the network, while the second model represents traffic generated during a day-to-day operation of a large group of independent sensors. The first scenario, truly new for RACH, has attracted a lot of attention from the research community[1]. However, the second scenario could also present a significant burden on the RACH, since the massive number of devices raises the basic load level on RACH compared to the typical cell nowadays. The amount of the requests generated by the normal operation of the M2M User Equipments (UEs) creates a high total load composed of independent requests. Due to the sporadic nature of message transmissions in many M2M applications it is not prudent to keep radio resources continuously reserved. Instead, most M2M devices must complete the Random Access Procedure (RAP) before sending a message.

For the system to efficiently support the cumulative load, an appropriate amount of *resources*, or Physical Random Access CHannel (PRACH) preambles, must be allocated. The amount of preambles can be adjusted by changing the PRACH configuration index (hence, allocating more instances of PRACH in the frame), or by adjusting the split between the preambles reserved for contention-based and contention-free access. In addition to that, whenever there are multiple Quality of Service (QoS) classes in the system, separating the preambles becomes an efficient tool for prioritization. Unlike probabilistic prioritization, such as Access Class Barring (ACB) and Extended Access Class Barring (EAB), resource separation can provide full isolation between different QoS classes. Consequently, we formulate the first and second research questions addressed in this chapter: **(i)** How to **allocate the preambles** in order to maximize the RACH performance? **(ii)** How can preamble split and **preamble allocation be used for prioritization**?

In some cases, however, the existing amount of PRACH preambles is insufficient, e.g., in order to support very high load in large cells. In that case, approaches to reuse the preambles or expand their set come into play. A promising way to do it is to introduce intermediate aggregators, collecting connection requests and forwarding them to the Next

[1] Chapters 4 and 5 of the thesis are concerned with this scenario.

Generation Node B (gNB). On the one hand, aggregation helps offload the gNB, on the other hand, it also adds extra delay for collecting the requests, and this trade-off has to be carefully evaluated. We take on this evaluation as a third research question in this Chapter: **(iii)** how does the **aggregation impact the performance** of LTE RACH?

3.1 Contributions and Structure of the Chapter

First, in Sec. 3.2, we detail the prior work on RACH. We separately review the literature on prioritization through random access procedure manipulation, preamble separation, and aggregation for RACH.

In the first main part of the chapter (Sec. 3.3), for the constant traffic setting, we examine the effects of preamble separation on the RACH throughput, delay, and request drop ratio for two UE request classes. Class I represents delay-intolerant UE requests and class II represents delay-tolerant UE requests. One can imagine a class mapping to Quality of Service Class Indicator (QCI) classes [3GP15a] or a mapping to Human-to-Human (H2H) and M2M devices [AAF16; Fod+16; LKY11]. We quantify the throughput, delay, and drop ratio trade-offs of separating the preambles into two disjoint sets. For underloaded systems, we find that there is a "safe" allocating region, where class I prioritization is relatively harmless for class II. Also, we quantify an allocation region where the overall throughput is increased due to preamble separation. Based on these insights, we develop the Load-Adaptive Transmission-MAximizing Preamble Allocation (LATMAPA). LATMAPA is based on a throughput maximization principle and automatically adapts the number of preambles allocated to the high- and low-priority classes according to their load levels. Our evaluations indicate that LATMAPA effectively ensures high throughput as well as low delays and drop probabilities for the high priority class across a wide load range.

In the second main part of the chapter (Sec. 3.4), we address a scenario where the total amount of resources is insufficient, and the intermediate aggregation is used to support larger number of UEs. We study the medium access aspects of the cluster-based aggregation scheme for connection establishment of M2M UEs. Our contributions are (i) analysis of the aggregation process and a study on how it influences the connection from clusterhead to gNB, and (ii) an accurate joint medium access model of the RAP *within a finite-user cluster*, considering the cross-impact and interrelation between aggregation and random access procedures. A byproduct of the joint analysis is an accurate finite-user steady-state model of the RAP without aggregation. The models are verified with simulations and compared to the state-of-the-art. They allow accurate performance predictions and provide insights on the dimensioning and resource allocation for clusters.

The content of this chapter is based on our work published in [Vil+17b] and [VK17b].

3.2 Related Work

Random access in cellular networks has been studied from a variety of angles. In the following, we briefly review the categories most closely related to our study. For a general overview of the related work, we refer the reader to the earlier Sec. 2.4. We first review studies on QoS provisioning and prioritization in LTE and NR random access. Then, we review the studies about the aggregation in random access.

3.2.1 Prioritization through Random Access Procedure Manipulation

Several studies, e.g., [CYZ03; RALRCP09; SL11; Xia05], have investigated random access prioritization through manipulations of the random access contention procedures or parameters, such as transmission attempt limit and backoff window duration, on a given set of preambles. Moreover, as a refinement of ACB, EAB has been introduced by Third Generation Partnership Project (3GPP) in Release 11 [3GP11]. EAB enables prioritization through assigning different barring probabilities to the different UE classes [AG17; ZGA16]. The adjustment of the random access contention, e.g, through EAB, on a given set of preambles is complementary to our approach of conducting the random access contention of the different priority classes on separate sets of preambles. In particular, the random access contention could be differentiated within a given preamble set to achieve further QoS differentiation. Generally, methods that manipulate the random access contention, such as EAB, are designed for non-persistent temporary UE request traffic burst [Dua+16]; Whereas we focus on persistently high UE request traffic loads.

3.2.2 Prioritization through Preamble Separation

A few prior studies have examined different forms of preamble separation. In particular, some studies have split the preambles into distinct sets for contention-based random access and for non-contention (dedicated) access [HLR11; KKA13; YHH11]. Chu *et al.* have developed a general model of resource allocation in slotted ALOHA (whereby a preamble can be considered a resource) through a matrix representation [Chu+15]. Complementary to these studies, we focus on contention-based random access.

Initial studies of the prioritization of contention-based random access through separating preambles have been conducted by Lee *et al.* [Lee+12], Kalalas *et al.* [KVGAZ16], and Lin *et al.* [CLL11; Lin+14]. The prioritization through preamble separation has also been covered in the patent [CY+14]. These initial studies have only examined throughput for pre-configured fixed static preamble separation. In contrast, we consider dynamic adaptive preamble separation according to the traffic loads for the priority classes according to the LATMAPA approach introduced in this study. Moreover, we conduct an in-depth evaluation of LATMAPA that considers throughput, delay, and drop probabilities.

Zhao *et al.* [ZZF14] have proposed a heuristic load-adaptive preamble allocation rule,

which we consider as a benchmark in our evaluations, see Section 3.3.5.3. Zhao *et al.* have incorporated the heuristic preamble allocation rule into an overall protocol with a variant of binary exponential backoff. In this study, we focus on examining the effects of preamble allocation for prioritizing random access. We do not vary the backoff process; rather we consider the standard LTE and NR uniform random backoff throughout.

Du *et al.* [Du+16] have proposed an approach for PRACH resource allocation, aiming at minimizing the contention resolution time. The approach relies on real-time knowledge of the number of contending UEs in every PRACH slot, and on numerical solvers for calculating the optimal split for certain load values. In contrast, LATMAPA requires only average load as an input, and provides a closed-form expression for the optimal split. We compare LATMAPA to the approach by Du *et al.* [Du+16] in Section 3.3.5.3.

3.2.3 Aggregation for Random Access

As mentioned above, congestion control methods are suited only to deal with the temporary overload in the channel, and do not bring a qualitative change in PRACH capacity. The authors in [Jan+14] have shown that the spacial distribution of devices can help to increase PRACH resource capacity. Several more invasive and not standard-compliant methods have been identified as mM2M-enablers for 5G networks. Among them, there are novel access schemes, such as Non-Orthogonal Multiple Access (NOMA), successive interference cancellation, and aggregation-based medium access [ICT15]. Data aggregation models for M2M have been already studied in [Sha+15; TST12; Meh+15]. Differently from them, our work tackles *aggregation of connection requests*, that is, assuming that a cell has enough uplink resources to meet the demands, but not enough PRACH resources for all the M2M devices to establish a connection. Possible RACH throughput gains from clustered aggregation have been simulatively studied in [Wan+13], and a similar Device-to-Device (D2D) based group paging has been introduced in [DW15]. However, still missing is the analytical modeling and understanding on how the connection request aggregation process influences RAP performance.

3.3 LATMAPA: Load-Adaptive Throughput MAximizing Preamble Allocation

In this section, we approach the first two research questions of this chapter, (i) how to allocate the preambles to maximize the performance of the system, and (ii) how to use preamble allocation for prioritization. We develop a throughput maximizing preamble allocation policy, and then study its impact and applications for prioritization. The section is organized as follows. We recap the basics of connection establishment procedure in 3.3.1. Sec. 3.3.2 gives the background on the concept of preamble separation. Sec. 3.3.3 analyzes how the number of allocated preambles affects the RACH performance for individual classes and for the entire system over a range of UE request loads. Sec. 3.3.4 examines preamble allocation methods that strive to meet a delay target or to maxi-

mize throughput; the throughput maximization approach results in the Load-Adaptive Throughput MAximizing Preamble Allocation (LATMAPA) policy. Sec. 3.3.5 evaluates LATMAPA through analysis, simulations, and benchmark comparisons.

3.3.1 Connection Establishment in LTE and NR: Recap

We briefly summarize the RAP in this section. For the detailed description of the procedure, we refer the reader to Chapter 2 and respective illustrations 2.4 and 2.3. The procedure of establishing a radio interface connection from a UE to the gNB, also known as RAP, is performed if the UE is newly connecting to the network, or has remained inactive for a sufficiently long time (regulated by UE Inactivity Timer). RAP starts with the UE listening for broadcast messages of the gNB, in which the latter advertises synchronization data and parameters of the PRACH. After discovering the PRACH parameters, the UE knows when to attempt a connection, which frequency resources to use, and which preamble sequences are available for sending as a first handshake message. Preamble sequences are codewords used to extend PRACH capacity: the gNB is able to detect all activated preambles, hence, a collision is only occurring if multiple UEs choose the same preamble during the same slot. In the following we adopt the notion that a time x frequency x preamble is called Random Access Opportunity (RAO). After detecting activated preambles, the gNB replies with RA reply message, which contains information when to send the actual Radio Resource Control (RRC) Connection Request. Then, the gNB either confirms with the reply that the connection is established, or, in case a collision has occurred, the UE waits for a timeout before going to attempt the connection establishment once more after a back-off [Tya+15].

3.3.2 Preamble Separation

3.3.2.1 Preamble Assignment Options

In general, several options of allocating preambles can be considered. Conventionally, there is **no separation**, meaning all devices compete in the entire set of preambles and can collide with each other. Another option is fixed, **non-overlapping assignment**, where both classes have their own preamble set, thus, competing only with the devices from the same class. **Overlapping assignment** [CLL11; Lin+14] assumes that prioritized UEs can compete in the entire set, whereas non-prioritized UEs can only use a predefined fraction of the preambles.

In this work, we compare the steady-state performance of the system for the no separation and non-overlapping assignment allocation options. The separation of the preambles into two sets involves a number of trade-offs. By allocating more preambles to class I, we degrade the performance of class II.

Table 3.1: Summary of model notations for Chapter 3.

$\mathbf{M}$	Set of all preambles available in each slot
M	Total number of preambles available in each slot (= 54 if not stated otherwise)
$m_I,\ m_{II}$	Numbers of preambles allocated for class I, class II
W	Maximum number of allowed transmission attempts (= 8 if not stated otherwise)
$B_{\max}$	Max. back-off value in slots (= 20, default)
$\lambda_I,\ \lambda_{II}$	Poisson arrival rates of class I, class II UE req./slot
$\rho = \frac{\lambda}{M}$	Poisson process arrival rate, normalized for one preamble
f	Steady-state UE request success probability in one attempt
x	Expected number of UE request (incl. initial arrivals + retransmissions) contending for preambles in a slot.
T	Steady-state throughput of UE request per preamble per slot
D	Average steady-state delay (in slots)
δ	Steady-state UE request drop prob., after max. of W attempts
$\hat{\rho}$	Normalized Poisson process arrival rate achieving the max. throughput, referred to as *peak throughput load*
$\hat{D}$	Steady-state del. (in slots) for successful UE request if $\rho = \hat{\rho}$
$\hat{\delta}$	Steady-state ratio of dropped UE request if $\rho = \hat{\rho}$

3.3.2.2 Modeling RACH with Preamble Separation

Generally, the RACH can be represented as a multichannel slotted Aloha system, with a slot representing one time-domain RACH opportunity, and a channel representing one RACH preamble [AK16; Tya+15]. In our model, we consider two device classes, both with an infinite number of UEs (infinite source model) and finite total request arrival rates. That is, the numbers of arriving requests per slot are modeled by independent Poisson distributions, with the expected values λ_I and λ_{II} for class I (delay-intolerant devices) and class II (delay-tolerant devices), respectively.

The UEs of both classes attempt to send a RACH MSG1, which consists of a RACH preamble chosen uniformly out of the available sets $\mathbf{M_I}$ and $\mathbf{M_{II}}$, respectively, as illustrated in Fig. 3.1. Any request that has collided in a first transmission attempt is re-transmitted again up to the maximum of W transmission attempts. The re-transmission proceeds after a back-off time that is uniformly chosen from the interval 0 to $B_{\max}$. If a

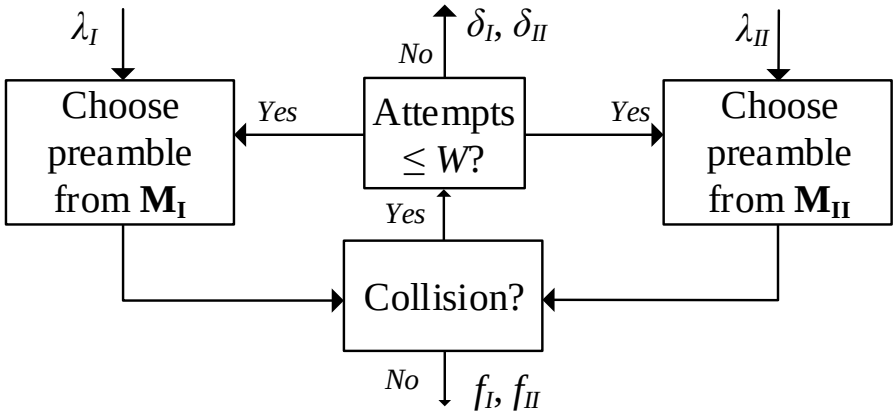

Figure 3.1: Illustration of two-class fixed-assignment RACH model with preamble sets $\mathbf{M_I}$, $\mathbf{M_{II}}$: UE requests arrive with rates λ_I and λ_{II} for the two classes and select preambles from their respective fixed-assigned sets $\mathbf{M_I}$ and $\mathbf{M_{II}}$. Preamble transmissions without a collision result in successes. Collided preamble transmissions are retransmitted until W attempts are reached and then dropped (if the Wth attempt collides).

request has collided W times, it is considered as dropped. We denote δ for the request drop probability. We denote f for the probability of success in one attempt. The average delay D measures the average number of slots from the first request transmission attempt until the successful reception of the request. Note that the delay D does not take the unsuccessful (dropped) requests into account.

Since we consider infinite sets of devices, the arrival rates of the initial (new) requests remain constant, while the retransmissions increase the total number of UEs attempting access up to x for the steady-state [Tya+15]. A summary of system model notations in presented in the Tab. 3.1. We note that some MAC and physical layer considerations have not been captured in our system model, since we focus on the preamble contention aspect. We acknowledge that, in general, the neglected parameters, such as UE location [Ko+12], inter-cell interference, or access barring [Tya+17], can influence quantitatively influence the RACH behavior.

3.3.3 Analysis of Random Access System

In this section, we analyze the influence of the numbers of preambles assigned to the two classes on the key performance metrics throughput, delay, and drop ratio. Initially, as groundwork, we analyze the random access system without preamble separation. Then, we proceed to examine preamble separation.

3.3.3.1 Overall System Without Preamble Separation

3.3.3.1.1 Review of Steady State Analysis. Utilizing the notation summarized in Table 3.1, we first briefly review the steady-state analysis of the system without pream-

ble separation [SH00; Tya+15]. In steady-state,

$$f = e^{-\frac{x}{M}} \text{ and} \tag{3.1}$$

$$\frac{x}{\lambda} = \frac{1-(1-f)^W}{f}. \tag{3.2}$$

As there is no closed-form solution for Eqn. (3.2) with respect to x, f, numerical methods have to be used to obtain f and x from the system of Eqns. (3.1), (3.2). The obtained f and x values are used to calculate the performance metrics as [SH00; Tya+15]:

- Drop ratio δ: ratio of the requests that did not succeed in any of the W transmission attempts to the total number of initial requests transmitted:

$$\delta = (1-f)^W. \tag{3.3}$$

- Throughput T: ratio of successfully received requests to the total number of transmission opportunities:

$$T = \frac{\lambda}{M}(1-\delta). \tag{3.4}$$

- Delay D: time period from the first transmission attempt until the request is successfully received by the gNB . Since the number of PRACH slots in a given LTE frame depends on the PRACH configuration, we measure the delay in units of PRACH slots:

$$D = \left(1+\frac{B_{\max}}{2}\right)\frac{1}{f-1} \times \frac{1+(W-1)(1-f)^W - W(1-f)^{W-1}}{1-(1-f)^W}. \tag{3.5}$$

The resulting dependency of the total throughput and drop ratio on the total normalized load ρ is depicted in Fig. 3.2. We observe that there are two distinct operating regions: an underloaded region to the left of point **A** in Fig. 3.2, and an overloaded region to the right of point **A**. The underloaded region is characterized by linear increase of the throughput and steady low drop ratio. On the other hand, in the overloaded region, the drop ratio increases rapidly as the throughput drops.

Our hypothesis is that the preamble separation into two device classes has different effects and involves different trade-offs depending on whether the total system load is in the underloaded or overloaded region. Hence, it is important to exactly know the load value at the border between these two regions. Therefore, we find in the next subsection the normalized load value $\hat{\rho}$ corresponding to the maximum throughput at point **A** in Fig. 3.2.

3.3.3.2 Peak Throughput Load

Considering the total normalized load $\rho = \lambda/M$, we evaluate the load value $\hat{\rho}$ that achieves the peak throughput, i.e., corresponds to point **A** in Fig. 3.2.

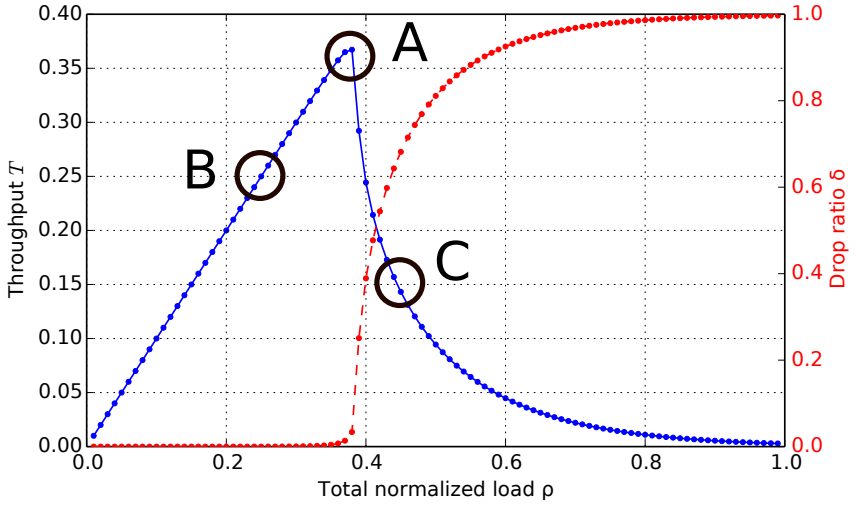

Figure 3.2: Throughput T and drop ratio δ vs normalized arrival rate ρ in the system without preamble separation, $W = 8$ transmission attempts.

Theorem 1. *The peak throughput load, i.e., the (normalized) load value which achieves the maximum normalized throughput is found as:*

$$\hat{\rho} = \frac{1}{e(1-(1-1/e)^W)}. \tag{3.6}$$

Proof. The theorem is proven by analyzing the function $T(\rho)$ given by Eqn. (3.4). After solving Eqn. (3.1) for x and substituting it in Eqn. (3.2), considering that $\rho = \lambda/M$, we obtain:

$$\rho = \frac{f\ln(f)}{(1-f)^W - 1}. \tag{3.7}$$

From Eqns. (3.4) and (3.7):

$$T = -f\ln(f). \tag{3.8}$$

Now, we can find the value of f maximizing the throughput through differentiation

$$\frac{dT}{df} = f\frac{d(\ln(f))}{df} + \ln(f) = \ln(f) + 1 \tag{3.9}$$

and setting Eq. (3.9) to zero. Thus,

$$f = 1/e \tag{3.10}$$

attains the maximum throughput. By substituting $1/e$ for f in Eqn. (3.7), we obtain the *peak throughput load* as in Eqn. (3.6). □

We observe that $\hat{\rho}$ depends only on the number of allowed transmission attempts W, and asymptotically reaches $1/e$ for $W \to +\infty$, see Fig. 3.3.

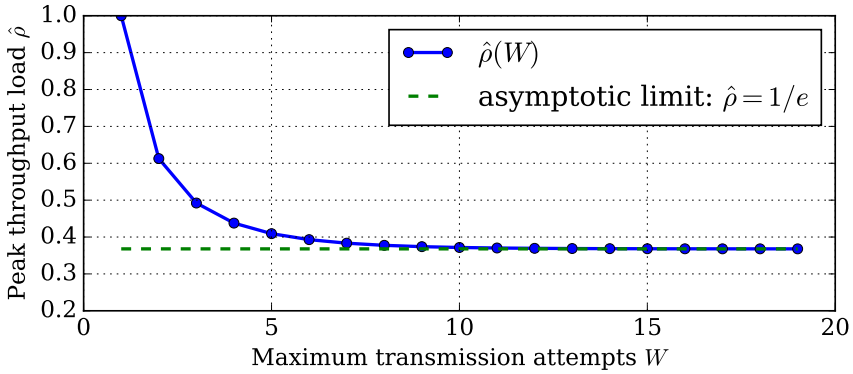

Figure 3.3: Peak throughput load value $\hat{\rho}$ achieving maximum throughput as a function of maximum number of transmission attempts W.

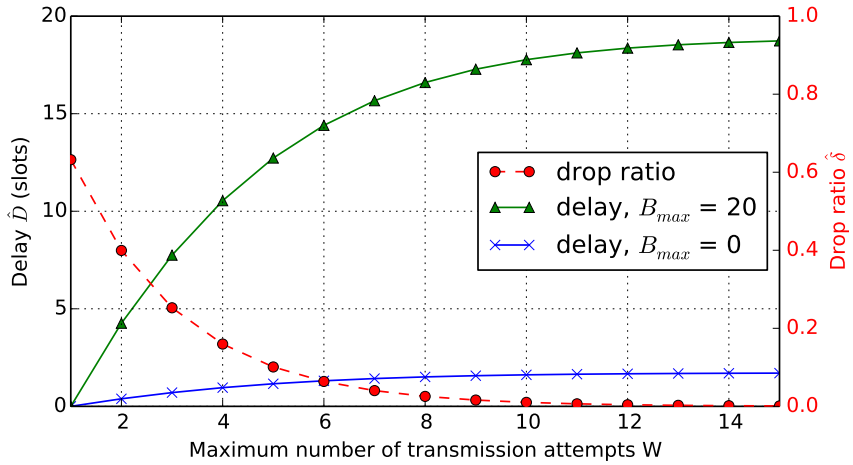

Figure 3.4: Delay $\hat{D}$ and drop ratio $\hat{\delta}$ achieved for the maximum throughput (peak load $\hat{\rho}$) vs. maximum number of transmission attempts W, for different values of $B_{\max}$.

Corollary 1. *For the peak throughput load, the steady-state performance of the system is:*

$$\hat{D}=\left(1+\frac{B_{\max}}{2}\right)(e-1)\frac{1+(W-1)(1-1/e)^W-W(1-1/e)^{W-1}}{1-(1-1/e)^W} \tag{3.11}$$

$$\hat{\delta}=(1-1/e)^W, \tag{3.12}$$

$$\hat{T}=1/e. \tag{3.13}$$

Proof. Obtained by substituting f with (3.10) in Eqns. (3.5) and (3.3). □

The resulting dependencies are presented in Fig. 3.4. It is intuitively clear that increasing W increases the delay, while decreasing the drop ratio. Note that the steady-state throughput does not depend on the $B_{\max}$ with our model assumptions. (This observation does not hold for the general case with finite number of UEs in the cell or with a varying arrival rate λ, see Section 3.4.)

From Fig. 3.4, we also observe that, if the parameters $B_{\max}$ and W are properly chosen (e.g., $W = 8$, $B_{\max} = 0$), the system performance at the peak load point is characterized by moderate delays and low drop probability.

3.3.3.3 Prioritization with Preamble Separation

We now proceed to analyze the fixed-assignment preamble separation, i.e., we examine the split of the preambles into two non-overlapping sets $\mathbf{M_I}$ and $\mathbf{M_{II}}$. Since the two sets are non-overlapping, we can consider them as two independent systems. Thus, their performance metrics can be obtained via Eqns. (3.1)–(3.5), whereby we replace M in Eqns. (3.1) and (3.4) by m_I and m_{II}, respectively.

In the next subsections, we examine the separation effects in the underloaded and overloaded regions with two example cases of the total initial arrival rate: $\rho = (\lambda_I + \lambda_{II})/M = 0.25$ (point **B**) and $\rho = (\lambda_I + \lambda_{II})/M = 0.45$ (point **C** in Fig. 3.2). For ease of illustration, we set the absolute arrival rates of both classes to be equal, i.e., $\lambda_I = \lambda_{II}$.

3.3.3.3.1 Underloaded Region. The plots in Fig. 3.5 represent the performance of the RACH for class I and class II with a fixed preamble assignment (separation), where m_I (on the x-axis) is the number of preambles assigned to class I; thus, $m_{II} = M - m_I$ preambles are assigned to class II. The throughput is normalized with respect to all $M = |\mathbf{M}|$ available preambles. We observe from Fig. 3.5a, that for the underloaded case there exists a region $m_l \leq m \leq m_r$ where the total throughput with preamble separation matches exactly the total throughput without separation. This region is bounded by the number of preambles m_l and m_r achieving the peak throughput of class I and II respectively:

$$m_l = \left\lceil \frac{\lambda_I}{\hat{\rho}} \right\rceil \text{ and } m_r = \left\lfloor M - \frac{\lambda_{II}}{\hat{\rho}} \right\rfloor . \tag{3.14}$$

(3.15)

Hence, the width Δm of this region is:

$$\Delta m = m_r - m_l = M - \left\lceil \frac{\lambda_{II}}{\hat{\rho}} \right\rceil - \left\lceil \frac{\lambda_I}{\hat{\rho}} \right\rceil . \tag{3.16}$$

The region width Δm is zero, if $(\lambda_I + \lambda_{II})/M = \hat{\rho}$, i.e., when the total load equals the peak load; which corresponds to point **A** in Fig. 3.2. Figs. 3.5b and 3.5c indicate that prioritization within this region moderately decreases the delay in one class, while keeping the drop ratio very low. Even though underloaded systems do not pose performance challenges for the RACH in practice, our analysis shows that an efficient delay-targeted prioritization for class I can be performed within the region $m_l \leq m \leq m_r$ without a significant performance degradation for class II.

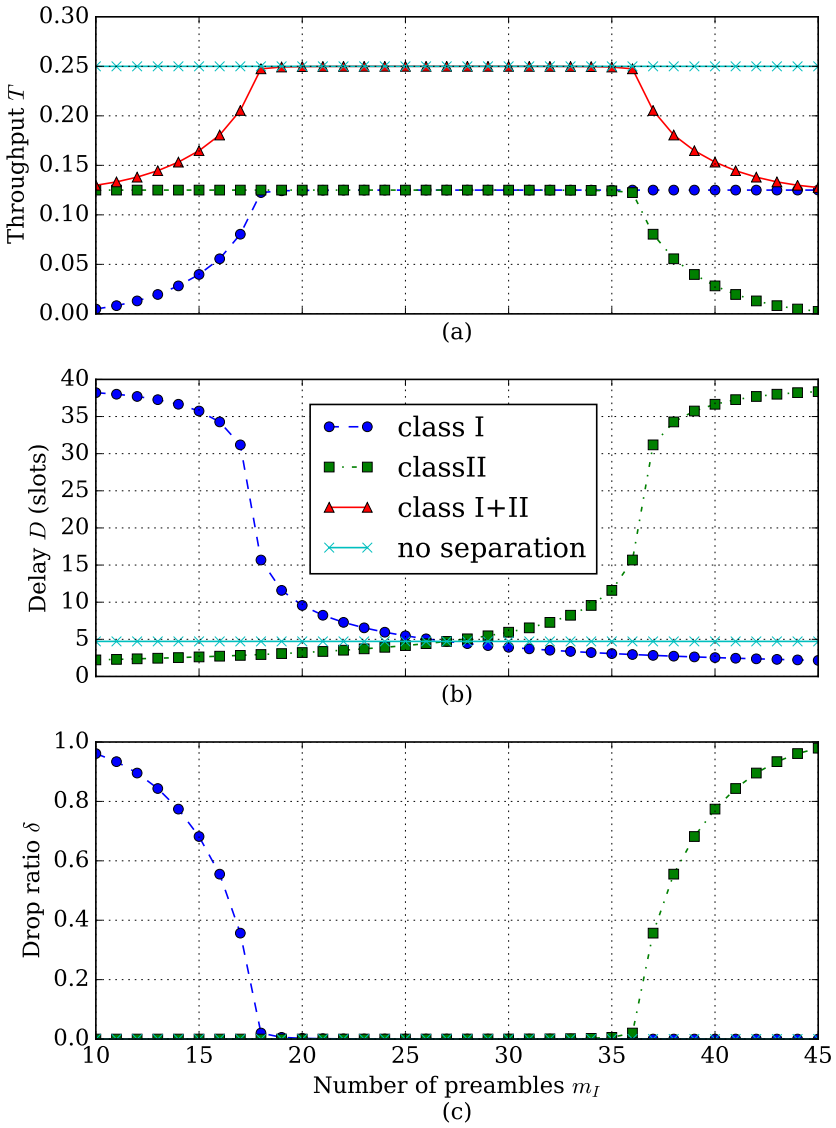

Figure 3.5: System performance vs. number m_I of preambles allocated to class I, for underloaded $\lambda = \lambda_I + \lambda_{II} = 0.25M$ scenario (point **B** in Fig. 3.2). Fig. (a) shows throughput T, Fig. (b) delay D, and Fig. (c) shows drop probability δ. System parameters: $M = 54$ preambles, $B_{\max} = 20$ slots, $W = 8$ transmission attempts.

3.3.3.3.2 Overloaded Region. We first observe for the overloaded region in Fig. 3.6(a) that the total (aggregate) system throughput (of both class I and class II) is higher or equal to the throughput without preamble separation. There are two throughput peaks on the plot, corresponding to $m_I = 22$ and $m_I = 32$. These peaks correspond to preamble allocations maximizing the throughput of class I and class II respectively (i.e., point **A** in Fig. 3.2). Since $\lambda_I = \lambda_{II}$, the peaks are of equal magnitude.

For a general case, the total throughput is calculated from the throughputs of class I

T_I and class II T_{II} as follows:

$$T = T_I m_I + T_{II}(M - m_I). \tag{3.17}$$

The maximum total throughput depends on the λ_I/λ_{II} ratio. If $\lambda_I/\lambda_{II} > 1$, then the maximum total throughput corresponds to the point when the number of preambles allocated to class I maximizes its performance. The maximum achievable total throughput is $T^{\max} = 1/e$, and is possible whenever $T_I = \hat{T}$, and $\lambda_I = \hat{\rho}M$.

Now we turn to the delay and drop ratio of the overloaded region. From Fig. 3.6(b), we conclude that any prioritization of class I (for $m_I > 27$) results in a significant delay decrease for class I, with a slight delay increase for class II. Importantly, the delay reduction for class I comes at the expense of an increased drop probability for class II, as shown in Fig. 3.6(c).

3.3.4 Preamble Allocation Methods

The goals of prioritization on the RACH can be both to increase the number accepted UEs (throughput), as well as to decrease the access delay. In this section we consider two approaches for calculating the number m_I of preambles for the prioritized class I: based on delay requirement matching and based on throughput maximization.

3.3.4.1 Matching the Target Average Delay

If the devices in a delay-intolerant class have a common delay requirement, then it can be beneficial to dimension the RACH according to this requirement denoted as $\bar{D}$. In this study, we consider delay in slots, therefore translation into the actual time domain requires knowledge of the PRACH configuration parameters. For instance, the PRACH configuration index 7 [3GP15b], results in one RACH opportunity per frame; thus, the length of one slot is 10 ms. Following the analysis in Sec. 3.3.3, we can calculate the required minimum number of preambles $m_I^{\min}$ in order to achieve a target average delay. Specifically, substituting x obtained from Eqn. (3.2) into Eqn. (3.1), and solving the resulting expression for m gives

$$m_I^{\min} = \left\lceil \lambda_I \frac{(1-f)^W - 1}{f \ln(f)} \right\rceil. \tag{3.18}$$

Eqn. (3.18) establishes the relation between $m_I^{\min}$ and f. Now, to obtain $m_I^{\min}$ as a function of a given delay requirement $\bar{D}$, Eqn. (3.5) must be solved for f. There is no closed-form relation between f and a given delay requirement $\bar{D}$, however, Eqn. (3.5) can be solved numerically for f given $\bar{D}$.

The method of using target delay for allocating preambles suffers from several drawbacks, which are illustrated in Fig. 3.7 where $m_I^{\min}/M$ is plotted as a function of the delay requirement for different load values λ_I. First, the delay parameter does not account for dropped requests δ, and, thus, does not represent a good standalone metric for

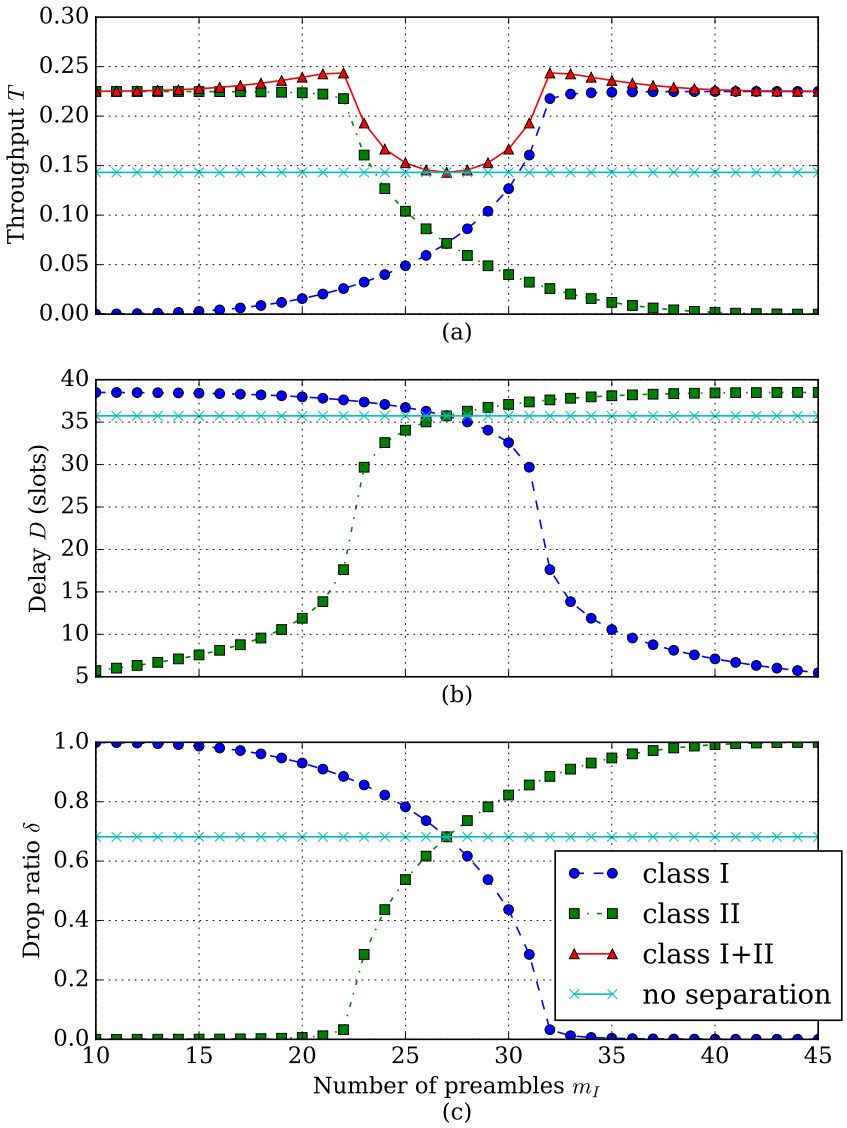

Figure 3.6: System performance vs. number of preambles allocated to class I m_I, for overloaded $\lambda = \lambda_I + \lambda_{II} = 0.45M$ scenario (point **C** in Fig. 3.2). Fig. (a) shows throughput T, Fig. (b) delay D, and Fig. (c) shows drop probability δ. System parameters: $B_{\max} = 20$ slots, $W = 8$.

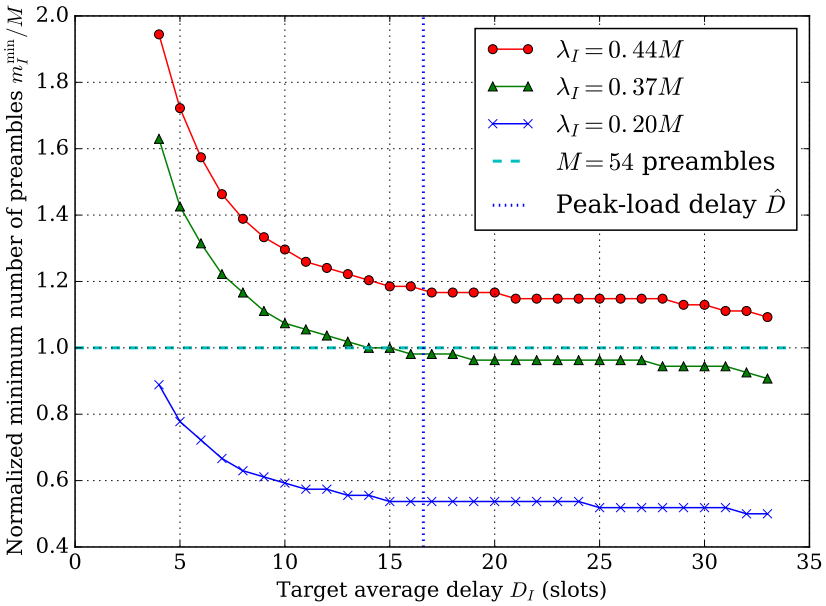

Figure 3.7: Normalized minimum number $m_I^{\min}/M$ of preambles necessary for allocation to class I (y-axis) in order to meet the delay requirement (x-axis, in slots) for different load values λ_I expressed as a fraction of the total number $M = 54$ of available preambles. The region above the $M = 54$ line is not achievable due to the insufficient available preambles. The region right to the peak-load delay $\hat{D}$ line is achieved in the overloaded system with high drop ratio δ. System parameters: $B_{\max} = 20$ slots, $W = 8$.

the performance: for given system parameters W and $B_{\max}$, the target delay requirement can be located in the overloaded region (on the right from the peak-load delay $\hat{D}$ in the Fig. 3.7) and, thus, can be accompanied by a high drop ratio. If $m_I^{\min} > M$ (see $M = 54$ line in Fig. 3.7) the target delay cannot be achieved at all for a given λ_I. Moreover, since $B_{\max}$ has no influence on the throughput or drop ratio (see Sec. 3.3.3.2), a better adjustment for the average delay can be achieved through a proper $B_{\max}$ setting.

3.3.4.2 Throughput Maximization: LATMAPA

Alternatively, the preamble-based prioritization can target the throughput (and corresponding drop ratio) as performance metric. The goal for setting the minimum necessary number of preambles $m^{\min}$ is to keep the throughput of the corresponding class at its highest value.

Corollary 2. *The minimum necessary number of preambles $m^{\min}$ to maximize the throughput of the corresponding class is found as:*

$$m^{\min} = \lceil \lambda e(1 - (1 - 1/e)^W) \rceil. \tag{3.19}$$

Proof. The maximum throughput per preamble is achieved if λ, normalized by the num-

Algorithm 1 Load-Adaptive Throughput-MAximizing Preamble Allocation (LATMAPA).

procedure LATMAPA
UE req. arrival rates: λ_I for high prior. class I; λ_{II} for low prior. class II;
RACH parameters: W transm. attempts, M preambles;
Prioritization factor r, $r \in [0,\ 1]$;
Calculate $m_I^{\min}$, $m_{II}^{\min}$ for λ_I, λ_{II} via Eqn. (3.19)
    if $m_I^{\min} \leq M - m_{II}^{\min}$ **then**
        $m_{II} \leftarrow m_{II}^{\min}$; $m_I \leftarrow M - m_{II}$
    else
        $m_{II} \leftarrow \max\left(\left\lceil \frac{M r m_{II}^{\min}}{m_I^{\min} + m_{II}^{\min}} \right\rceil,\ M - m_I^{\min}\right)$
        $m_I \leftarrow M - m_{II}$
    end if
return Preamble numbers for classes I and II: m_I, m_{II}
end procedure

ber of allocated preambles $m^{\min}$, is equal to the peak throughput load $\hat{\rho}$ (3.6), i.e.,

$$m^{\min} = \left\lceil \frac{\lambda}{\hat{\rho}} \right\rceil. \tag{3.20}$$

From Eqn. (3.20) and (3.6) we obtain Eqn. (3.19). □

As shown in Fig. 3.2, the drop ratio can be kept low as long as the throughput of class I remains less than or equal to the peak throughput. However, if we allocate more than $m_I^{\min}$ preambles to class I, the overall throughput decreases while having almost no effect on the throughput and drop ratio of class I. Thus, by choosing $m_I > m_I^{\min}$, the performance of class II is unnecessarily degraded.

Following these observations, we propose the Load-Adaptive Throughput MAximizing Preamble Allocation (LATMAPA) for determining the necessary amount of preambles (see Algorithm 1). LATMAPA requires UE request arrival rate estimates which can be obtained with combinations of existing short [Mad+15] and long [Cho+11; LCW16] timescale prediction techniques. The core idea of LATMAPA is that for the given arrival rates λ_I, λ_{II} we calculate the respective necessary number of preambles $m_I^{\min}$, $m_{II}^{\min}$ using Eqn. (3.19). If there are enough resources to meet the demand of both classes (underloaded case, Section 3.3.3.3.1), i.e., if $M \geq m_I^{\min} + m_{II}^{\min}$, then we allocate to class II its required number of preambles, i.e., $m_{II} = m_{II}^{\min}$, and allocate the remaining preambles to class I:

$$m_I = M - m_{II}^{\min}. \tag{3.21}$$

Thus, the number m_I of preambles allocated to class I is a least as large as necessary ($m_I^{\min}$). Hence, class I is prioritized compared to class II.

Next, consider the overloaded case (see Section 3.3.3.3.2) when there are not enough preambles to satisfy the demand of both classes, i.e., if $M < m_I^{\min} + m_{II}^{\min}$. In order

to maintain a prescribed level of performance for class II, we introduce a prioritization factor r, $r \in [0,1]$, that regulates the fairness for the number of preambles allocated to class II. Smaller r increases the fairness and reduces the prioritization effects, while larger r is decreasing the fairness. $r = 1$ corresponds to strict prioritization. In particular, we allocate to class II the portion r of the proportional allocation of the M preambles according to the ratio $m_{II}^{\min}/(m_I^{\min} + m_{II}^{\min})$ of the required preambles for classes I and II, i.e., we allocate $rMm_{II}^{\min}/(m_I^{\min} + m_{II}^{\min})$ preambles to class II. On the other hand, if the prioritization factor r is so low that the allocation according to r would gives less preambles to class II than are left after allocating $m_I^{\min}$ preambles to class I, then we allocate the remaining $M - m_I^{\min}$ preambles to class II. Thus, overall, we allocate the number of preambles specified in Step 9. of Algorithm 1 to class. As specified in Step 10. of Algorithm 1, we then allocate the remaining $M - m_{II}^{\min}$ preambles to class I.

3.3.5 Evaluation

3.3.5.1 Simulation Set-up

We implemented the simulation models with an event-based OMNeT++ framework (C++) [Var+01]. We collected and processed the statistics with Python-based open-source SciPy [JOP+01] libraries. We simulated the RAP at the level of detail corresponding to our model. In particular, we simulated one gNB with either infinite-source (UE) assumption (i.e., back-logged request do not reduce the arrival rate) or with a finite large number $N \in \{1000, 5000, 10000, 30000\}$ of UEs. An RA request is considered as collided if two or more UEs select the same preamble in the same time slot. No propagation or interference effects are considered. The 95 % confidence intervals are less than 3 % of the corresponding sample means and are not plotted to avoid visual clutter.

3.3.5.2 LATMAPA: Analysis vs. Simulation

Fig. 3.8 shows the LATMAPA performance as a function of the normalized class I arrival rate $\rho_I = \lambda_I/M$ for a fixed class II arrival rate $\rho_{II} = \lambda_{II}/M = 0.15$. We have set the prioritization factor to the relatively small value $r = 0.02$ so as to initially consider a scenario with pronounced prioritization. We examine the impact of r in detail in Section 3.3.5.4. In Fig. 3.8, we investigate the performance of LATMAPA and compare our analytical model for an infinite UE population with simulations for finite UE populations.

From Fig. 3.8(a), we observe that for increasing class I traffic load ρ_I, LATMAPA sustains a nearly linearly increasing class I throughput almost up to the load point $\hat{\rho} = \lambda_I/M$. Note that at the $\hat{\rho}$ load point, the number $m_I^{\min}$ of preambles required for class I reaches the total number of available preambles M. We observe from Fig. 3.8(c) that the class II throughput starts to drop when the class I load approaches $\hat{\rho} - \rho_{II} = 0.22$. This is because the pronounced prioritization for the considered small $r = 0.02$ adaptively takes preambles from the low-priority class II and assigns the preambles to the high-priority class I as the class I traffic load increases.

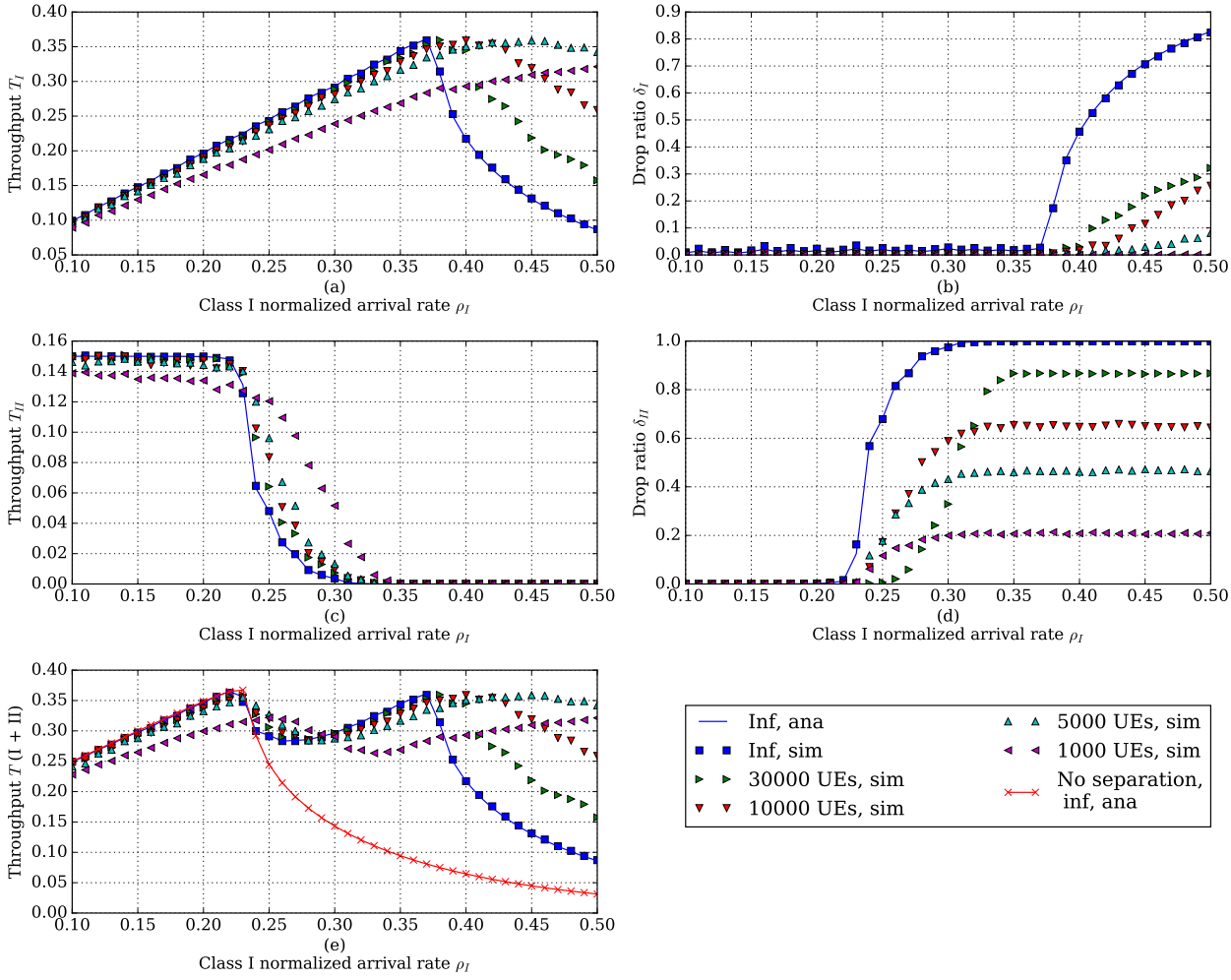

Figure 3.8: LATMAPA throughput T and drop ratio δ for class I (a, b), class II (c, d) and the total throughput (e) as a function of normalized class I arrival rate $\rho_I = \lambda_I/M$; $\lambda_{II}/M = 0.15M$, fixed; System parameters: $M = 54$ preambles, $W = 8$, $r = 0.02$. LATMAPA is used to calculate the number of preambles m_I and m_{II}. Model verification for infinite and finite number of UEs.

Similarly, we observe from Fig. 3.8c that LATMAPA maintains a nearly constant high class II throughput until the total required number of preambles $m_I^{\min} + m_{II}^{\min}$ exceeds the number of available preambles M, i.e., until the RACH becomes overloaded.

We observe a positive side effect of prioritization with LATMAPA in Fig. 3.8e, which shows the total throughput for both classes. In the overloaded region, we observe that prioritizing class I leads to an increase of the total throughput with LATMAPA compared to the total throughput without separation (which is plotted as the "No separation, inf, ana" curve). This throughput increase achieved with LATMAPA prioritization corresponds to the throughput increase achieved with preamble separation in the overloaded region (see Sec. 3.3.3.3.2 and Fig. 3.6(a)). We also observe a slight "dip" (decrease) in the total throughput in the load range between $\hat{\rho} - \rho_{II} = 0.22$ and $\hat{\rho}$. This dip effect

is due to different slopes to the left and right of the maximum throughput region (**A** in Fig. 3.2): Class II throughput decreases faster (slope to the right of **A**) than class I gains throughput (slope to the left of **A**).

Regarding the accuracy of the analysis, we observe from Fig. 3.8 that the simulation for the infinite UE population model essentially coincides with the analysis for the infinite UE population model. We also observe from Fig. 3.8(a), (c), and (e) that the finite UE population throughputs are approximated by the infinite UE population analysis. The discrepancy in throughputs between simulation and analysis increases with decreasing number of UEs. However, the analysis gives a meaningful approximation and lower throughput bound down to 10,000 UEs. We observe from Figs. 3.8(b) and (d) that the drop ratios from the finite-UE simulations deviate significantly from the analytical infinite-UE results. However, the infinite-UE analysis provide an upper bound of the drop ratios.

Importantly, LATMAPA still maintains low drop ratio for the prioritized request class I. LATMAPA inherently excludes any cross impact between the two UE classes, i.e., the QoS levels of the two request classes are isolated from each other. Therefore, quality of service, resulting in low drop ratio and, hence, low delay, can be guaranteed for class I as long as there are enough preambles (i.e., for low r, we need $M \geq m_I$). The QoS level isolation achieved with preamble separation is fundamentally different from prioritization methods that manipulate the random access on a given set of preambles, e.g., methods that manipulate the access barring, backoff window, or number of transmission attempts, because these prioritization methods do not eliminate contention of the different classes for the same set of preambles. Also, the preamble separation approach allows for effective prioritization during long periods of overload and for steady-state operation, where the access barring based approaches fail [Dua+16].

3.3.5.3 LATMAPA: Comparison with Other Allocation Methods

We compare LATMAPA with the two existing load adaptive preamble allocation mechanisms in [ZZF14; Du+16]. With ρ_I and ρ_{II} denoting the normalized loads of high-priority class I and low-priority class II UE requests, respectively, the Zhao2014 allocation mechanism [ZZF14] allocates $m_I = \min\{\lfloor 1.5\rho_I M \rfloor,\ \lfloor\ M w \rho_I/(\rho_I + \rho_{II})\rfloor\}$ preambles to the high-priority class I. The weight parameter w is varied in the range $(0, 10]$. The remaining $m_{II} = M - m_I$ preambles are allocated to the low-priority class II.

The Du2016 allocation mechanism [Du+16] considers the access barring factor b_I, and the number x_I of contending UEs in a given slot for class I. The Du2016 approach calculates an optimal split $\beta^\star = m_I/m_{II}$ as follows:

$$\beta^\star = \begin{cases} \frac{x_I(1-b_I)}{M\log x_I(1-b_I) - x_I(1-b_I)} & \text{if } x_I(1-b_I) \in [3, +\infty) \\ \frac{x_I(1-b_I)}{M\log x_I(1-b_I)/2 - x_I(1-b_I)} & \text{if } x_I(1-b_I) \in (1, 3). \end{cases} \tag{3.22}$$

From the optimal split $\beta^\star$, the Du2016 approach allocates $m_I = M\beta^\star/(1+\beta^\star)$ preambles to the high-priority class I and $m_{II} = M - m_I$ preambles to the low-priority class II. Note

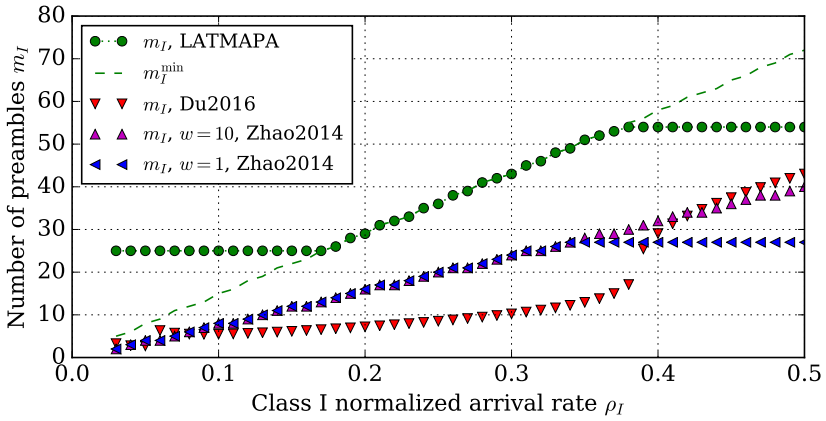

Figure 3.9: Comparison of the number of preambles m_I allocated to the high priority class I by LATMAPA, Zhao2014 [ZZF14], and Du2016 [Du+16] as a function of the normalized class I UE request arrival rate ρ_I. The minimum required number of preambles $m_I^{\min}$ from Eqn. (3.18) is plotted as a reference. Class II load is kept constant at $\rho_{II} = 0.2$.

that the Du2016 approach utilizes information about the exact number of contending UE requests in the upcoming slot. It is not realistic to obtain this number for every slot; however, the expected number of contending UE requests can be obtained as a function of the arrival rate λ_I by numerically solving Eqns. (3.1) and (3.2). For a fair comparison with LATMAPA and Zhao2014, we use this expected number of arrivals x_I for obtaining the optimal split as in Eqn. (3.22), and set the barring factor $b_I = 0$.

We compare the preamble allocation for class I resulting from LATMAPA, Zhao2014 [ZZF14], and Du2016 [Du+16] in Fig. 3.9, with the fixed class II arrival rate $\rho_{II} = 0.2$. We observe that both Zhao2014 and Du2016 do not allocate enough preambles to the high priority class I. For Zhao2014 [ZZF14], we observe that changing the weight parameter w only influences the allocation high loads $\rho_I \geq 0.35$. In contrast to the Zhao2014 and Du2016 allocation methods, LATMAPA allocates the required minimum number m_I^{min} of preambles to the high-priority class I as long as the available number of preambles M and traffic load ρ_I permit; hence, LATMAPA more effectively prioritizes the high-priority class I traffic than the prior Zhao2014 and Du2016 approaches.

3.3.5.4 LATMAPA: Impact of Prioritization Factor r

The prioritization factor $r \in [0, 1]$ controls the minimum level of service provided to class II. It only plays a role if the overall amount of preambles is insufficient to satisfy the traffic load of both classes. That is, if $r = 1$, the available preambles are allocated proportionally to two classes. If $0 < r < 1$, class II only obtains an r portion of the proportional preamble allocation. In the other extreme case, if $r = 0$, class I gets all available the resources.

Fig. 3.10(a) shows the number of allocated preambles as a function of the class I

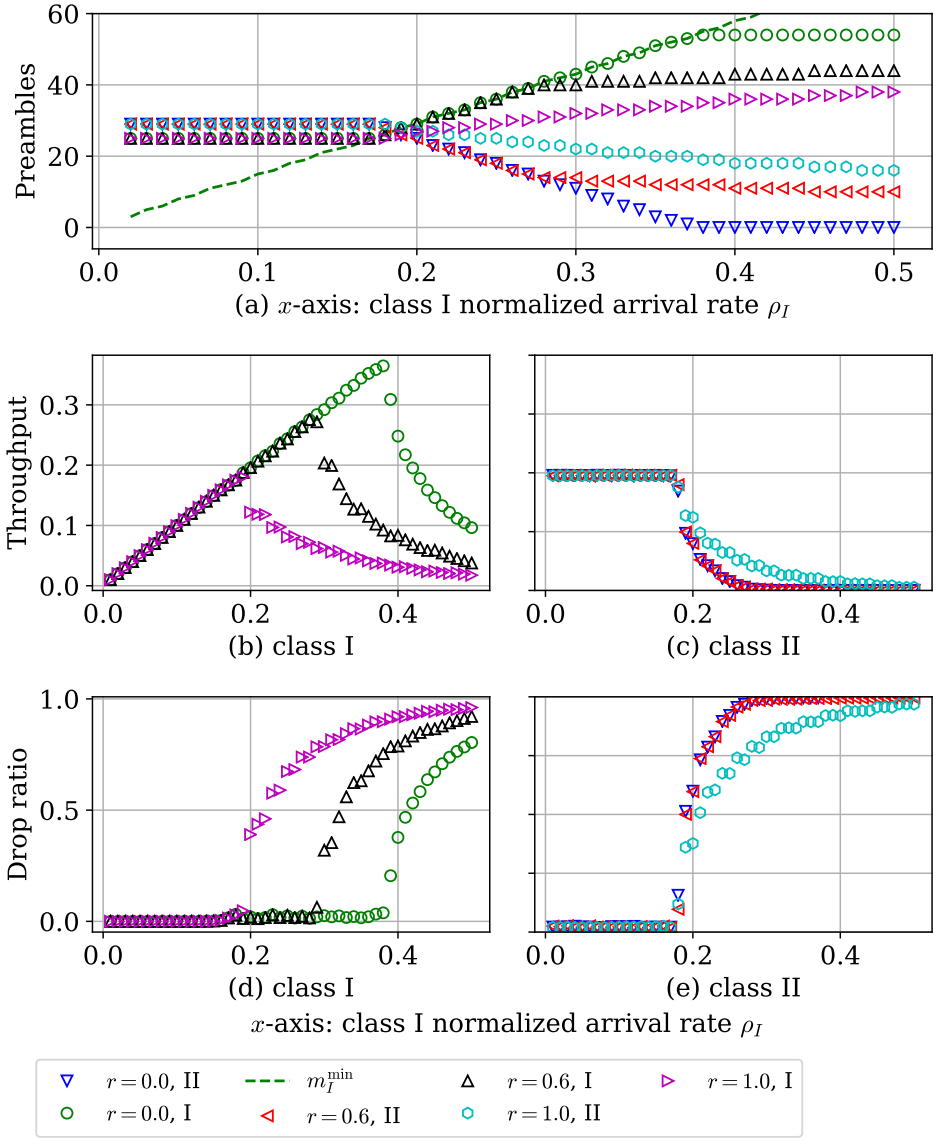

Figure 3.10: LATMAPA performance with different prioritization factors r: (a) number of preambles allocated to classes I and II; (b), (c) throughput of class I and II respectively; (d), (e) drop ratio for class I and II respectively. X-axis is class I normalized arrival rate ρ_I; class II traffic load $\rho_{II} = 0.2$, fixed.

normalized arrival rate (load) ρ_I, with class II arrival rate fixed at $\rho_{II} = 0.2$. We observe that the preamble allocation is static until the arrival rate reaches the point where $m_I^{\min} + m_{II}^{\min} = M$ at $\rho_I = 0.17$: class II gets only necessary number of $m_{II} = m_{II}^{\min}$ preambles, and class I receives the remaining $m_I = M - m_{II}^{\min}$ preambles.

The prioritization factor starts playing a role once the arrival rate of class I increases above $\rho_I = 0.18$, as can be observed from Figs. 3.10(b-e). For $r = 1$ both classes are treated equally and share the available M preambles proportionally to their loads ρ_I and ρ_{II}. For $r = 0.6$, we observe a shift in the preamble allocation towards class I: class I

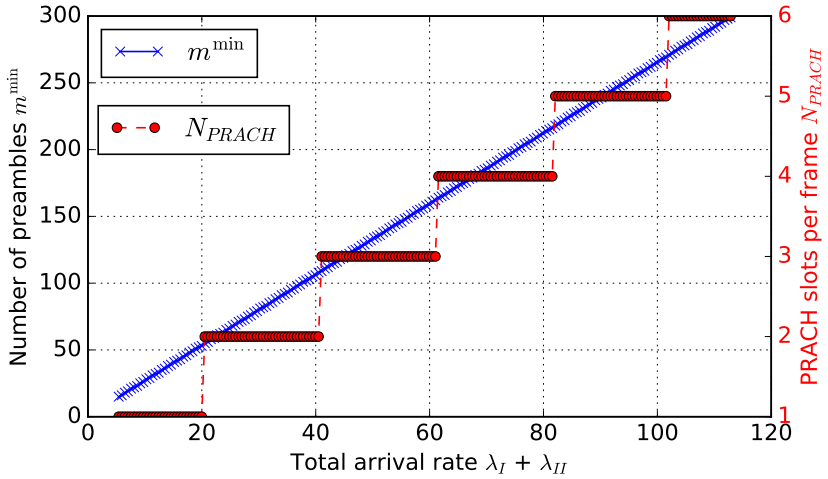

Figure 3.11: Minimum number of required preambles $m^{\min}$ and PRACH slots per frame N_{PRACH} as a function of total arrival rate λ [arrivals per frame].

is prioritized, hence, the gap between m_I and m_{II} is larger than for $r = 1$. For $r = 0.0$, class I is first allocated its required minimum number of preambles $m_I^{\min}$ (up to the available M preambles), and any remaining preambles are allocated to class II. Thus, the setting $r = 0$ corresponds to strict prioritization.

3.3.5.5 Tuning PRACH Configuration Index

In the preceding sections, we have analyzed and evaluated scenarios for a prescribed fixed PRACH configuration index. However, practical scenarios require the tuning of the PRACH configuration index, which corresponds to the number of PRACH slots available in a given frame [YHH11; Yun12]. Our model can be readily extended to tune the PRACH configuration index. The tuning allows to choose the optimal index in order to properly provision the channel. In particular, if $m^{\min} = m_I^{\min} + m_{II}^{\min} \geq M$, then a larger number of PRACH slots per frame N_{PRACH} is needed:

$$N_{\text{PRACH}} = \left\lceil \frac{m_I^{\min} + m_{II}^{\min}}{M} \right\rceil . \tag{3.23}$$

Fig. 3.11 shows the required minimum number of preambles $m^{\min}$ and the required number of PRACH slots per frame N_{PRACH} as a function of the total arrival rate $\lambda_I + \lambda_{II}$. The number of PRACH slots can be used to determine the PRACH configuration index, e.g., index 0 for $N_{\text{PRACH}} = 1$ or index 12 for $N_{\text{PRACH}} = 5$.

Remark 3 (On comparison with RAP manipulation methods)**.** *Generally, random access performance has been studied for two main settings: constant (steady-state) traffic, where the system behavior is studied for long periods of constant UE request load, and bursty traffic, where the system is studied for temporary (sudden) overload periods. The constant traffic studies have mainly focused on evaluating steady-state performance aspects and influencing parameters [Tya+15]. On the other hand, the bursty traffic studies*

have focused on methods for the efficient resolution of large amounts of simultaneous (one-shot) or nearly simultaneous (mostly modeled as beta-distributed with a prescribed activation time) UE request arrivals [Dua+16]. State-of-the-art methods for prioritizing random access through the manipulation of the RAP on a given set of preambles, such as EAB, belong to the category of bursty traffic studies. That is, random access parameter manipulation methods, such as EAB, have been developed for temporary, non-persistent overload conditions. Thus, these random access parameter manipulation methods are not suitable for addressing persistent, constant overload conditions (which are the focus of this present study). For instance, studies [Tya+15; Tya+17] have demonstrated that neither access barring or tuning of the back-off parameters change the steady-state throughput or drop ratio of systems with constant traffic loads. Therefore, LATMAPA, which has been developed for steady-state (constant) traffic, can not be directly compared with prioritization methods that manipulate the RAP on a given set of preambles so as to address non-persistent traffic bursts. In the next chapters, we will return to the question of preamble (resource) allocation for the non-persistent traffic scenarios.

3.4 Random Access with Spatial Aggregation

As we show in Sec. 3.3.5.5, with the increasing load PRACH configuration index could be tuned to increase the number of available preambles per frame, and, hence, keep the performance at acceptable levels. However, according the specification, the amount of PRACH allocations per frame is limited [Cox12]. To support higher loads, methods to increase the number of preambles or reuse them within a cell are needed. A promising approach for it is to reuse the preambles using spatial aggregation: To introduce intermediate aggregators, which collect the requests locally and forward them to the gNB. As we review in Sec. 3.2.3, the aggregation of RACH requests has been previously studied in the literature. However, still missing is the analytical modeling and understanding on how the connection request aggregation process influences RACH performance, in terms of the delay and request drop ratio. This section is hence dedicated to analytical performance assessment of RACH with aggregation.

The remainder of the section is organized as follows. Background and system model are introduced in 3.4.1. Next, Sec. 3.4.2 is presenting the analytical model. Finally, Sec. 3.4.3 shows simulation results, where we benchmark our analytical model with event-based simulations.

3.4.1 Scenario and Protocol: Aggregation of Connection Requests

To extend the capacity of PRACH via spatially reusing the resources, we propose aggregation of requests. We split all M2M UEs into clusters, and every M2M UE, instead of sending RRC Connection Request directly to the gNB, first forwards it to the respective clusterhead (CLH). The link between UE and CLH (see Fig. 3.13) is a standard

RAP with available number of preambles denoted throughout the rest of the chapter by M_C. We assume that M_C is advertised in gNB broadcasts, and that it can be chosen based on the number of clusters and the number of devices in a cluster. It can be at most $M_C \leq M$, where M is the total amount of preambles for contention-based RAP (typically 54 [3GP11]). Throughout the section, we consider the number of preambles reserved for reuse to be relatively low, e.g., $M_C = 0.1M \approx 6$. In general, it has to be calculated according to the ratio of clustered M2M UEs and other backgroud UEs (e.g., for H2H applications). After the connection UE to CLH is established, the CLH aggregates a certain amount of connection requests from UEs within a cluster and then forwards them collectively to the gNB as an *aggregated packet*. Interference on PRACH among the clusters can be kept at minimum by interference-aware clustering and proper power control [Wan+13]. The protocol is depicted in Fig. 3.13.

The protocol can be implemented in a distributed fashion using Device-to-Device (D2D) neighbor discovery [TPM16], where any UE can serve as a temporary prearranged clusterhead (possibly rotating). Alternatively, a predefined M2M gateway can serve at a clusterhead [TST12]. In this work, we deal exclusively with the medium access aspects of clustered connection establishment, hence, we assume that the clusters are formed in advanced and the clusterhead is predefined. Consequently, we neglect possible effects of interference due to non-perfect cluster formation or power control problems in our analysis. We acknowledge that, in general, these effects, along with the possible effects of cluster formation, should be considered for a complete feasibility analysis of the proposed solution.

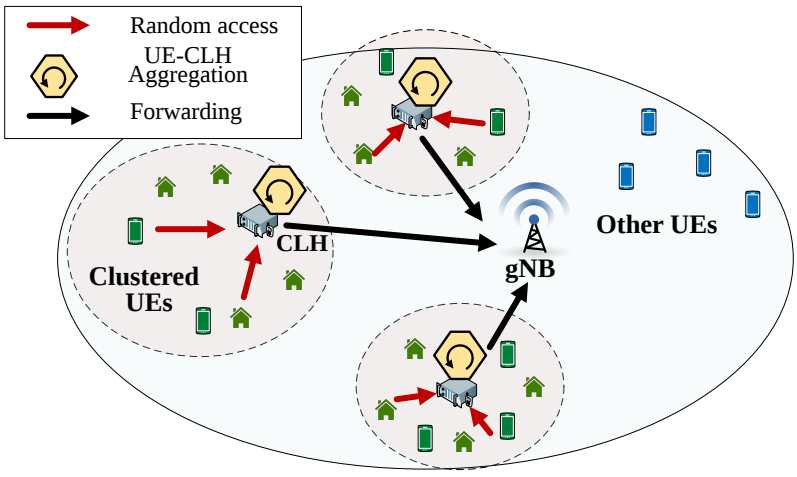

Figure 3.12: Cluster-based connection request aggregation architecture. All M2M UEs are grouped into clusters, and are establishing a connection to clusterheads (CLHs) first. The CLHs are aggregating requests within the cluster, before forwarding them to the gNB.

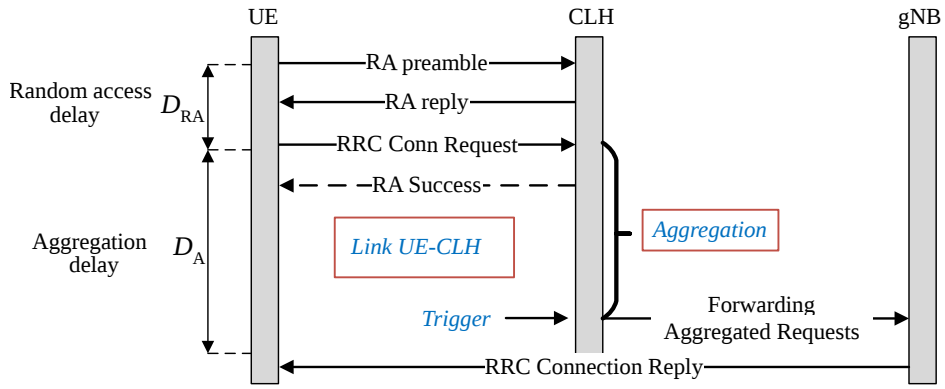

Figure 3.13: Connection establishment protocol: UE connects to CLH via RAP, then the request waits for aggregation, and then it is forwarded to the gNB.

3.4.2 Performance Analysis

In this section, the two-part analysis is presented: (3.4.2.1) the aggregation process standalone and its implications and (3.4.2.2) the joint model of RAP and aggregation within a single cluster.

3.4.2.1 Aggregation on Clusterhead

We start with analyzing the delay due to aggregation process. In fact, for analyzing this process, we need to consider what is the *aggregation trigger*, namely, what is the condition under which the clusterhead decides to forward aggregated requests. Commonly assumed triggers are either with deterministic [Meh+15] or Markovian timers [TST12], meaning that the packets are sent after the expiration of an aggregation timer. However, these triggers do not consider an upper limit on the size of the aggregated packet. That is, if an RRC Connection Request containing UE Identity and Connection Cause is of size 80 bits [3GP16], and the maximum size of an aggregated packet is 200 bytes, at most $N_A = 20$ packets can be sent at once. We denote N_A as an *aggregation factor* and use it as a more realistic triggering condition in the following analysis.

The arrival process of the requests for aggregation is the output process of successful UE-CLH connection establishment. Hence, arrival rate of the requests for aggregation λ_A is calculated from the performance parameters of RAP as $\lambda_A = TM_C$, where T is the normalized throughput of RAP (ratio of successful request divided by the total number of RAO), and M_C is the number of available preambles per slot. We approximate the output process of the RAP stage with the Poisson distribution with the mean value λ_A.

Clusterhead to gNB connection. It is straightforward to see that, since individual requests are arriving according to Poisson distribution, aggregated packets are ready to be sent at every N_A occurrence of a Poisson events. Hence, by definition, the inter-arrival times (IAT) of aggregated packets d_{ia} are Erlang-N_A distributed with parameters (N_A,λ_A):

$$f_{ia}(d_{\text{ia}}; N_A, \lambda_A) = \frac{(\lambda_A)^{N_A} d_{\text{ia}}^{N_A-1} e^{-\lambda_A d_{\text{ia}}}}{(N_A-1)!}$$
$$\text{for } d_{\text{ia}}, \lambda_A \geq 0, \quad \text{with } \mathbb{E}[d_{\text{ia}}] = \frac{N_A}{TM_C} t_{\text{slot}}, \tag{3.24}$$

where t_{slot} is the length of a PRACH slot, typically $t_{\text{slot}} = 10$ ms.

Mean inter-arrival time of the aggregated requests is depicted in Fig. 3.14(a) as a function of the number of UEs in a cluster with a fixed arrival rate per UE. As expected, we observe that aggregated arrivals occur significantly more often than the arrivals within any individual UE. This provides an insight into the connection establishment on the link CLH-gNB: since IAT is significantly less than the typical values of 10s for the UE Inactivity Timer in current LTE systems [3GP12; GXK16], it is unlikely that the clusterhead undergos a RAP on this stage (see CDF in Fig. 3.14(b)). Moreover, deploying a combination with time trigger can fully eliminate the need for RA. Following these observations, we assume that CLHs are always staying in RRC-CONNECTED mode and thus the *aggregated requests do not experience RA delay* on the link CLH-gNB. Hence, there is *no contention between the clusterheads*. In this case, delay for forwarding the requests to gNB is fully dependent on the load of the scheduler and on the scheduler

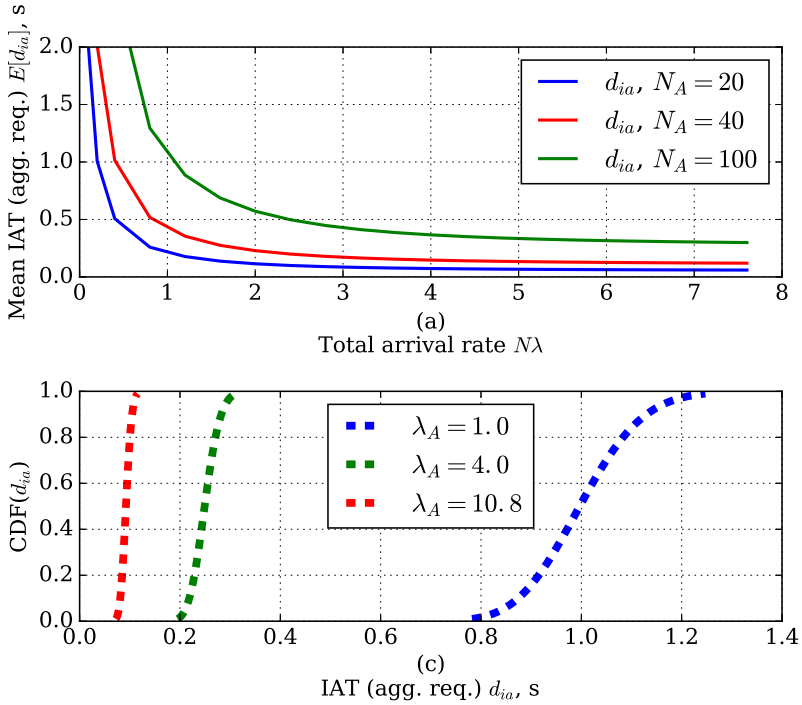

Figure 3.14: Delay analysis of the aggregation stage: (a) mean inter-arrival time of aggregated requests d_{ia} vs. total arrival rate, (c) cumulative distribution function of d_{ia} aggregated requests vs. output rate of RAP.

choice, and its analysis is intractable for a general case. Since we assume that the system has enough uplink resources, we can expect the serving times to be less than a frame size of 10ms. Hence, for further analysis we assume it negligible.

Waiting time of a single connection request. Note that d_{ia} is different from the waiting times of an arbitrary connection request D_A, as it can arrive at any of $i \in [1, N_A]$ position in the aggregation buffer. If a packet is arriving at a position i, it has to wait for $N_A - i$ other packets, hence, its waiting time is distributed according to Erlang $(N_A - i, \lambda_A)$ process. The distribution of the waiting time D_A can thus be obtained:

$$\begin{aligned} f_{D_A}(x) &= \tfrac{1}{N_A}\left(f_{ia}(x; N_A - 1, \lambda_A) + \cdots + f_{ia}(x; 1, \lambda_A)\right) = \\ &= \tfrac{\lambda_A e^{-\lambda_A x}}{N_A} \textstyle\sum_{n=0}^{N_A-2} \tfrac{(\lambda_A x)^n}{n!}. \end{aligned} \tag{3.25}$$

Note that every N_A^{th} packet always has zero waiting time. Since we are interested in the average values of the delay, we approximated it as an average over expectations given by Eqn. (3.24) as:

$$\mathbb{E}[D_A] \approx \frac{t_{\text{slot}}}{N_A} \sum_{n=1}^{N_A-1} \frac{n}{TM_C} = \frac{t_{\text{slot}}(N_A - 1)}{2TM_C}. \tag{3.26}$$

3.4.2.2 Markov Chain Model of RAP and Aggregation

Several approaches to RACH modeling are present in the literature. They can be classified into two main groups: one-shot arrivals, modeled with beta distribution, and steady-state modeling, where the classical Poisson arrival process is assumed. We consider a case of constantly high load on PRACH, hence, we resort to Poisson arrival process of individual UEs. Moreover, since we cannot assume sufficiently large amount of UEs in a cluster, infinite source models, such as devised by Tyagi *et al.* [Tya+15] cannot provide sufficient accuracy.

As discussed in the previous subsection, aggregation process is dependent on the output of the random access procedure. However, the longer the request stays in the aggregation, the less new requests it generates. Hence, there is an inter-dependency between two processes, and separate analysis of the procedures is imprecise. Therefore, our Markov chain model includes aggregation process as one of its states.

The joint Markov chain model of individual UE's states is based on the Distributed Coordination Function (DCF) modeling first introduced in [Bia00]. Its illustration is given in Fig. 3.15. Input parameters of the model are: expected arrival rate from an individual UE λ, number of UEs in a cluster N, number of preambles available for clustered access M_C, maximum back-off $B_{\max}$.

The model considers every UE to be in one of the states: CONN (connected), AGG (connection request is waiting for aggregation on CLH), DROP (dropped after W transmission attempts), (i, k) (i^{th} re-transmission attempts with k^{th} back-off slots remaining), and OFF (no pending requests). We denote the probability of a collision as p_c and the

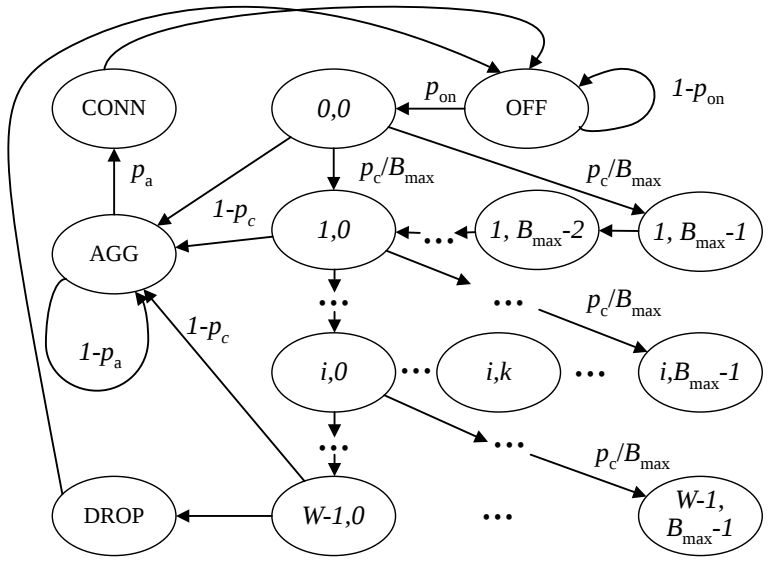

Figure 3.15: Markov-chain model of an UE considering aggregation process.

probability of connecting while being in the aggregation buffer as p_a. We further denote the probability of a transition between any pair of states k and m as $\mathbb{P}[\text{state } k|\text{state } m]$ and the steady-state probability of any state k as $\tilde{p}_{\text{state } k}$.

The probability of generating new request while being in the OFF state is given by:

$$\mathbb{P}[0,0|\text{off}] \;=\; p_{\text{on}} = 1 - e^{-\lambda}. \tag{3.27}$$

Transition probabilities are computed as:

$$\begin{aligned}
\mathbb{P}[i,k|i-1,0] &= \frac{p_c}{B_{\max}},\\
\mathbb{P}[\text{conn}|i,0] &= 1-p_c,\\
\mathbb{P}[\text{drop}|W-1,0] &= p_c,\\
\mathbb{P}[\text{conn}|\text{agg}] &= p_a,\\
\mathbb{P}[\text{off}|\text{drop}] &= \mathbb{P}[\text{off}|\text{drop}] = 1.
\end{aligned} \tag{3.28}$$

We proceed by computing steady-state probabilities using the global balance equations:

$$\tilde{p}_{0,0} \;=\; p_{\text{on}}\tilde{p}_{\text{off}}, \tag{3.29}$$

$$\tilde{p}_{i,k} \;=\; \tilde{p}_{i-1,k}\frac{p_c}{B_{\max}} + \tilde{p}_{i,k-1} = \frac{B_{\max}-k}{B_{\max}}p_c^i\tilde{p}_{0,0}, \tag{3.30}$$

$$\tilde{p}_{i,0} \;=\; p_c\tilde{p}_{i-1,0} = ... = p_c^i\tilde{p}_{0,0}. \tag{3.31}$$

$$\tag{3.32}$$

Following that, we derive the remaining steady-state probabilities $\tilde{p}_{\text{conn}}$, $\tilde{p}_{\text{drop}}$ as a function of $\tilde{p}_{\text{off}}$, $\tilde{p}_{\text{agg}}$:

$$\tilde{p}_{\text{conn}} = \sum_{i=0}^{W-1}(1-p_c)\tilde{p}_{i,0} = \sum_{i=0}^{W-1}(1-p_c)p_c^i\tilde{p}_{0,0} = p_{\text{on}}\tilde{p}_{\text{off}}(1-p_c^W), \tag{3.33}$$

$$\tilde{p}_{\text{drop}} = p_c\tilde{p}_{W-1,0} = p_c^W p_{\text{on}}\tilde{p}_{\text{off}}, \tag{3.34}$$

$$\tilde{p}_{\text{agg}} = \frac{p_{\text{on}}}{p_a}\tilde{p}_{\text{off}}(1-p_c^W). \tag{3.35}$$

Now, the $\tilde{p}_{\text{off}}$ can be calculated by imposing normalization condition:

$$1 = \tilde{p}_{\text{agg}} + \tilde{p}_{\text{off}} + \tilde{p}_{\text{conn}} + \tilde{p}_{\text{drop}} + \sum_{i=0}^{W-1}\sum_{k=0}^{B_{\max}-1}\tilde{p}_{i,k}. \tag{3.36}$$

From it, we derive the equation for $\tilde{p}_{\text{off}}$

$$\tilde{p}_{\text{off}} = \frac{2p_a(1-p_c)}{2(1-p_c)(p_a(1+2p_{\text{on}})+p_{\text{on}}(1-p_c^W))+p_a p_{\text{on}} p_c(1+B_{\max})(1-p_c^{W-1})}. \tag{3.37}$$

Let us denote the effective arrival rate, including re-transmissions and activity time, from a single UE as $\tilde{\lambda}$. Then, expected number of contending UEs in a given slot is given by $\tilde{\lambda}N$, and the collision probability is given by:

$$p_c = 1 - \left(1 - \frac{1}{M_C}\right)^{\tilde{\lambda}N-1}. \tag{3.38}$$

The effective arrival rate is equal to the probability of being in any of the $(i,0)$ states:

$$\tilde{\lambda} = \sum_{i=0}^{W-1}\tilde{p}_{i,0} = \frac{1-p_c^W}{1-p_c}\tilde{p}_{0,0} = \frac{1-p_c^W}{1-p_c}p_{\text{on}}\tilde{p}_{\text{off}}. \tag{3.39}$$

Substituting p_c and $\tilde{p}_{\text{off}}$ by Eqn. (3.37) and (3.38), we obtain $\tilde{\lambda}$ as a function of p_a. Now we compute an approximation for p_a using Eqn. (3.26), we can estimate probability of a transition from the aggregation state as:

$$\mathbb{E}[D_A] = \sum_{n=0}^{\infty} n\,\mathbb{P}[n] = \frac{1-p_a}{p_a} \overset{!}{=} \frac{N_A-1}{2TM_C}, \tag{3.40}$$

$$p_a = \frac{2TM_C}{N_A-1+2TM_C} = \frac{2\tilde{\lambda}N(1-p_c)}{N_A-1+2\tilde{\lambda}N(1-p_c)}. \tag{3.41}$$

Finally, using Eqns. (3.39) and (3.41), and simplifying the result we obtain:

$$\tilde{\lambda} = \left(\frac{1}{N}\right)\frac{p_{\text{on}}(1-p_c^W)(2N+1-N_A-2\tilde{\lambda}N(1-p_c))}{2(1-p_c)(1+2p_{\text{on}})+p_{\text{on}}p_c(1+B_{\max})(1-p_c^{W-1})}. \tag{3.42}$$

Eqn. (3.42) can be solved numerically for $\tilde{\lambda}$ with iterative methods. Using $\tilde{\lambda}$, p_c and the results of the previous subsection, delay D (in ms), outage probability δ, and throughput T (per PRACH slot) of the aggregated procedure are computed as:

$$\begin{aligned} D &= D_A + D_{\mathrm{RA}} = \\ &= t_{\mathrm{slot}} \frac{N_A + 1}{2TM_C} + t_{\mathrm{slot}} \left(\frac{p_c(B_{\max} - 1)}{2(1 - p_c)} \right) \left(\frac{1 + (W-1)p_c^W - Wp_c^{W-1}}{1 - p_c^W} \right), \\ & \delta = \tilde{p}_{\mathrm{drop}} / (\tilde{p}_{\mathrm{drop}} + \tilde{p}_{\mathrm{conn}}) = p_c^W, \\ & T = \tilde{\lambda} N (1 - p_c) / M_C. \end{aligned}$$

In the next section, we study the predicted performance of the joint procedure, and verify the model simulatively.

3.4.3 Evaluation

Evaluations were performed using event-based simulator. Duration of performed simulations is fixed to 3000 PRACH slots. We simulate one cluster with $N = 200$ UEs, $M_C = 6$ available preambles, and PRACH slots available every frame, i.e., $t_{\mathrm{slot}} = 10$ms. The reception is assumed successful if no collision has occurred, hence, no propagation effects are included in the simulation model.

3.4.3.1 Random Access without Aggregation

First, we benchmark our model's performance without the aggregation state with a similar model by Madueno *et al.* [Mad+16] (denoted MChain) and with infinite-source model provided by Tyagi *et al.* [Tya+15] (denoted InfSrc). Fig. 3.16 shows throughput versus total arrival rate, and verifies our model against simulation. It is observed that our model provides a more accurate estimate of the throughput for a both 200 and 1000 UEs. Both MChain [Mad+16] and InfSrc [Tya+15] models provide a good approximation for large number of UEs, however, significantly underestimate the throughput for $N = 200$. Although the model from [Mad+16] resembles ours without aggregation, it has fundamentally different results since it is insensitive to the number of users N, and, hence, similar to infinite-source model.

3.4.3.2 Performance with Aggregation

Performance results are shown in Fig. 3.17 in terms of throughput T (a), delay D (b) and drop probability δ (c). We compare results of the analytical model and simulation for different values of N_A. From Fig. 3.17(a), it is observed that increasing the aggregation factor leads to the shift of the peak throughput point to the right. This means that, with higher aggregation factor, higher throughput is achieved in the overload region (right from the maximum throughput point), and lower throughput is achieved in underloaded

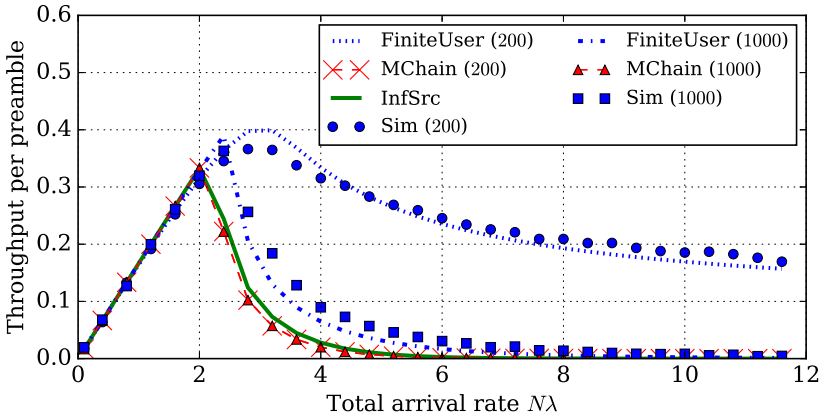

Figure 3.16: Throughput per preamble vs. total arrival rate. Comparison for three models: FiniteUser (our model), MChain [Mad+16], InfSrc [Tya+15]. Parameters: $N = [200, 1000]$, $M_C = 6$, $B_{\max} = 20$ slots, $W = 8$.

region (left from the maximum throughput point). Hence, tweaking N_A could be helpful in case the system is overloaded. To illustrate this effect further, the dependency of total arrival rate λN corresponding to maximum throughput $T_{\max}$ from N_A is plotted in Fig. 3.18. From a practical point of view, this effects also means that with the higher aggregation factor, higher intra-cluster arrival rates can be supported without overloading the system.

In Fig. 3.17(c) we observe that the drop ratio δ is decreasing with increasing aggregation factor. Since UEs spend longer time in AGG state, less new requests are generated, and, hence, collision probability p_c is reduced. This, in turn, reduces δ. Decreasing drop rate is traded for increasing delay, as observed from Fig. 3.17(b). The increase is significant if the arrival rates are low. It is intuitively explained by the fact that low arrival rate results into longer waiting time until N_A requests are aggregated.

3.4.3.3 Aggregation Delay vs. Random Access Delay

To study the delay effects further, we plot both D_A and D_{RA} in Fig. 3.19. First, we verify the validity of D_A approximation against simulation results in Fig. 3.19(a). From Fig. 3.19(b), we observe that, for $N_A = 20$ in the low arrival rate region full delay D is dominated by aggregation process, but in the high arrival rate region ($N\lambda \geq 2$) random access procedure delay D_{RA} is significantly larger. Similar result is obtained for larger $N_A = 100$, however, D_A remains at least as high as D_{RA}. Random Access delay D_{RA} is also smaller for larger aggregation factor.

Very high delay due to aggregation for low-arrival-rate region suggest that a load-adaptive triggering policy with variable $N_A \leq N_A^{\max}$ can be used. Alternatively, a hybrid combination with time-based triggering policy is an option to limit maximum waiting time due to aggregation. Note that, although we do not study time-based or hybrid aggregation policies, they can be easily accommodated into the presented model.

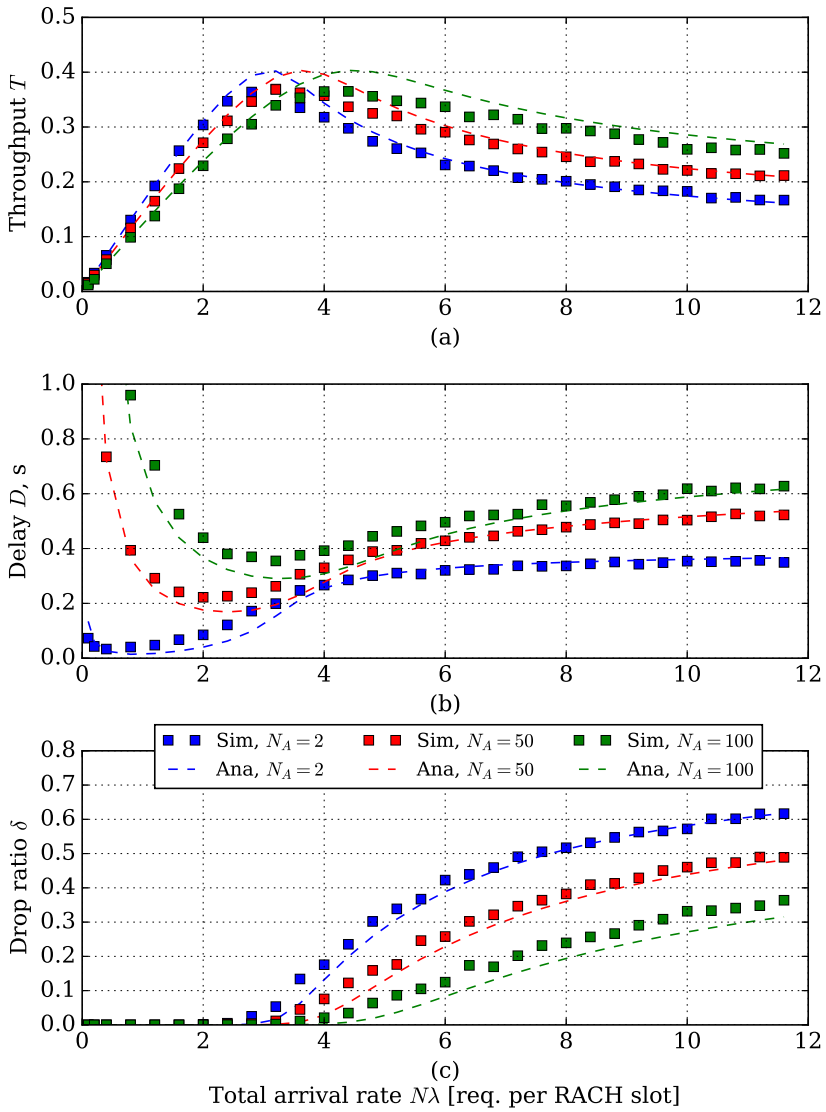

Figure 3.17: Performance of RAP with aggregation vs. total load per PRACH slot $N\tilde{\lambda}$: (a) throughput T, (b) delay D, (c) drop ratio δ. Performance is shown aggregation factors $N_A = [2, 50, 100]$ packets. Parameters as in Fig. 3.16.

3.5 Summary

For the setting of steady-state UE request arrival load in future 5G wireless systems that have evolved from LTE, we have examined the approaches to improve the performance via the preamble manipulations. In the first part of the chapter, we have answered the questions of how to allocate the preambles to maximize RACH performance (Theorem 1), and studied how separation of the preambles into two classes, a high-priority class I and a low-priority class II, can be used to prioritization. For underloaded traffic conditions we have determined a safe prioritization region Δm, within which delay decreases for class I are not accompanied by noticeable performance degradations for class II. For overloaded

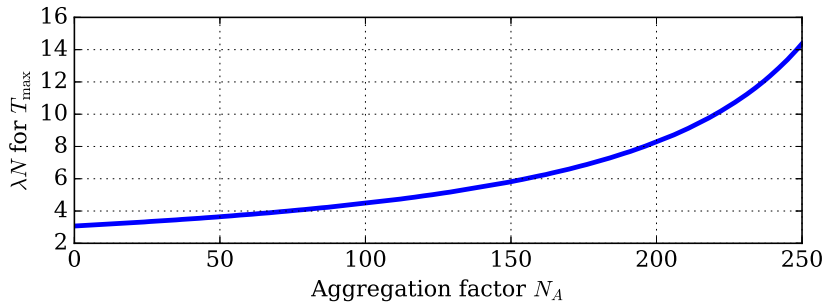

Figure 3.18: Arrival rate per UE λ, achieving maximum throughput, vs. aggregation factor N_A. Same parameters as for Fig. 3.16.

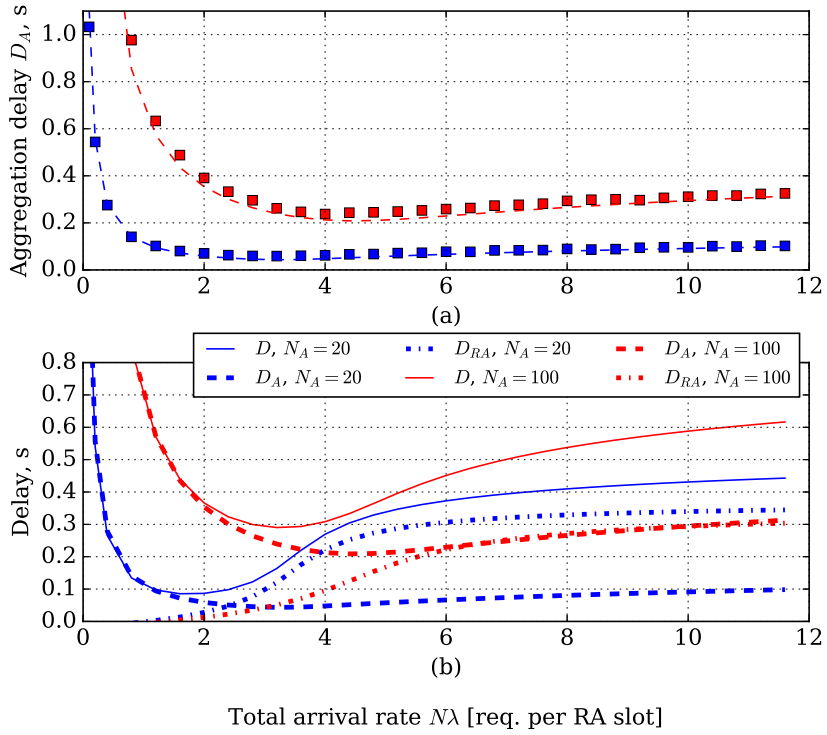

Figure 3.19: Delay illustrations: (a) confirming aggregation delay with simulation, (b) comparison of D, D_A, and D_{RA}. Same parameters as for Fig. 3.16.

traffic conditions we have demonstrated that preamble separation can increase the total (aggregate) throughput. Prioritization of class I in the overloaded region comes at the cost of increasing the ratio of dropped requests for class II, but can significantly decrease the delay and throughput for class II.

We have further investigated two possible preamble allocation methods for prioritization. The first approach matches the average access delay of the prioritized class, but turned out to be not practical. The second method, Load-Adaptive Throughput-MAximizing Preamble Allocation (LATMAPA) strives to maximize the system throughput. We demonstrated that LATMAPA gives favorable performance up to the exhaustion of available preambles by the prioritized class I.

Future research can investigate the combination of our LATMAPA preamble separation approach for steady-state (constant) overload traffic conditions with methods that manipulate the random access on a given set of preambles, such as Extended Access Barring, for mitigating temporary arrival bursts. There is also a need to design a practical protocol for informing UEs about the available $\mathbf{M_I}$ and $\mathbf{M_{II}}$ preamble sets, which could be achieved by adding the preamble set information to the broadcasted system information blocks.

In the second part of the chapter, we have studied the case where the total number of preambles is insufficient to serve the load in the cell, and analyzed a cluster-based aggregation scheme for connection establishment between machine type UEs and a gNB, as a possible solution to increase the number of available preambles. Assuming well-performed clustering, and, hence, no interference among UEs, such scheme allows spacial reuse of random access resources, where each clusterhead aggregates the request from underlying UEs. We use Markov chains to model the RAP within the cluster together with the aggregation process to account for finite number of UEs in the cell. The medium access model is confirmed to be accurate via event-based simulations. Its analysis shows that the aggregation process can decrease drop rate and increase throughput of RAP, while increasing delay. However, cluster-based connection establishment architecture needs further investigations. That is, our modeling does not include possible effects of inter-cluster interference. Optimizing cluster formation and designing clustering protocols for decreasing inter-cluster interference presents a challenge for future work.

Chapter 4

Efficient Resource-Aware Burst Resolution in M2M Random Access

In the previous chapter, we have addressed the topic of steady-state, "long-term" performance of Random Access Procedure (RAP), a key performance metric for large networks with a massive number of User Equipments (UEs) with independent packet arrivals. In this chapter, we turn our attention to another use case of Machine-to-Machine (M2M) applications, the correlated arrivals from a large group of devices. We refer to such events as **burst arrivals**. They arise from simultaneous triggering of a multitude of UEs, initiated either by the network (i.e., group paging [HHN13]) or by an external event. Examples are: triggering of alarm sensors detecting the same disturbance; a power blackout causing the devices to re-connect to the Next Generation Node B (gNB). Such events might be infrequent, but they lead to long connectivity outages and failures of systems relying on it, because random access protocols are inherently inefficient at handling correlated load [CS88].

Analysis of burst arrivals requires to consider **transient performance** of the protocols and respective performance metrics. As a main metric, state of the art assumes **burst resolution time**, the time it takes to connect all or a specific ratio of the UEs involved in the event. Optimizing for the burst resolution time is a combinatorial problem with the complexity quickly exploding with the burst size [WBC15]. Instead, our approach is to break the problem down into **contention rounds** and adjust contention parameters, namely access probability and the number of Physical Random Access CHannel (PRACH) preambles, for every round. Such approach simplifies the formulation yet generalizes well for the full burst resolution and yields significant gains in burst resolution time, as we will demonstrate in the course of the chapter. However, the burst resolution also involves a number of trade-offs: Setting higher access probability and allocating more preambles leads to higher resource consumption. It is a largely neglected yet important for the network dimensioning aspect: The more resources does the connection establishment procedure consume, the less are remaining for the data transmission. We evaluate this trade-off in the first part of the chapter and use its insights in the second part to develop novel overload control algorithm based on the standardized access control parameters: Access Class Barring (ACB) and preamble allocation. In the final part of the chapter, we exploit the fact that burst arrivals are often **correlated in space**, not just in time, and thus we propose to apply binary countdown contention resolution,

as an advanced listen-before-talk technique, to further boost the efficiency of the burst resolution.

4.1 Contributions and Structure of the Chapter

The contributions of the present chapter are split into three main sections. First, in Sec. 4.3, we define the concept of resource consumption of the RAP and derive its relation to the contention parameters. We demonstrate that the resource consumption has two main components: deterministic (preambles) and stochastic (Physical Uplink Shared CHannel (PUSCH) resources). We then present two different approaches to incorporate resource consumption into the RAP optimization: based on the **resource efficiency** and based on **Pareto optimality**.

Second, in Sec. 4.4, we use the insights of the Pareto analysis to devise a **Pareto Optimal Channel allocation – Access barring (POCA)** algorithm. The algorithm finds a solution to the multi-objective optimization problem belonging to the Pareto set. Combined with the state of the art backlog estimation technique, this solution is then used to dynamically optimize the contention parameters for every round. As we demonstrate by the means of numerical simulations, POCA reduces the average burst resolution time compared to the baseline algorithms for the case of resource-constrained operation of RAP.

Third, in Sec. 4.5, we exploit the fact that the burst arrivals are typically spatially correlated and thus the UEs are close to each other and can potentially overhear other transmissions. Therefore, we propose an efficient Listen Before Talk (LBT) scheme, **Binary Countdown Contention Resolution (BCCR)**, to aid the conventional RAP. We devise a **modified RAP**, where a preamble contention is followed by a contention resolution, which reduces collision probability of MSG3 and lowers the burst resolution time. We analyze the performance and efficiency of the novel RAP and, using the concept of Pareto optimal RAP introduced in Sec. 4.3, we develop a Dynamic Binary Countdown - Access barring (DBCA) algorithm for even faster burst resolution.

The content of this chapter is based on our published works [VK17a; VRK17; VRK19b; VRK19a].

4.2 Related Work

In this section, we review the state of the art closely related to the work in this chapter. For the general state of the art review on M2M random access, we refer the reader to Chapter 2. We group the related work into two categories: resource consumption analysis of RAP (corresponding to contributions in Secs. 4.3 and 4.4) and binary countdown contention resolution (corresponding to contributions in Sec. 4.5).

4.2.1 Resource Consumption and Random Access Procedure

There exist two different ways to consider resources in the RAP procedure in the state-of-the-art: by considering resource efficiency or by viewing resources as a constraint. The resource efficiency is often defined as the normalized throughput, i.e., throughput per preamble [Tya+15; LKY11], see also Chapter 3. Typically, allocation of preambles is considered static in the system [KVGAZ16; CLL11], and throughput is optimized given this static constraint. An extended approach is taken in [Dua+16], where dynamic ACB is combined with dynamic allocation of preambles. The approach however optimized access probability and preamble allocation separately, unlike in our work where they are optimized jointly. We consider the approach by [Dua+16] as a benchmark to our algorithms in this chapter.

A trade-off between preambles spent of contention-free and contention-based access is treated in [KKA13]. A hard constraint on downlink resources for MSG2 is considered in the analytical models developed in [WBC15; Con+16a]. The constraint is limiting the number of MSG2 replies, thus the number of scheduled MSG3 transmissions, reducing in turn the expected throughput of RAP. These works are complementary to our and their insights can be additionally accommodate into our modeling and optimization approach.

In contrast to the state-of-the-art, we offer a more elaborate view on the resource consumption of RAP. We notice that, since RAP is a four-way handshake, there are two components to the resource consumption: deterministic (preambles) and stochastic (PUSCH resources). Furthermore, since the amount of allocated PUSCH resources directly depends on the contention parameters, a collision is less favorable than an idle preamble. It is contrasted with conventional slotted ALOHA protocols, where idle and collided preambles are equally harmful to the system. This difference therefore must be considered in the analysis and optimization.

4.2.2 Binary Countdown for Contention Resolution

The majority of the RAP improvements proposed in the state oft the art are focusing on the *preamble contention* step. We observe however that the actual collision, although being a direct consequence of the preamble collision, is occurring at the MSG3 transmission. Therefore, our approach in Sec. 4.5 aims at resolving the MSG3 collision while allowing the preamble collision. For that, we invoke the BCCR protocol prior to MSG3 contention [VRK17]. The addition of BCCR makes our approach largely orthogonal to the state of the art, as the preamble contention could thus be optimized independently of the MSG3 contention.

Binary Countdown Contention Resolution belongs to the group of access reservation protocols and dates back to the early works on the random access [Tan02]. The core idea is to resolve a contention using a series of short LBT messages prior to the actual transmission. The BCCR protocol is known primarily from the Controller Area Network (CAN) bus systems, but it has been also employed in powerline communication [Geh+14], and studied academically for ad-hoc networks [YYH03; HH10]. Recently,

BCCR has been independently revisited as a possible option for access reservation in the next generation Wi–Fi-like networks [Bai+17; SRCN11].

4.3 Resource Consumption of RAP

In this section, we define the system model used in the remainder of the chapter (4.3.1), and formulate our first contribution by defining resource consumption and developing the optimization framework for resource-aware RAP (4.3.2).

4.3.1 System Model and Preliminaries

We consider a burst arrival scenario as proposed in [3GP11], where N UEs in a cell with one gNB are semi-synchronously activated. At time $t < 0$, all UEs are disconnected from the gNB. During the interval $0 \leq t < T_a$, every UE commences the connection procedure at a random time t with probability distribution function $g_a(t)$. The probability distribution is representing an arrival process with three main possibilities: beta-distributed, uniformly random, and simultaneous "spike" arrivals with $T_a = 0$.

We denote the periodicity of PRACH in the resource grid as a *PRACH slot*, or *slot*[1]. It has been shown that in the current networks the periodicity of a collision feedback and/or system information broadcast might exceed PRACH slot duration [LM+17; WBC15]. To generalize the analysis accounting for different possible Random Access CHannel (RACH) implementations, we define a *contention round* and consider per-contention-round performance metrics (see also Sec. 2.3 and Fig. 2.5 for illustrations). Hence, we denote the minimum period within which the contention parameters can be adjusted and the collision feedback can be received as a contention round. A single contention round could comprise one or multiple PRACH slots.

Prior to any contention round i, every activated UE undergoes an ACB check: with the access probability p_i it proceeds to contend, and with barring probability $1 - p_i$ it skips the upcoming round. In other words, ACB represents a geometric random back-off. It is a possible back-off option, and it can also serve as an approximation for other back-off schemes (exponential, uniform, etc.) or combinations thereof. If the ACB check is passed, UE chooses (uniformly random) a j^{th} preamble, with $j \in \{1, ..., M_i\}$, where M_i is the total number of preambles available in a single contention round. Each preamble can have one of three possible outcomes: **idle** if no device occupies the preamble; **successful** if one and only one device chooses the preamble; and **collided** otherwise.

For every available preamble, there is one Random Access Opportunity (RAO) associated to it. The number of successful UEs in a contention round is limited to the number of RAOs, and it can be at most one success per RAO since we assume no interference cancellation capabilities. Denoting the number of UEs choosing a given preamble j as

[1]The definitions of a slot, contention round, and Random Access Opportunity (RAO) correspond to definitions 1, 3, and 2 in Chapter 2.

Table 4.1: Summary of main model notations in Chapter 4.

M_i	Number of preambles (channels) available per contention round i
$M_I^{(i)}/M_O^{(i)}/M_C^{(i)}/M_S^{(i)}$	Number of idle / occupied / collided / successful preambles (channel) after a contention round i
p_i	Access probability
N	Total number of UEs / Burst size
n_i	Number of contending (backlogged + newly arrived) UEs prior to contention round i
s_i / S	Number of successful UEs / Expected number of successful UEs in a contention round (Throughput)
r_i / R	Consumed resources / Expected consumed resources during a contention round i
r_I / r_O	Uplink resources (PRACH+PUSCH) consumed per idle/occupied preamble (channel)
$\bar{r}$	Resource constraint

$m_{i,j}$ and the outcome of a RAO as $x_{i,j}$, we define a collision channel model as:

$$x_{i,j} \triangleq \begin{cases} 1 & m_{i,j} = 1 \\ 0 & \text{otherwise.} \end{cases} \tag{4.1}$$

As the focus of the chapter is on the performance of contention resolution mechanisms, we make an additional assumption that the downlink channel resources are sufficient and do not pose a performance bottleneck. An extension to account for it would be straightforward.

Now consider a single contention round i. In the beginning of it, n_i backlogged UEs, accounting for both previously unsuccessful and newly activated UEs, are competing for M_i preambles. Instantaneous performance of RAP in a contention round i is characterized with the following two performance metrics: number of successful UEs s_i, and the **resource consumption** r_i. We define the expectation of s_i as **throughput**, which is a function of n_i, p_i, M_i:

$$S(p_i, M_i|n_i) \triangleq \mathbb{E}[s_i]. \tag{4.2}$$

The respective single contention round optimization problem is:

$$\underset{p_i, M_i}{\text{maximize}} \quad S(p_i, M_i|n_i), \tag{4.3}$$

where we use the notation $S(\cdot|\cdot)$ to emphasize that p_i and M_i are optimization variables and n_i is the input condition. Typically, in the state of the art, only p_i is considered as

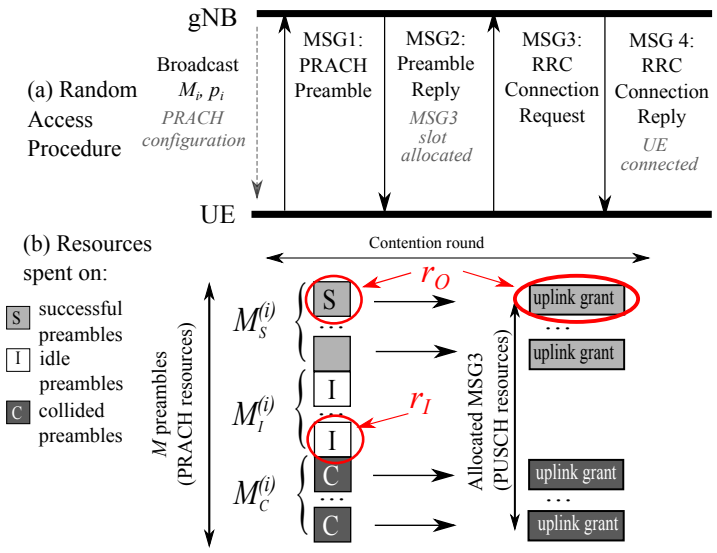

Figure 4.1: (a) Four step RAP; (b) Uplink resource consumption corresponding to MSG1 (deterministic component) and MSG3 (stochastic component).

a parameter to be adjusted, and M_i is set equal to 64, the amount of preambles in one PRACH. In general, however, M_i can also be adjusted dynamically and communicated to the UEs via gNB broadcast [Dua+16; Vil+17b]. gNB can control the number of PRACH allocations per frame, thus reducing or increasing the number of available preambles. M_i might also be set lower than the PRACH allocation allows, as the remaining preambles might be used for contention-free RACH, for another Quality of Service (QoS) class, or for another network slice [Vil+17b]. Note that solving the problem (4.3) requires the knowledge of n_i, which is typically not available. Instead, an estimation of n_i is used. For simplicity, we do not make a distinction between n_i and its estimate throughout the next sections, but we deploy the estimation for practical evaluations in the later Secs. 4.4.2 and 4.5.4.

Remark 4. *As the reader has probably noticed, we have swapped the original problem of reducing the average burst resolution time with the problem of optimizing per-contention-round throughput. This allows for a lean and lightweight approach to optimize the contention parameters, and it can be easily shown that the problems are equivalent as long as the arrival is independent of the throughput, which is typically the case for real systems. We will return to the original problem and will use average burst resolution as a metric for the numerical evaluations in the later sections.*

Following an analysis similar to [WBC15], it is straightforward to see that in the case of access barring without the re-transmission limit, the function (4.2) is expressed as follows:

$$S(p_i, M_i | n_i) = n_i p_i \left(1 - \frac{p_i}{M_i}\right)^{n_i - 1}. \tag{4.4}$$

Most of the state of the art papers search for the optimal access probability $p_i^\star$ and the number of preambles $M_i^\star$ in various settings. Jin *et al.* [Jin+17] propose a dynamic

adaptation of access probability maximizing S

$$p_i^\star = \min(1, M_i/n_i). \tag{4.5}$$

A similar policy is adopted by Duan *et al.* [Dua+16], with the additional step of allocating the preambles $M_i^\star$. None of the approaches, however, considers resource consumption constraint. In the next subsection, we strictly define the resource consumption and then present two approaches how the resource consumption can be considered in designing a resource-aware RAP.

4.3.2 Efficiency vs. Pareto Optimality

The per-preamble definition of efficiency is insufficient for RAP, where the collision happens on the resources allocated *after the initial preamble contention*. Additional PUSCH resources are spent on MSG3 *for every activated* preamble (i.e., occupied channel), making a collision less favorable than an idle preamble from the resource consumption perspective. If a preamble is idle, no resources for MSG3 are allocated, but if a preamble is collided, resources are allocated but wasted due to the collision. In Fig. 4.1, the resource consumption of RAP is illustrated.

To account for additional consumption, we first define variables r_I and r_O as the amount of resources spent per every idle or occupied preamble, respectively. We assume that collision and success consume equal resources. As illustrated in Fig. 4.1(b), r_I includes only PRACH resources spent on 1 preamble, and r_O includes PRACH+PUSCH resources.

Then, define the vector $\mathbf{M}_i = [M_C^{(i)}, M_S^{(i)}, M_I^{(i)}]$, such that $M_i = M_C^{(i)} + M_S^{(i)} + M_I^{(i)}$, as an outcome of the ith contention round. In other words, $\mathbf{M}_i$ is a "split" of the M_i preambles into collided, successful, and idle. This split is determining the total amount of resources consumed during a contention round. Formally, $M_I^{(i)} \triangleq \sum_{j=1}^{M_i} \mathbb{1}_{m_{i,j}=0}$, $M_O^{(i)} \triangleq M_S^{(i)} + M_C^{(i)} = \sum_{j=1}^{M_i} \mathbb{1}_{m_{i,j}\geq 1}$, where $\mathbb{1}_X$ is the indicator function of a subset defined by condition X.

The resource consumption r_i of a contention round i is then defined as

$$r_i \triangleq (M_S^{(i)} + M_C^{(i)})r_O + M_I^{(i)} r_I, \tag{4.6}$$

which makes instantaneous consumption r_i a random variable dependent on $\mathbf{M}_i$, with expectation $R(p_i, M_i|n_i) \triangleq \mathbb{E}[r_i]$.

Definition 4 (Efficiency)**.** *The efficiency of RAP is defined as the ratio of the expected throughput to the expected resource consumption of a contention round:*

$$T(p_i, M_i|n_i) \triangleq \frac{S}{R}. \tag{4.7}$$

For the system model, the efficiency is found as:

$$T(p_i, M_i|n_i) = \frac{\mathbb{E}[s_i]}{r_O M_i + \mathbb{E}\left[M_I^{(i)}\right](r_I - r_O)}. \tag{4.8}$$

With S given via (4.4), remaining is to find the expression for the expected number of occupied $\mathbb{E}[M_O^{(i)}]$ and idle $\mathbb{E}[M_I^{(i)}]$ preambles.

Lemma 1. *Given n_i backlogged UEs and the access probability p_i, the expected number of occupied preambles $\mathbb{E}\left[M_O^{(i)}\right]$ and idle preamble $\mathbb{E}\left[M_I^{(i)}\right]$ in the i^{th} contention round is:*

$$\mathbb{E}\left[M_O^{(i)}\right] = M_i - M_i\left(1 - \frac{p_i}{M_i}\right)^{n_i}, \tag{4.9}$$

$$\mathbb{E}\left[M_I^{(i)}\right] = M_i\left(1 - \frac{p_i}{M_i}\right)^{n_i} \tag{4.10}$$

Proof. Consider a single preamble j first. Denote by $y_{i,j} \triangleq \mathbb{1}_{m_{i,j}\geq 1}$ the binary random variable indicating occupation of the preamble j in the round i. The probability that a given preamble is idle can be obtained then as:

$$\mathbb{P}[y_{i,j} = 0] = \left(1 - \frac{p_i}{M_i}\right)^{n_i}. \tag{4.11}$$

Using the sum of expectations, we obtain (4.9) as:

$$\mathbb{E}[M_O^{(i)}] = M_i \sum_j \mathbb{E}\left[y_{i,j}\right] = M_i\left(1 - \mathbb{P}[y_{i,j} = 0]\right) = M_i - M_i\left(1 - \frac{p_i}{M_i}\right)^{n_i}. \tag{4.12}$$

Similarly, we obtain Eqn. (4.10). □

Using Eqn. (4.4), (4.8), and the results of the lemma, we derive resource consumption and efficiency as:

$$R = M_i\left(r_O + (r_I - r_O)\left(1 - \frac{p_i}{M_i}\right)^{n_i}\right), \tag{4.13}$$

$$T = \frac{n_i p_i \left(1 - \frac{p_i}{M_i}\right)^{n_i - 1}}{M_i\left(r_O + (r_I - r_O)\left(1 - \frac{p_i}{M_i}\right)^{n_i}\right)}. \tag{4.14}$$

Our revised definition captures the difference in the resource consumption of idle and occupied channels by weighting them differently in the expected outcome.

To illustrate the difference between throughput and efficiency, we plot the respective functions (4.4) and (4.14) against p_i for different values of M_i in Fig. 4.2. We choose the exemplary values r_I, r_O considering that PRACH in occupies 6 Resource Blocks (RBs) and an uplink packet least 1 RB, hence, $r_I = 6$ RBs/64 preambles ≈ 0.09 (only 1 PRACH preamble), and $r_O = r_I + 1$ RB ≈ 1.09 (1 PRACH preamble + 1 PUSCH RB). We observe that for the same value of M_i, different values of access probability $p_i^\star$ are needed to maximize S and T. This result comes from the fact that the consumed resources R are coupled with the channel split $\mathbf{M}_i$, making a collision less favorable and, hence, reducing

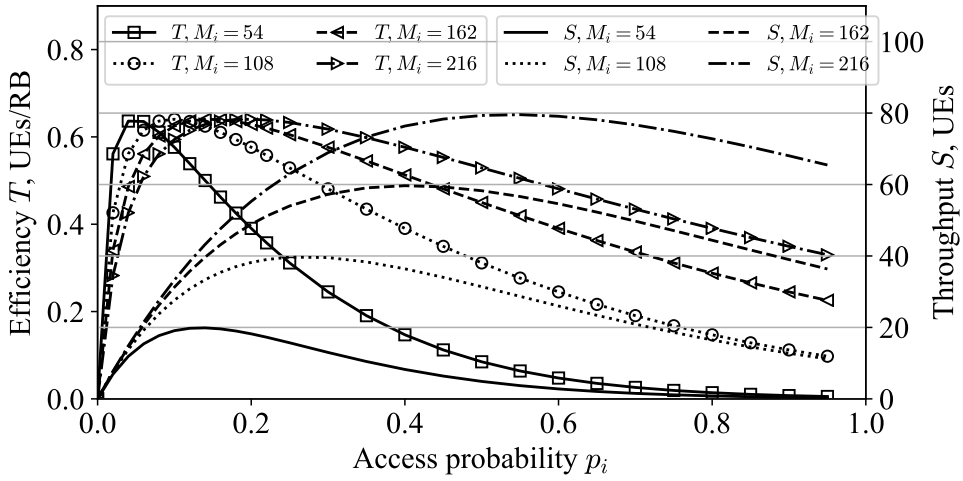

Figure 4.2: Resource efficiency T (left y-axis) and throughput S (right y-axis) vs. access probability p_i for different M_i values. Parameters: $r_I = 0.09$ RBs, $r_O = 1.09$ RBs, $n_i = 400$ UEs.

the optimal access probability maximizing the efficiency. Consequently, the policy (4.5) used in the state of the art [Jin+17; Dua+16] is suboptimal in terms of efficiency T.

One approach to use this result would be to design an algorithm adjusting (p_i, M_i) to maximize the efficiency. However, computing a jointly optimal solution is a non-linear mixed-integer problem and requires numerical methods. The worst-case complexity of such algorithm is not guaranteed to be polynomial. Additionally, since efficiency is a composite objective function, its usage presents *a compromise between two competing metrics*, throughput and resource consumption. In the next sections, we present an alternative approach: To explore the contradicting nature of both metrics, we treat each of them as a separate objective in the framework of a bi-objective optimization problem.

4.3.2.1 Pareto Optimal Random Access Procedure

Instead of considering a single metric as an objective, we formulate a multi-objective optimization problem with two competing objective functions, throughput S and resource consumption R. We aim to maximize the throughput and minimize the resource consumption on the same time, by manipulating optimization variables p_i and M_i.

We formulate the bi-objective optimization problem as follows:

$$\underset{p_i, M_i}{\text{maximize}} \ \{S(p_i, M_i|n_i), -R(p_i, M_i|n_i)\}, \tag{4.15a}$$

$$\text{s.t. } p_i \in (0, 1] \tag{4.15b}$$

$$M_i = h\tilde{M}, \ h \in \mathbb{N}_{++}, \tag{4.15c}$$

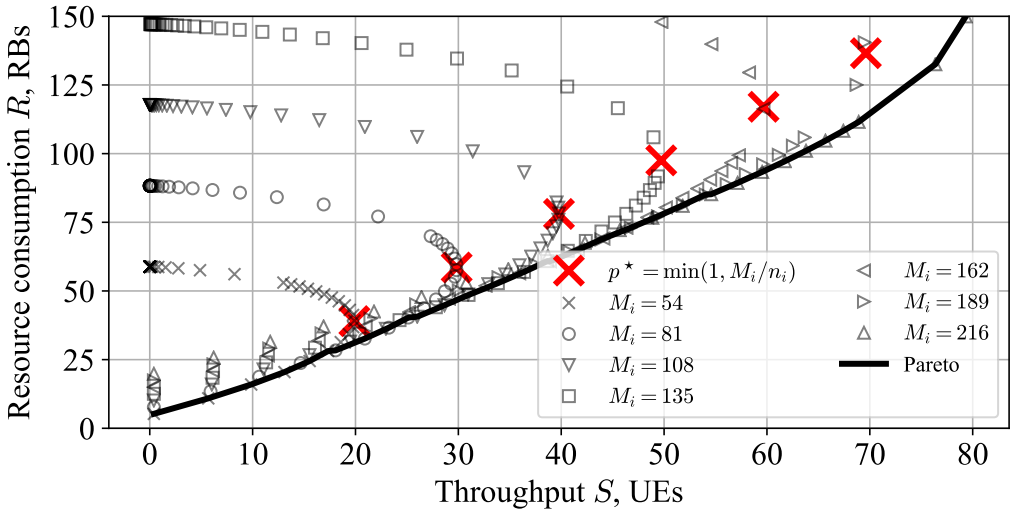

Figure 4.3: Competing objective functions for varying (p_i, M_i) for the problem (4.15) with Pareto set. Parameters: $r_I = 0.09$, $r_O = 1.09$, $n_i = 400$.

With (4.15c), we impose an arbitrary constraint on the preamble allocation granularity: Preambles must be allocated as multiple integer of $\tilde{M} \geq 1$.

Now, we are looking for the Pareto set: Values of (p_i, M_i) for which none of the two objective functions can be increased without decreasing another objective. The sample solution space with the Pareto set is illustrated in Fig. 4.3. Every curve corresponds to a fixed value M_i with varying p_i to produce individual points on a curve. All the points on the lower border of the solution space form the Pareto set. We observe that the points $p_i^\star = \min\left(1, \frac{M_i}{n_i}\right)$, delivered by the state-of-the-art ACB policy (4.5), do not belong to the Pareto set and hence are sub-optimal. The optimality gap can be read from Fig. 4.3 as a respective projection of $p_i^\star$ points on the Pareto frontier.

The problem (4.15) can be solved by *scalarization* [Mie08]: Converting it into a single objective problem using the preferences between the objectives. We choose the ϵ-constraint method: Set a constraint ϵ on one objective, and optimize for the second. A constraint can be set either on the minimum throughput or on the maximum resource consumption, depending on the practical use case and preferences of a generic decision maker. We take the second approach as an example and devise a practical algorithm to implement it in the next section.

4.4 POCA: Pareto Optimal Channel allocation – Access barring algorithm

In this section, we design Pareto Optimal Channel[2] allocation – Access barring (POCA) algorithm to close the optimality gap, which existence we demonstrated in the previous section, and obtain the results from the Pareto set.

4.4.1 Constrained Optimization Problem

There are two ways to re-formulate (4.15) as a constrained problem: (i) To consider the expected throughput as a target, while minimizing resource consumption; Or (ii) to choose the resource consumption as a constraint and maximize the expected throughput S. From a practical point of view, both approaches cloud be valid, and respective solutions in terms of (p_i, M_i) belong to the Pareto set. Here we choose the latter approach. We treat RAP as a constrained optimization problem, maximizing the throughput S given a certain constraint $\epsilon \equiv \bar{r}$ on the expected resource consumption[3].

By imposing the constraint $R \leq \bar{r}$, we reformulate (4.3) as:

$$\begin{aligned} &\underset{p_i, M_i}{\text{maximize}} \; S(p_i, M_i | n_i), \\ &\text{s.t. } R \leq \bar{r} \text{ and (4.15b), (4.15c).} \end{aligned} \tag{4.16}$$

The optimization problem is non-linear and mixed-integer, but a polynomial time solution can be found. First, note that for a fixed M_i, corresponding $p^\star$ has a closed form solution.

Lemma 2. *For fixed M_i, the optimal solution $p^\star(M_i)$ to the problem (4.16) is found as:*

$$p^\star(M_i) = \min\left(\frac{M_i}{n_i}, p_{\max}\right), \tag{4.17}$$

$$\text{where} \quad p_{\max} \triangleq \begin{cases} M_i - M_i \left(\frac{r_O - \bar{r}/M_i}{r_O - r_I}\right)^{\frac{1}{n_i}} & \text{if } r_O \geq \bar{r}/M_i, \\ 1 & \text{if } r_O < \bar{r}/M_i. \end{cases}$$

Proof. Obtained by re-formulating the constraints. □

Second, the solution space with respect to M_i is limited by the granularity $\tilde{M}$ and by $M_{\max} = \bar{r}/r_I$, obtained by setting $p_i = 0$. Hence, the solution to the full problem reduces

[2] We use a more general term "channel" instead of "preamble" in the algorithm's name to emphasize the applicability to other multi-channel slotted ALOHA systems besides LTE RAP.

[3] Separate constraints can also be enforced on PRACH and PUSCH resources. We argue however that a total constraint on the resources spent on connection establishment is more interesting. Both approaches can be accommodated in the proposed framework with minor modifications.

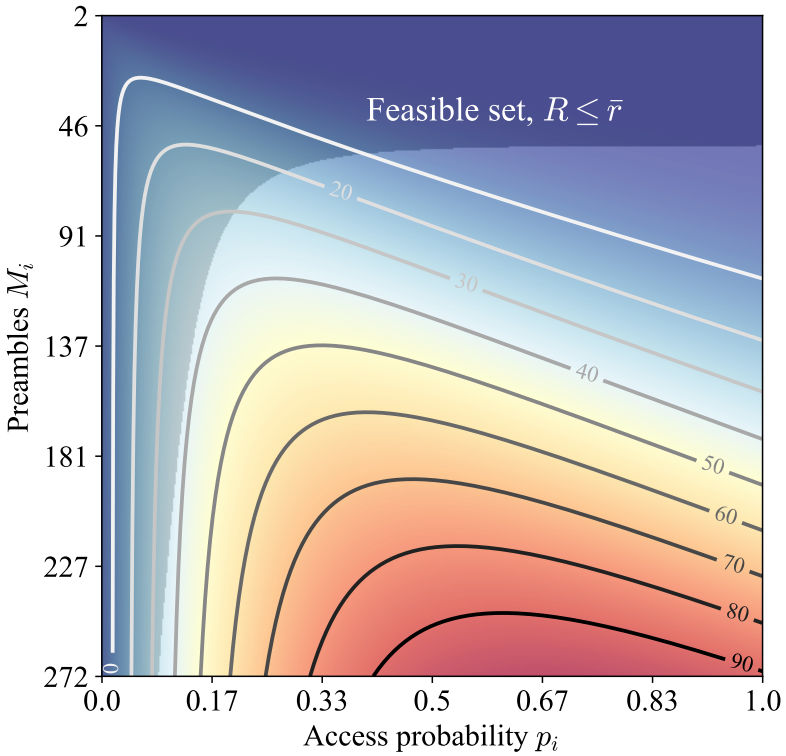

Figure 4.4: Exemplary solution space for the relaxed (continuous preamble allocation) problem (4.16) with $n_i = 400$, $\bar{r} = 60$.

to the search in h, with computation of $p^\star$ for every iteration. Resulting complexity of the algorithm is hence at most linear $\mathcal{O}(h_{\max})$, where $h_{\max} = \lfloor \frac{\bar{r}}{r_I \tilde{M}} \rfloor$. Moreover, it can be shown that the problem is quasi-concave in h, since the Pareto frontier is found as a maximum of monotonic functions $S(h)$.

The resulting pseudocode for POCA is outlined in Alg. 2.

Remark 5. *Since the preamble split* $\mathbf{M}_i$ *is random, we have considered* $\bar{r}$ *as a constraint on expectation, which is a soft constraint. It is a constraint which can be imposed by the desired dimensioning of resources in a system with dynamic scheduling. If desired is a hard constraint instead, it should be accommodated into* S *and treated as unconstrained optimization problem instead.*

4.4.2 Performance Evaluation

In this subsection, the performance of POCA is evaluated and benchmarked with two related solutions. A custom event-based simulator is used, where the implementation assumptions correspond to the system model.

As benchmarks for POCA we choose pseudo-Bayesian approach (Lin17) [Jin+17], dynamic access barring with fixed resource allocation (Duan16-F) and with dynamic resource allocation (Duan16-D) [Dua+16]. The first two, Lin17 and Duan16-F, do not

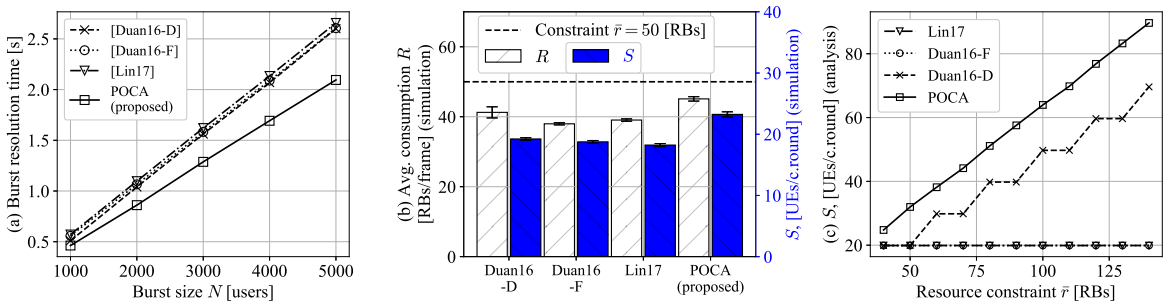

Figure 4.5: Evaluation results: (a) average burst resolution time vs. burst size N, simulated; (b) average resource consumption (left y-axis) and throughput (right y-axis) per contention round R, simulated; and (c) analytically obtained values for throughput vs. resource consumption constraint $\bar{r}$. Burst size for (b) is $N = 2000$, number of back-logged UEs for (c) is $n_i = 400$; For all plots: Activation time $T_a = 10$ ms, $\bar{r} = 50$, $r_O = 1.09$, $r_I = 0.09$.

adjust the number of preambles and only tweak the access probability, while the latter is optimizing both p_i, M_i. We aided the benchmarked algorithms with explicit constraint on the consumption $\bar{r}$. We set the constraint high enough for Duan16-F and Lin17 to deliver feasible solution. For Duan16-D, we find the best solution satisfying the constraint via exhaustive search. We have compared the performance in terms of burst resolution time, throughput, and average resource consumption for a burst arrival scenario [3GP11]. All algorithms require knowledge of current backlog n_i, i.e., how many nodes will attempt the transmission in the next step. The estimation based on the observations of the channel split $\mathbf{M}_i$ [Jin+17] is used for POCA, as it performed best in our simulations. The simulation set-up follows the assumptions in 4.3.1, capturing only the Medium Access Control (MAC) layer effects with main parameters summarized in Fig. 4.5.

In Fig. 4.5a, average burst resolution time is plotted as a function of the burst size N. We observe that for $N = 1000$ UEs the proposed algorithm achieves 9% lower time

Algorithm 2 POCA: Pareto Optimal Channel allocation – Access barring

for every contention round **do**
  Input: n_i, $\bar{r}$, m (resource granularity). Set: $\hat{k} \leftarrow k_{\max}$, $M^\star \leftarrow \hat{k}m$;
  Compute $p^\star = f(M^\star)$ via (4.17), $S^\star(p^\star, M^\star | n_i)$ via Eqn. (4.4).
  while $\hat{k} > 0$ **do**
    Set: $\hat{k} \leftarrow \hat{k} - 1$, $\hat{M} \leftarrow \hat{k}m$.
    Compute $\hat{p} = f(\hat{M})$ via (4.17), and $\hat{S}(\hat{p}, \hat{M} | n_i)$ via Eqn. (4.4).
    if $\hat{S} > S^\star$ **then**
      Set: $p^\star \leftarrow \hat{p}$, $M^\star \leftarrow \hat{M}$, $S^\star \leftarrow \hat{S}$
    end if
  end while
return $p^\star, M^\star$
end for

compare to the closest Duan16-D algorithm, and the gain grows to 19% for large bursts of $N = 5000$ UEs. To study why is it the case, we plot the measured throughput and average resource consumption per contention round in Fig. 4.5b for an exemplary burst size $N = 2000$. We observe that POCA achieves higher throughput and better utilizes the resources within the constraint $\bar{r}$. The deficiency of the benchmark solutions comes from the sub-optimality of the policy (4.5) and from the separate preamble-access probability optimization. In contrast to it, POCA jointly considers preamble allocation and access probability, hence, delivering the solutions from the Pareto set. We further study the difference in throughput as a function of the resource constraint in Fig. 4.5c. Duan16-F and Lin17 do not adjust M_i, hence they achieve the same throughput independent on the constraint. The throughput difference between Duan16-D and POCA is growing with the constraint, which is well inline with the observations in Fig. 4.3, where the optimality gap is increasing with resource consumption.

4.5 Binary Countdown Contention Resolution for RAP

In this section, we go beyond the optimization of the standard-compliant RAP. We propose a novel RAP, where Binary Countdown Contention Resolution (BCCR) is used prior to MSG3 to resolve a possible collision, exploiting the fact that the burst arrivals are typically spatially correlated. First, we recap the BCCR protocol and explain the modified RAP in 4.5.1. For the modified RAP, we study the joint operation of ACB and BCCR in dense networks under burst arrival scenario. We analyze the performance of BCCR and its joint performance with ACB in 4.5.2. Then, we proceed to apply the insights from Pareto optimal approach developed earlier in Sec. 4.3 to the modified RAP, following up with a proposal of Dynamic Binary Countdown - Access barring (DBCA) algorithm. We evaluate DBCA performance in 4.5.4.

4.5.1 Binary Countdown Contention Resolution

In this section, we explain the basics of BCCR protocol (4.5.1.1), its integration into the RAP of NR (4.5.1.2), discuss possible options for priority assignment (4.5.1.3), and give an illustrative example of RAP operation with BCCR (4.5.1.4).

4.5.1.1 Recap: Binary Countdown Protocol

The core idea of BCCR is to use short Contention Resolution Slots (CRSs) prior to the packet transmission to probabilistically "decide" in a distributed fashion which of the contending UEs transmits the packet.

To explain the protocol, we denote the number of CRSs in a given contention round i as k_i, and the number of associated *priority levels* $l_i \triangleq 2^{k_i}$. Before the start of the BCCR procedure, each contending UE u uniformly at random chooses a priority level $\mathbf{p}^{(u)}$. The

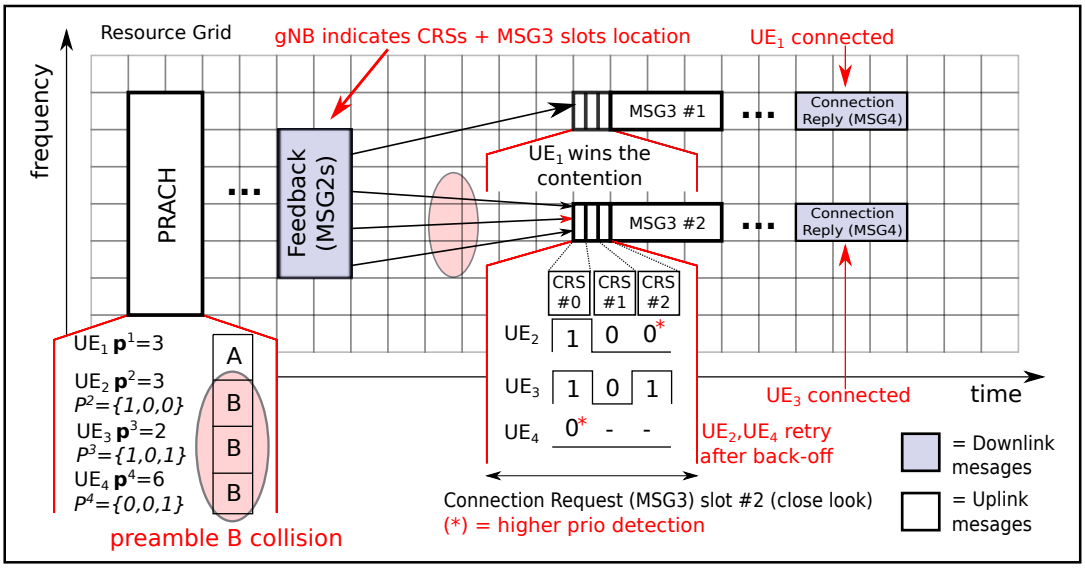

Figure 4.6: Exemplary operation of RA procedure with Binary Countdown Contention Resolution shown for two activated preambles A and B. For every activated preamble, additional resources are allocated for $k_i = 3$ CRSs prior to MSG3 transmission. CRSs remain unused in the case of preamble A, whereas CRSs for preamble B the CRSs are used to resolve the contention.

selected level $\mathbf{p}^{(u)}$ is represented as a k_i-digit binary sequence $\mathbf{P}^{(u)} = \left[P_0^{(u)}, \ldots, P_{k_i-1}^{(u)}\right]$, corresponding to the base-2 representation of $\left(l_i - 1 - \mathbf{p}^{(u)}\right)$, where $P_j^{(u)} \in \{0, 1\}$ and $0 \leq j \leq k_i - 1$. As an example, if we set $k_i = 2$, we have the highest priority level $\mathbf{p}_{\max,i} = 0$ represented by $\mathbf{P}_{\max,i} = [1, 1]$, while the lowest priority $\mathbf{p}_{\min,i} = 3$ is represented by $\mathbf{P}_{\min,i} = [0, 0]$. Here, we follow the convention that 0 *is the highest priority*.

The binary sequence, generated from the chosen priority level, is then used by the UE to decide its behavior in any CRS $\#j$. Starting from the CRS #0 onward, a contending UE u is either listening to the medium if $P_j^{(u)} = 0$, or transmitting a signal to inform other contenders of its presence if $P_j^{(u)} = 1$. If, in any CRS, a silent UE detects another UE transmitting, it assumes there is a contending UE with higher priority and *immediately* abandons the contention, i.e., it does not transmit in any later CRS regardless of its priority. If, on the contrary, a UE completes the k_i CRSs without having detected any UE with higher priority, it assumes that it is the winner of the contention and proceeds to send its packet.

4.5.1.2 Integration in RAP

In contrast to bus or Wi–Fi systems, contention in LTE and NR starts with sending a random PRACH preamble, which makes BCCR not applicable on the first step. However, as the actual collision occurs at the step three (MSG3), we propose to allocate PUSCH resources for BCCR prior to MSG3, hence, extending the MSG3 slot by k_i

CRSs [VRK17]. Thus, the resulting procedure combines two techniques: overload control prior to preamble transmission by the means of ACB and contention resolution prior to MSG3 transmission using BCCR (see the time-frequency grid illustration in Fig. 4.6).

The duration of a CRS has to take into account the granularity of resource allocation, required resources for MSG3 duration, switching time between reception and transmission. These factors are mostly limited by the technology standard. While for LTE the allocation granularity is conservative and limited to 1 sub-frame, in 5G NR smaller and more flexible CRS configurations are possible due to flexible frame structure and finer scheduling granularity down to 1 Orthogonal Frequency Division Multiplexing (OFDM) symbol [DPS18]. To stay inline with NR scheduling, we assume a CRS to consume 1 RB bandwidth x 1 symbol period per CRS basis, so that $t_{\mathrm{CRS}} = 1$ OFDM symbol.

Synchronization of UEs and gNB is handled by a specific Timing Advance (TA) for each device, compensating both the heterogeneity in the uplink propagation delay and its time variability. If multiple UEs are contending for MSG3, they all receive the same TA instructions via MSG2 and thus BCCR in the RAP is intrinsically unsynchronized[4]. Therefore, our proposed approach is not to transmit during the entire contention resolution slot time duration, t_{CRS}, but rather only during its part with duration t'_{CRS}. For the pair of devices $\mathrm{UE}_u, \mathrm{UE}_w$, $u \neq w$, contending to send MSG3 over the same PUSCH resources, we denote d_u, d_w as their respective distances to the BS and $d_{u,w}$ as the distance between each other. For the sake of robustness, it is important to ensure that every UE is able to hear the broadcast from all other contending UEs, arriving entirely within t_{CRS}. Thus, the worst case scenario is when the first contending UE starts transmitting (closest to gNB) and has to wait to hear the last UE (furthest). Denoting closest UE as $u = 1$ and furthest as $w = 2$, this restriction is expressed as:

$$t_{\mathrm{CRS}} \geq (d_2 - d_1)/c + t'_{\mathrm{CRS}} + d_{1,2}/c, \tag{4.18}$$

where c the signal propagation speed, approximately equal to the speed of light. Furthermore, given the triangle inequality $d_w - d_u \leq d_{u,w}$, we can obtain a more restrictive but simpler condition to work with, satisfying Eqn. (4.18): $t_{\mathrm{CRS}} \geq t'_{\mathrm{CRS}} + 2d_{u,wx}/c$.

This allows us to calculate the minimum BCCR broadcast diameter as a function of the ratio $t'_{\mathrm{CRS}}/t_{\mathrm{CRS}}$. Note that, assuming a fixed transmission power, the higher the ratio $t'_{\mathrm{CRS}}/t_{\mathrm{CRS}}$ is, the greater the robustness against Signal to Interference to Noise Ratio (SINR) degradation. E.g., for a ratio of 0.9, we obtain a broadcast distance of approximately 1 km; i.e., every device is able to contend at least with every other device less than 1 km away. Furthermore, it is important to note that aforementioned "bursty" arrivals are typically spatially as well as temporally correlated. Thus, moderate values of the hearing distance are likely to suffice in such scenarios. For larger events there might be performance penalties due to hidden terminal problem. The exact size of fully supported events depends on the UE distribution and placement, network density, etc. and is outside the scope of the chapter, but to be studied in future work.

[4]For static UEs, it is possible to circumvent the problem by storing the last TA value and re-using it during RAP [Ko+12].

4.5.1.3 Priority Assignment

Priority levels can be assigned in a number of different ways. Conventionally, binary countdown sequences and respective priorities are assigned to the users based on their application type. This is typical for binary countdown in CAN bus, because of the inherently hierarchical functioning of the system; i.e., nodes can be easily distinguished by the priority of their function [Geh+14]. In a similar way, LTE/NR UE's priority can be assigned based on QoS class, on a per-user, or even per-flow basis. On top of prioritization, full contention-free access could be potentially achieved, if a sequence is prepended with a unique user identifier. However, this might be hard to implement in in practice, since it requires many CRSs and raises fairness issues. The potential of BCCR for prioritization is studied in our earlier work [VK17a] but it is not covered in this chapter.

Instead, the priority assignment policy we consider here is *uniformly random choice of priorities*. This gives another "channel" dimension for multi-channel ALOHA, similarly to preambles, which allows to improve the overall throughput of RAP. Randomization and prioritization can be even implemented together at the expense of a longer contention resolution period.

4.5.1.4 Example: Random Access Procedure with BCCR

Fig. 4.6 shows an example of BCCR operation with $k_i = 3$ CRSs. After the preamble transmission is received, gNB allocates the resources for BCCR and MSG3 for every activated preamble and informs UEs about the allocated CRSs and their position in the time-frequency grid by MSG2 feedback. In the case of preamble A, it has only been activated by UE_1, so there is no collision to be avoided. Note, however, that UE_1 still needs to perform BCCR prior to sending MSG3, since the number of UEs occupying a certain preamble is unknown. Although unused CRSs introduce extra overhead, we will show in the later sections that this overhead is negligible compared to the gains of BCCR in high-load regime. Moreover, this overhead can be avoided if gNB can distinguish a collided from a singleton preamble during the first step of RAP [Mag+18].

In contrast, UE_2, UE_3, and UE_4 have all activated the same preamble B. They then perform BCCR with randomly chosen priorities $\mathbf{P}^{(2)} = [1, 0, 0]$, $\mathbf{P}^{(3)} = [1, 0, 1]$ and $\mathbf{P}^{(4)} = [0, 0, 1]$. UE_2 and UE_3 transmit a signal in CRS #0, which is sensed by the listening UE_4. Thus, UE_4 immediately abandons the contention and does not participate in any further CRSs, regardless of its priority. In CRS #1, both UE_2 and UE_3 remain silent and listen to the medium, and both detect no transmission. Finally, in CRS #2, UE_2 remains silent while UE_3 transmits a signal. Thus, UE_2 also abandons the contention, leaving UE_3 as sole winner, which then proceeds to sending MSG3 without collisions and to successfully connect with the gNB. In this case, BCCR has avoided what would otherwise have been a wasted RAO, turning it into a successful connection.

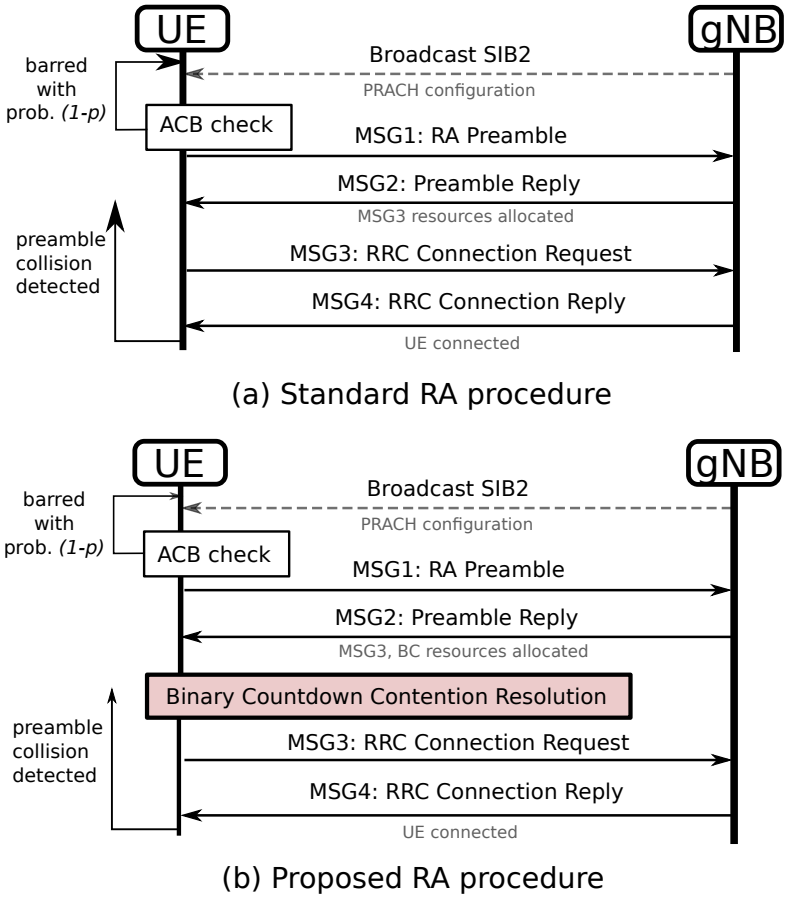

Figure 4.7: (a) Standard RA procedure; (b) Proposed RA procedure aided with an additional step of BCCR prior to MSG3 transmission.

4.5.2 Modeling and Performance Analysis

In this section, we analyse the performance of the joint access barring and binary countdown operation. First, the system model is described (4.5.2.1). Then, we study the performance of BCCR, to understand its gains and overhead (4.5.2.3). We expend the analysis to derive an expected throughput in a single contention round considering both BCCR and ACB (4.5.2.2) and further develop it towards bi-objective optimization problem (4.5.2.4). Finally, we provide an approach to generalize the analysis for the full burst resolution delay (4.5.2.5).

4.5.2.1 System Model

We follow the system model assumptions introduced earlier in Sec. 4.3.1, with the addition of BCCR as described in the joint procedure in 4.5.1. Here, we briefly summarize the model adjustments due to BCCR.

After the MSG2 reception, and prior to BCCR, each preamble can have one of three possible outcomes: **idle** if no device occupies the preamble; **successful** if one and only

one device chooses the preamble; and **collided** otherwise, in accordance with the channel model defined by Eqn. (4.1). Now, for any collided preamble j, at most one UE among those having chosen it, can be successfully resolved via BCCR. For every available preamble, we have one RAO $x_{i,j}$ associated to it. Recall that the number of UEs choosing a given preamble j is denoted by $m_{i,j}$, and outcome of a RAO by $x_{i,j}$. Resulting extension of the collision channel (4.1), to account for BCCR is

$$x_{i,j} \triangleq \begin{cases} 1 & m_{i,j} = 1 \ \cup \ (m_{i,j} > 1 \ \cap \ \text{resolved via BCCR}), \\ 0 & \text{otherwise.} \end{cases} \tag{4.19}$$

The contention in a collided preamble ($m_{i,j} > 1$) is defined as *resolved* via BCCR, if one of the UEs has uniquely chosen the highest priority among the set of the priorities selected by UEs occupying preamble j.

4.5.2.2 Joint ACB – BCCR Performance

In the following part of the section, we analyze the RAP under joint action of ACB and BCCR. Consider the system state prior to a contention round i. We denote the number of competing UEs at this point as n_i. Since we assume p_i-persistent ACB with no drops, n_i accounts both for backlogged users and newly arrived ones as there is no distinction between their behavior. The performance in terms of the throughput is then described by the Theorem 2.

Theorem 2. *Given n_i competing UEs, access probability p_i, and l_i BCCR priority levels, the expected number of successful UEs S_J in the contention round i is:*

$$S_J = \frac{n_i p_i}{l_i} \sum_{v=1}^{l_i} \left(1 - \frac{v}{l_i} \frac{p_i}{M_i}\right)^{n_i - 1}. \tag{4.20}$$

Proof. See App. 4.A. □

The implications of the theorem are illustrated in Fig. 4.8, where the expected number of successful UEs is plotted as a function of p_i for different values of l_i for a fixed $M_i = 54$ preambles and $n_i = 1000$ UEs. As expected, increasing l_i improves the performance, and increases the supported load by shifting the peak of the curve to the right. We also observe that the analytical results are closely matching the simulation.

Additionally, in Fig. 4.9, we simulatively study the effects of violating the assumption that all UEs are in the overhearing range of each other, which we have made since 4.5.1.2. We define a *distance penalty* p_f as the probability that a BCCR broadcast is not received (failure) on single link between any two UEs. Under a simple yet common assumption that the failures are Bernoulli-distributed with mean p_f, we simulate the same scenarios as in 4.8 for fixed $k_i = 3$ ($l_i = 8$) but vary p_f. We observe that even with $p_f = 0.5$, with half messages lost on average, the performance with BCCR is significantly better than the baseline.

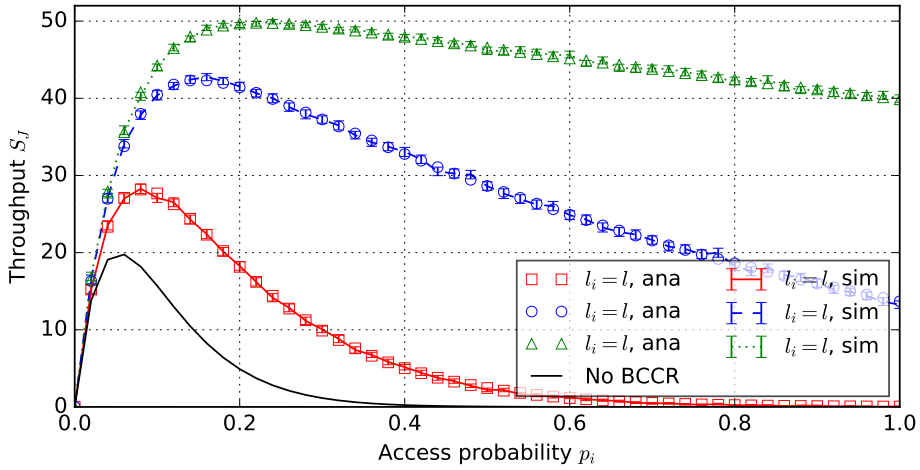

Figure 4.8: Expected number of successful UEs S_J vs. access probability p_i in a contention round i, $M_i = 54$ preambles, $n_i = 1000$ UEs, .95 confidence intervals.

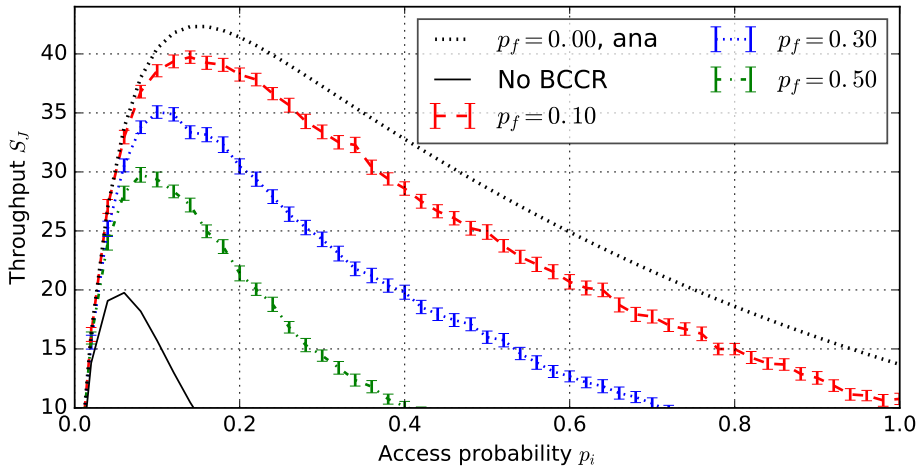

Figure 4.9: Expected number of successful UEs S_J with the probability of not hearing a BCCR broadcast $p_f \in \{0, 0.05, 0.1, 0.3\}$ (distance penalty); $M_i = 54$ preambles, $n_i = 1000$ UEs, $l_i = 8$ priority levels, .95 confidence intervals.

4.5.2.3 BCCR Overhead and Efficiency

Next, we assess the gains and trade-off of introducing BCCR in the system. Adding BCCR in RAP is introducing overhead due to the resources reserved for CRSs. BCCR operating with l_i priority levels requires $k_i = \lceil \log_2 l_i \rceil$ CRSs.

In order to assess the trade-off, we first need to quantify the number of resources consumed per contention round. We follow here the approach from Sec. 4.3 and focus only on the consumed resources in the uplink channels (PRACH and PUSCH). The number of occupied preambles is increasing with increasing access probability p_i. According to

the procedure, for every occupied (activated) preamble, resources for MSG3 + k_i CRSs transmissions are allocated. Using the results of Lemma 1, we can extend the expected resource consumption derivation given by Eqn. (4.13) to RAP with BCCR.

Corollary 3. *For a given contention round i, the expected uplink resource consumption of RAP with BCCR, as a function of the number of contending UEs n_i, is:*

$$R_{\text{J}} = M_i r_I + r_3 \left(1 + k_i \delta\right) \underbrace{\left(M_i - M_i \left(1 - p_i / M_i\right)^{n_i}\right)}_{\text{expected occupied preambles } M_O^{(i)}} \text{ RBs.} \tag{4.21}$$

where $M_i r_I$ are the resources consumed by PRACH, $r_3 \triangleq (r_O - r_I)$ the PUSCH resources consumed by per every MSG3 transmission and δ is the relative overhead introduced by each CRS with respect to r_3.

Proof. Follows directly from Lemma 1. □

Recall the definition of efficiency (4.7) $T \triangleq S/R$, the amount of successful requests normalized by the total number of resources spent, during a single contention round. Accordingly, for RAP with BCCR we get $T_{\text{J}} \triangleq S_{\text{J}}/R_{\text{J}}$. To evaluate the trade-off of introducing BCCR, we consider the ratio T_{J}/T, characterizing the efficiency *gain*. The gain is shown in Fig. 4.10 for the case of no ACB, i.e. $p_i = 1$. We fixed $M_i r_I = 6$ RBs, since PRACH typically occupies 6 RBs in LTE. The value for r_3 might in general vary due to protocol implementation and channel variation. Here, we assume $r_3 = 2$ RBs [JPS17].

We show the gain for three different per CRS overhead values δ: 0.07 (proposed option of 1 symbol per CRS with $r_3 = 2$ RBs), medium value 0.15, and very high value 0.5 (for $r_3 = 2$, it corresponds to 1 RB long CRS). We observe from Fig. 4.10 that for the proposed BCCR implementation, efficiency gain exists even for low number of contending UEs. However, for very high CRS duration $\delta = 0.5$, BCCR usage only makes sense if high number of contending UEs is high $n_i \geq 70$. The higher is the number of CRSs, the higher is the gain increase with the number of UEs n_i, however, also the higher is the minimum number of UEs where gain is larger than 1. This motivates the dynamic allocation of CRSs: The higher is the anticipated load in the contention round (i.e., the estimated back-log), the larger k_i should be allocated by the gNB.

4.5.2.4 Bi-objective Optimization

When applying ACB in the RACH, the *access probability* p_i must be chosen and broadcast by the gNB prior to every contention round i. Adding BCCR introduces a new design parameter into the problem, namely the number of CRSs k_i. Its value must also be chosen by the gNB and communicated to the UEs along with the p_i, so that it is known by all the participants prior to MSG3 transmissions. For a given n_i, we define a pair of values (p_i, k_i) as an *operating point*.

In the state of the art, RAP optimization is typically approached as a maximization of the throughput. With access probability p_i being the only design parameter (no BCCR),

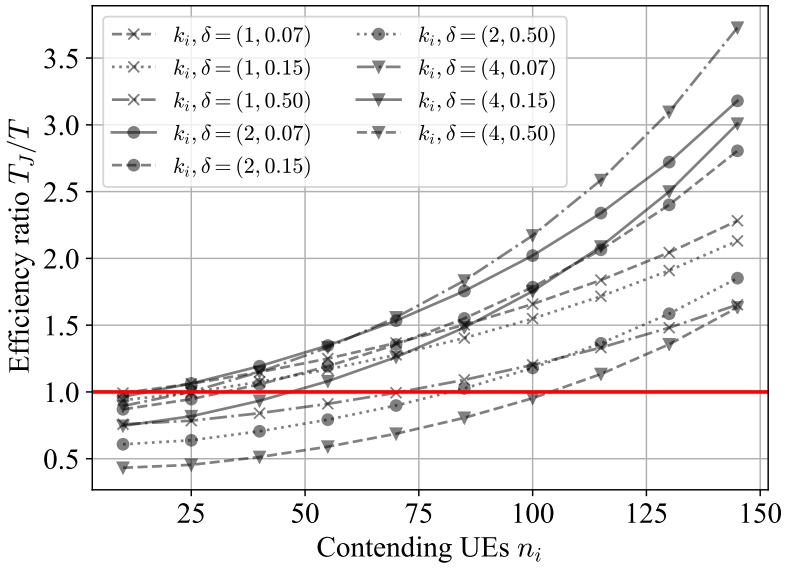

Figure 4.10: Ratio T_J/T quantifying efficiency gain of BCCR vs. number of contending (non-barred, i.e., $p_i = 1$) UEs n_i. $k_i \in \{1, 2, 4\}$ CRSs, relative overhead of CRSs $\delta \in \{0.07, 0.15, 0.5\}$; $M_i = 54$.

there is a single optimal point given by Eqn. (4.5). However, this approach is not directly applicable to our modified procedure. It is clear that, for $n_i > 1$, increasing k_i always has a positive effect on the throughput, and it is intuitively clear that BCCR can achieve an arbitrary small collision probability. However, this does not account for the fact that CRSs consume additional time-frequency resources, which introduces overhead compared to the ACB-only RAP. Hence, we face a fundamental trade-off between two competing optimization goals: maximizing the expected number of successes per RAP and minimizing the expected resource consumption. We thus apply the optimization approach analogous to Sec. 4.3, but with number of CRSs k_i as a second parameter instead of the number of preambles M_i[5]. We have evaluated this trade-off in terms of efficiency in 4.5.2.3, and here, we extend it towards a bi-objective optimization with S_J and R_J as competing objectives.

[5]It is possible to formulate a problem with three optimization variable (p_i, k_i, M_i), but we leave this as an opportunity for extension in future work.

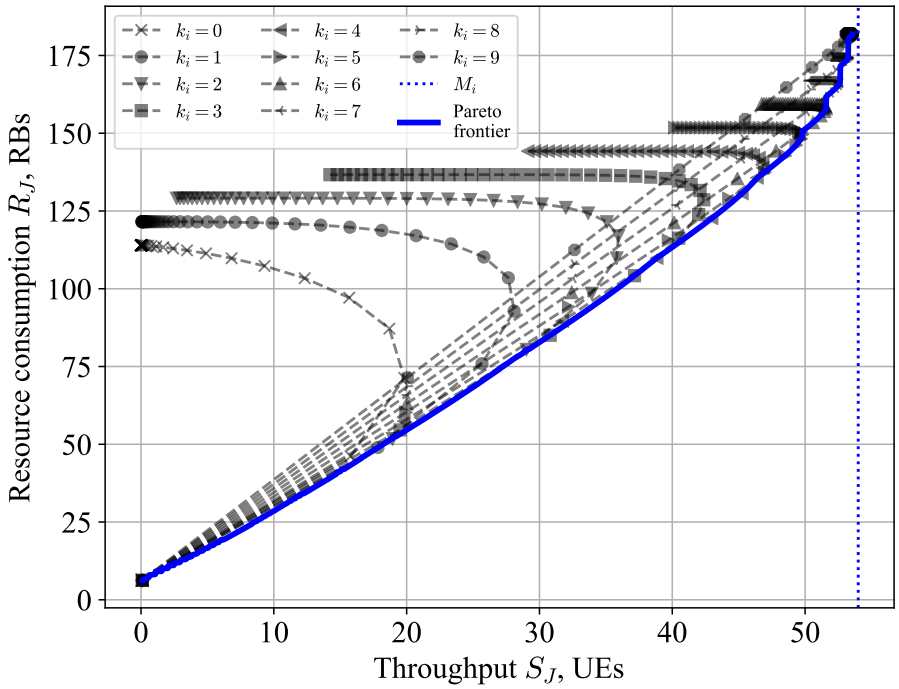

Figure 4.11: Expected consumed resources vs. expected throughput, showing the Pareto frontier. $M_i r_I = 6$ RBs, $r_3 = 2$ RBs, $\delta = 0.07$, $M_i = 54$ preambles, $n_i = 1000$ UEs.

The bi-objective optimization problem is formulated as follows:

$$\min_{p_i, k_i} \quad \{-S_J, R_J\} \tag{4.22a}$$

$$\text{with} \quad S_J = \frac{n_i p_i}{l_i} \sum_{v=1}^{l_i} \left(1 - \frac{v}{l_i} \frac{p_i}{M_i}\right)^{n_i - 1}$$

$$R_J = M_i r_I + r_3(1 + k_i \delta) \left(M_i - M_i \left(1 - \frac{p_i}{M_i}\right)^{n_i}\right)$$

$$\text{s.t.} \quad p_i \in (0, 1], \tag{4.22b}$$

$$k_i \in \mathbb{Z}_{\geq 0}. \tag{4.22c}$$

As it is a multi-objective optimization problem, we study Pareto optimal points, that is, solutions for which there is no other possible solution which simultaneously performs better with respect to one of the optimization goals without degrading the other. These points constitute the Pareto frontier. Introducing BCCR into the RAP dramatically modifies how the structure of Pareto frontier looks like (see Sec. 4.3 for comparison), but it does not modify the problem's dual nature of conflicting optimization goals.

Exemplary Pareto frontier produced numerically for the optimization problem defined by (4.22) is plotted in Fig. 4.11 for $n_i = 1000$, assuming that MSG3 takes

$r_3 = 2$ RBs (1 sub-frame in time domain), and a CRS occupies 1 OFDM symbol, i.e., $\delta = \frac{1 \text{ symbol}}{1 \text{ sub-frame}} \approx 0.07$. Every black curve corresponds to achievable performance for a fixed value of k_i, and varying the values of p_i. We observe that the Pareto frontier is a combination of achievable performance curves for different values of k_i. The ratio of S_J/R_J is almost constant through a large part of the Pareto frontier, and starts decreasing as throughput approaches the total number of preambles M_i. This means that for large values of k_i, increasing it further results only in marginal improvement of throughput.

We further observe that the Pareto frontier exhibits an asymptotic behavior at:

$$\lim_{k_i \to +\infty} S = \mathbb{E}[M_O^{(i)} | p_i = 1] = M_i - M_i \left(1 - 1/M_i\right)^{n_i}. \tag{4.23}$$

This follows since the expected throughput is constrained by the maximum expected number of occupied preambles, while the expected resource consumption, in our simplified model, is not constrained at all. It readily follows that:

$$\lim_{n_i \to +\infty} \mathbb{E}[M_i^O | p_i = 1] = M_i. \tag{4.24}$$

Hence, we can asymptotically achieve normalized per-preamble throughput of 1. This result coincides with analytical studies of other binary countdown-based protocols, showing that arbitrary small collision probability could be achieved [Bai+17].

We will return to the bi-objective optimization and design a practical burst resolution algorithm obtaining a Pareto-optimal solution later in Sec. 4.5.3.

4.5.2.5 Full Burst: Expected Resolution Time

To generalize the single contention round analysis towards the full burst resolution time t_{BR}, we apply a modified *drift approximation model* proposed by Wei *et al.* [WBC15]. Since describing the exact evolution of the backlog over time is a computationally complex problem, the authors [WBC15] propose to approximate it by considering only the evolution of the expectation of the backlog. Let us introduce additional notation of new arrivals during the contention round i as a_i. The backlog state at any time slot is thus represented by the following Lindley recursion:

$$n_{i+1} = n_i - \underbrace{s_i}_{\text{successful UEs}} + \underbrace{a_i.}_{\text{new arrivals}} \tag{4.25}$$

Now, to compute the expected burst resolution time, the recursion is approximated by its expectation:

$$\mathbb{E}[n_{i+1}] = \mathbb{E}[n_i] - \mathbb{E}[s_i] + \mathbb{E}[a_i]. \tag{4.26}$$

Expected success $\mathbb{E}[s_i] = S$ in a given round is computed via Theorem 2, and the expected arrivals in a round, dependent on the arrival process, are computed via the

probability density function of the activation time $g_a(t)$ as:

$$\mathbb{E}[a_i] = N \int_{(i-1)T_{\text{C.R.}}}^{iT_{\text{C.R.}}} g_a(t)dt, \tag{4.27}$$

where $T_{\text{C.R.}}$ denotes the duration of a contention round. Now, computing the expected burst resolution time $\mathbb{E}[T_{\text{BR}}]$ with an arbitrary precision of the backlog γ simplifies to an iterative application of (4.26) starting with $i = 0$ and with a stopping condition:

$$\mathbb{E}[T_{\text{BR}}] = i, \quad \text{if } \mathbb{E}[n_i] < \gamma \text{ and } \mathbb{E}[a_j] = 0 \; \forall \; j \geq i. \tag{4.28}$$

4.5.3 DBCA: Dynamic Binary Countdown - Access barring

In this section, applying the analytical results and observations from the previous sections for the practical design of RAP, we propose a DBCA protocol. In the core of the protocol is the idea to dynamically determine the values of p_i and k_i from the Pareto frontier for every contention round i. To make the protocol more practical, we also aid it with a backlog estimator, since backlog is unknown to the gNB in most of the scenarios.

DBCA protocol consists of the following main steps, repeated in every contention round:

I. Contenting UEs undergo ACB and (if successful) transmit MSG1s

II. gNB receives MSG1s and updates the estimate of the number of contending UEs $\hat{n}_i$.

III. Based on the estimate $\hat{n}_i$, gNB calculates the number of CRSs k_i to be used for MSG3 transmissions and informs contending UEs about it as part of the MSG2.

IV. UEs undergo BCCR and (if successful) transmit MSG3s. If unsuccessful, they back-off until the next round.

V. gNB receives MSG3s and updates the estimate of the number of backlogged UEs $\hat{n}^-_{i+1}$ for the next round.

VI. Based on the estimate $\hat{n}^-_{i+1}$, gNB calculates barring factor p_{i+1} for the next round, and informs UEs via system information broadcast.

The pseudocode for gNB-side of DBCA is presented in Algorithm 3. In the following, we explain in the choice of the operating point (p_i, k_i) given the backlog estimate in steps III and VI and the estimation procedure in steps II and V.

4.5.3.1 Choosing the operating point (p_i, k_i)

The core of the DBCA algorithm, corresponding to steps III and VI of the algorithm, is choosing the operating point on the Pareto curve. To obtain a Pareto-optimal solution, the respective bi-objective optimization problem is solved by scalarization, where we convert both objectives into one using the preferences of a decision maker. We apply

scalarization by ϵ-constraint method [Mie08]: A constraint is set on one objective function, and the system of optimized for the second objective. Either the minimum desired throughput or the maximum allowed resource consumption can be used as a constraint. Typically, however, the resource constraint $R_J \leq \bar{r}$ would be a major limiting factor. In this case, the optimization problem targets the maximization of the expected throughput subject to the constraint on the expected resource consumption. We formulate it as follows:

$$\max_{p_i,k_i} \quad S_J\left(k_i, p_i | n_i\right) \tag{4.29a}$$

$$\text{s.t.} \quad R_J\left(k_i, p_i | n_i\right) \leq \bar{r} \tag{4.29b}$$

$$k_i \in \mathbb{Z}_+,\ p_i \in (0,1] \tag{4.29c}$$

Remark 6 (On the Pareto optimality)**.** *ϵ-constrained method ensures at least weak Pareto optimality [Mie08]. If multiple optimal solutions to the problem (4.29) are found, strong Pareto optimality with respect to the original problem (4.22) can be enforced by choosing the solution with the lowest resource consumption.*

The operating point choice is split into two stages, since p_i and k_i must be allocated at different times: k_i prior to MSG3 and p_i prior to MSG1. First, consider the k_i allocation at stage II. To maximize the expected number of successes, gNB observes the outcome of the preamble transmission (number of activated preambles $M_I^{(i)}$) and decides k_i according to the estimated $\hat{n}_i$ subject to the resource consumption constraint $\bar{r}$. Setting $R_J = \bar{r}$, solving Eqn. (4.21) for k_i, and rounding to the nearest integer, we obtain the decision rule for the number of CRSs:

$$k_i = \left\lceil \frac{1}{\delta} \left(\frac{\bar{r} - M_i r_I}{r_3 \left(M_i - M_i \left(1 - p_i / M_i\right)^{\hat{n}_i} \right)} - 1 \right) \right\rfloor . \tag{4.30}$$

Then, k_i is communicated to the UEs as a part of the MSG2 alongside with the uplink grants for the MSG3 transmission.

Remark 7. *Note that we are again considering soft constraints, which apply only to the expectations. Here, hard constraint can be enforced by substituting the term $M_i \left(1 - p_i/M_i\right)^{\hat{n}_i}$ in Eqn. (4.30), representing the expected number of idle preambles, with the observed value $M_I^{(i)}$.*

Later, upon completion of the contention round (line 11 of the pseudocode), gNB observes the number of successful outcomes s_i and updates the backlog estimation. At this moment, p_{i+1} for the next cycle is decided, as a part of the solution to the problem (4.29), where we use a priori backlog estimate $\hat{n}_{i+1}^-$ for n_i. This solution can be found numerically, and we will return to the complexity of the solution in Sec. 4.5.3.3. This access probability is then broadcast before the $(i+1)^{\text{th}}$ contention round.

Algorithm 3 Pseudocode for Dynamic Binary Countdown - Access barring: gNB View.

Initialize $i = 0$, $\hat{n}_0^- = 1$, $p_0 = 1$, $q_0 = 0$
for every contention round i **do**
  Observe $M_I^{(i)}$ ▷ stage **II**
  Compute $\Delta\hat{n}_i$ via Eqn. (4.31)
  $\hat{n}_i = \hat{n}_i^- + \Delta\hat{n}_i$ ▷ update a posteriori backlog estimate
  Compute k_i via Eqn. (4.30) ▷ stage **III**
  Allocate resources for k_i CRSs and MSG3s
  if $\Delta\hat{n}_i > 0$ **then** ▷ stage **V**
    $q_{i+1} = q_i + 1$ ▷ correction for bursty arrivals
  else $q_{i+1} = 0$
  end if
  Observe successful MSG3 transmissions s_i
  $\hat{n}_{i+1}^- = \hat{n}_i^- + q_{i+1}\Delta\hat{n} - s_i$ ▷ update a priori backlog estimate
  Compute p_{i+1} via Eqn. (4.29) ▷ stage **VI**
end for

4.5.3.2 Estimating the Backlog $\hat{n}_i$

In most of the practical cases, the size of the backlog at any time step n_i is unknown to the gNB. Hence, we have to adapt the procedure in order to obtain an estimate of the backlog $\hat{n}_i$. There exist multiple state-of-the-art estimation technique, all relying on the observation of each contention round outcomes, i.e., number of idle $M_I^{(i)}$ and occupied $M_O^{(i)}$ preambles. In this work, we adapt the pseudo-bayesian estimation from [Jin+17] to the joint procedure.

The estimation of the backlog is reflected at two points in the algorithm: to decide the number of contention resolution slots (stage I) after observing the number of idle preambles $M_I^{(i)}$ (note that at this moment the number of successful UEs is unknown); and to decide the access probability for the $(i+1)^{\text{th}}$ contention round, after the number of successful UEs s_i is already known (stage V).

First, let us consider stage I. It calculates the a posteriori estimation $\hat{n}_i$ as a function of the a priori estimate $\hat{n}_i^-$ (which depends on the previous RAP round estimation, hence its recursiveness) and the number of idle preambles $M_I^{(i)}$. The backlog size in the i^{th} contention round is approximated by a Poisson random variable whose mean is the a priori estimate $\hat{n}_i^-$ and calculates the correction [Jin+17]:

$$\Delta\hat{n} = p_i\hat{n}_i^- \left(e^{-\frac{p_i\hat{n}_i^-}{M_i}} - \frac{M_I^{(i)}}{M_i} \right) \left(1 - e^{-\frac{p_i\hat{n}_i^-}{M_i}} \right)^{-1}, \tag{4.31}$$

The a priori estimation is then corrected:

$$\hat{n}_i = \hat{n}_i^- + \Delta\hat{n} \tag{4.32}$$

The stage V starts once the results of the complete i^{th} contention round are obtained.

A simple a priori estimate for the next contention round $i+1$ is computed as in [Jin+17]:

$$\hat{n}_{i+1}^{-} = \hat{n}_i + \alpha_{i+1}^{-} - s_i \tag{4.33}$$

where α_{i+1}^{-} is an *a priori* estimation of the *arrivals* during the next round. As we assume that no information about the arrivals distribution is available, we take α_{t+1}^{-} proportionally to the number of arrivals in the previous RA round and estimate it as $\alpha_i = \max(0, \Delta\hat{n}_i)$. As the estimation we use is an adaptation of the Enhanced Pseudo-Bayesian ACB algorithm from [Jin+17], we also use a heuristic involving a "boosting factor" q_{t+1} in the a priori estimation to better adjust for the burst arrivals:

$$\alpha_{i+1}^{-} = q_{i+1} \cdot \alpha_i = q_{i+1} \cdot \max(0, \Delta\hat{n}_i) \tag{4.34}$$

4.5.3.3 Complexity Discussion

Clearly, the algorithm complexity is dominated by the line 13 of the pseudocode, where p_{i+1} is computed as a solution to the problem (4.29). While this is a non-linear mixed integer problem, it is possible to find a solution efficiently, considering that if we fix k_i, the resulting problem of finding optimal $p_i^\star$ has a unique solution. We state the second fact as Lemma 3.

Lemma 3. *Given fixed number of CRSs $k_i = \bar{k}$, and the number of UEs $n_i = \bar{n} > 2$, the optimization problem*

$$\max_{p_i} \; S_J, \quad \text{s.t. } R_J \le \bar{r}, \quad p_i \in (0, 1], \tag{4.35}$$

has a unique solution $p_i^\star$.

Proof. See Appendix 4.B. □

The Lemma 3 implies that for any fixed k_i, we can find optimal $p_i^\star$ fast with any numerical methods or local search, e.g., gradient descent. To further simplify the problem, we convert it to a root finding problem in Lemma 4.

Lemma 4. *The problem defined by Eqn. (4.35) can be equivalently solved by*

$$p_i^\star = \min\left(\frac{\phi^\star M_i l_i}{n_i}, p_{\max}^{(J)}\right), \tag{4.36}$$

where $p_{\max}^{(J)}$ is given by (4.52) and $\phi^\star$ found either as a root of

$$(1-\phi) + e^{-\phi i}\left((1-\phi l_i)\left(e^{-\phi}-1\right)+\phi\right) - e^{-\phi} = 0, \tag{4.37}$$

or as $\phi^\star = \frac{n_i}{M_i l_i}$ if no roots exists for $\phi \in \left(0, \frac{n_i}{M_i l_i}\right]$.

Proof. See Appendix 4.C. □

Practically, a naïve Python implementation according to Lemma 4 based on scipy.optimize package [JOP+01] yields $20-35$ μs average execution time. Using the fact the objective function is increasing in k_i and in any realistic implementation k_i is upper-bounded by $k_{\max}$[6], the optimal operating point $(p_i^\star, k_i^\star)$ can be found with a search over $[0, k_{\max}]$. Hence, the worst-case complexity of the step is $\mathcal{O}(k_{\max})$.

4.5.4 Simulations and Performance Evaluation

In this section, we present simulative evaluation of the performance of DBCA and compare it to the baseline of Dynamic Access Class Barring (d-ACB) [Dua+16] and Q-ary Tree Resolution Algorithms (Q-TRA) [MSP14] by the means of a custom event-based simulator. We present the simulation set-up (4.5.4.1), metrics (4.5.4.2), and finally the results (4.5.4.3).

4.5.4.1 Simulation Set-up

We simulate a burst arrival scenario, with three burst arrival distributions [3GP11]: delta, uniform activation time distribution, and beta arrivals:

$$g_a(t) = \begin{cases} \frac{t^{\alpha_B-1}(T_a-t)^{\beta_B-1}}{T_a^{\alpha_B+\beta_B-2}B(\alpha_B,\beta_B)}, & 0 \le t < T_a \quad \text{(if beta)}, \\ \frac{1}{T_a} & 0 \le t < T_a \quad \text{(if uniform)}, \\ 1 & t = 0 \quad \text{(if delta)}, \\ 0 & \text{otherwise.} \end{cases} \tag{4.38}$$

where $B(\alpha_B, \beta_B)$ denotes the Beta function.

We simulate the four-way handshake of RAP using the collision channel model as defined by (4.19), hence, our simulation only captures MAC layer effects. The simulation is organized in contention rounds, where UEs are assumed to receive MSG4 if successful. If no MSG4 is received by the end of a contention round, UE assumes a collision. For simplicity, we assume that a contention round takes one PRACH slot, and a PRACH slot length is assumed to be equal to 10 ms, which corresponds to one PRACH allocation per frame (e.g., configuration index 5 in LTE or 18 in NR) [3GP15b; 3GP18a]. Considering a more practical model would potentially have a quantitative effect on the results, but it is left aside in order to obtain more illustrative performance evaluation. We choose an exemplary value of $k_{\max} = 14$ by assuming that the amount of resources spent on BCCR is at most equal to the resources spent on MSG3, i.e., $k_{\max}\delta = r_3$, and the duration of one CRS equals to one OFDM symbol. The simulation parameters are summarized in the Table 4.2. We present average values obtained from at least 30 Monte-Carlo simulations for each data point, with 95 % confidence intervals not exceeding 1.1 % of the mean.

[6] In general, resource grid and resource management might impose different granularity constraints on the allocation of CRS. As this constraints are implementation specific and hard to model realistically, in this work we assume that k_i could be allocated with granularity 1, i.e, any number of slots up to $k_{\max}$ can be allocated.

Table 4.2: Summary of simulation parameters.

Contention round / PRACH slot $T_{\text{C.R.}} = T_{\text{PRACH}}$	10 ms
Preambles per slot M_i	54
Number of UEs N	$500 - 10000$
Act. time T_a: uniform, beta	1 s / 100 c.rounds
Act. time T_a: delta	1 c.round
Beta distribution parameters (α_B, β_B)	(3,4) [3GP11]
Resource constraint proportionality constant C	$1.0 - 1.8$
Maximum number of CRSs $k_{\max}$	14
CRS allocation granularity	1
Resources per PRACH channel $M_i r_I$	6 RBs
Resources per MSG3 r_3	2 RBs [JPS17]
Single CRS relative overhead $\delta = r_{\text{CRS}}/r_3$	0.07
CRS duration t_{CRS}	1 OFDM symbol

Remark 8. *We only evaluate our approach for the case of one burst without any background traffic and assume that the amount of background traffic is negligible, as is common in the literature, e.g., [Dua+16; G¨17b; WBC15]. Although DBCA is not optimized for the presence of background traffic, the estimation steps II and V would implicitly take it into account by over-estimating the number of back-logged UEs involved in a contention. Alternatively, if a localized burst arrival is detected or anticipated (e.g., during group paging), some preambles could be reserved specifically for the burst resolution and advertised in the system broadcast respectively [Vil+17b].*

4.5.4.2 Performance Metrics

Three performance metrics are investigated: mean service time $\bar{t}_s \triangleq \sum_{j=1}^{N} t_s^j / N$ (time until a UE successfully completes RAP), mean consumed uplink resources $R_{\sum}$ throughout the whole burst resolution duration, and mean resource efficiency: successful outcomes, normalized by the consumed uplink resources. To provide a fair comparison, the per-contention round resource constraint for DBCA is set proportional to the expected resource consumption of a d-ACB algorithm [Dua+16] under the same conditions:

$$\bar{r} = C \times R\left(\hat{n}_i, p_i^{\star} | k_i = 0\right), \tag{4.39}$$

where C is the proportionality constant and $p_i^{\star}$ is the access probability maximizing the expected success rate for RA without BCCR as defined by (4.17). Intuitive meaning of the proportionality constraint is the following. The case of $C = 1$ corresponds to

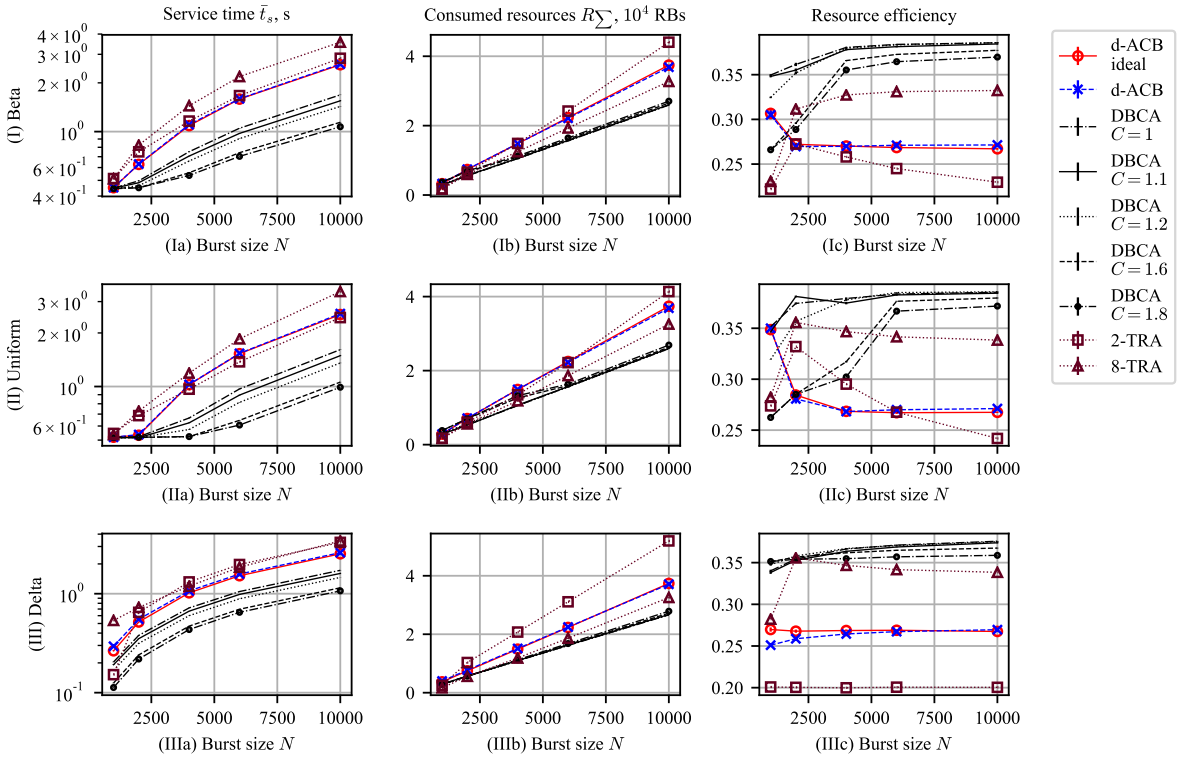

Figure 4.12: DBCA evaluation and comparison with the baseline [Duan16-F] for different values of C. (a) Mean consumed resources, (b) Mean service time $\bar{t}_s$, (c) Mean resource efficiency vs. burst size N. The 95 % confidence intervals do not exceed 1.1 % of the mean value. They are thus omitted to avoid visual clutter.

the case where the proposed algorithm DBCA cannot consume more resources per contention round than the baseline d-ACB, so it provides a fair comparison. The case of $C > 1$ studies how can DBCA benefits from the additional resources, which d-ACB cannot make use of. It is not straightforward to enforce resource constraint on the TRA, therefore we simulate TRA without resource constraint, giving it an advantage. We chose to simulate TRA with branching factors $Q \in \{2, 8\}$. For illustrative purposes, we also include an ideal version of d-ACB in the evaluation, where a perfect knowledge of the state information is assumed.

4.5.4.3 Results

In Fig. 4.12, we see how the proposed algorithm performs for different values of the proportionality constant C compared to the baseline algorithms. From Figs. 4.12(Ia-IIIa), we observe that DBCA provides lower average service time that the baseline for most of the arrival distributions. Only in the case of uniform arrivals with low load $N \leq 2000$ UEs, DBCA performs similar to d-ACB with small service times for all algorithms. Overall, service times grow almost linearly with N in high load regime for all arrival distributions

and algorithms, whereas a non-linear behavior is noted for low-to-medium load $N \leq 4000$ UEs for uniform and beta arrivals. Q-ary TRA with high branching factor $Q = 8$ does not provide significant advantage, while low branching $Q = 2$ performs well for uniform arrivals, however, still worse than DBCA for medium-to-high load.

In Figs. 4.12(Ib-IIIb), we observe the relationship between total resource consumption and load (burst size). First of all, we note dominantly linear growth of resource consumption with the load, with less steep slopes for DBCA protocols. As a result, DBCA consumes similar amount of resources as the baselines in low load regimen, and significantly less in high load regimen. Even in the case where $C = 1$, where DBCA is constrained to consume no more resources than d-ACB per contention-round, DBCA still yields lower overall consumption for full burst resolution due to shorter resolution times. This follows since DBCA is capable of obtaining more throughput out of the same consumed resources as d-ACB, thus, wasting less resources to collisions. Interestingly, binary TRA consumes consistently more than other baselines, while 8-ary – consistently less. This is a counter-intuitive observation, however, it is easily explained: While higher branching factor provides sub-optimal throughput, it produces mostly idle preambles, and idle preambles consume significantly less resources than collided. This result is confirmed in the Figs. 4.12(Ic-IIIc), where we see that 8-ary TRA is very resource efficient. Overall, we observe that DBCA with low C performs most efficient reaching efficiency ≥ 0.35 for medium-to-high load.

Another counter-intuitive observation is that the overall resource consumption of DBCA exhibits relatively low variation for the different values of resource constraint CR, especially for higher values of N. However, there is indeed a great difference in the mean service time, which is the lower the higher C is. This is explained by the fact that, if we look into Fig. 4.11, we see that the ratio S_J/R_J is relatively steady along most of the Pareto frontier for high values of n_i, which is the condition where most of the resource consumption takes place. Thus, we can trade consuming higher amounts of resources for a shorter period of time, or lower amounts for a longer period of time, without severely affecting the resulting resource efficiency.

4.6 Summary

This chapter has been dedicated to the analysis and development of methods for enhancing the transient performance of the RAP. We have approached the task by splitting it into per contention round optimization problem.

First, we have defined the resource consumption of the RAP and proposed two methods of incorporating it into the performance optimization: based on efficiency and based on Pareto-optimality. We demonstrated that the latter method is superior, as it allows devising an algorithm with polynomial complexity and it considers both throughput and resource consumption simultaneously. Next, we have proceeded to devise POCA algorithm, which delivers a Pareto-optimal solution to the resource constrained problem of optimizing the throughput. We have shown that algorithm performs better than state

of the art under the equal resource constrained conditions.

Finally, in the last part of the chapter, we have introduced a novel RAP, aided with Binary Countdown Contention Resolution, and analyzed the performance of joint ACB and BCCR operation. We applied the framework of Pareto optimal RAP, and, based on it, proposed Dynamic Binary Countdown - Access barring for fast and efficient M2M burst resolution. DBCA has been benchmarked via an event-based simulation against other state of the art solutions for different burst arrival processes and is shown to achieve up to twice lower average burst resolution delay.

Since BCCR relies on all UEs listening to each other, its performance is at its best in the highly dense networks with spatially correlated burst arrivals, and this is the scenario we have targeted in this chapter. Future work could address the scenarios of partial overhearing, there the UEs are not always close to each other, and access BCCR gains for such scenarios. BCCR could be also useful beyond the burst arrivals, for steady-state RAP performance improvements. Finally, BCCR asymptotically allows to achieve arbitrary low collision probability. This property could be utilized in for Ultra reliable ultra low latency (URLLC) applications [Pop+17], to design RAP with reliability guarantees.

Appendix

4.A Proof of Theorem 2.

Proof. By the definition (4.19), the outcome of an arbitrary RAO j is successful, $x_{i,j} = 1$, if a unique UE chooses the RAO, or if it it wins a contention by choosing the highest priority level. For an arbitrary priority level $\mathbf{p}' = v - 1$ with $v \in \{1, 2, \ldots, l\}$, there are v levels with equal or higher priority. E.g., $\mathbf{p}' \triangleq 0 \Rightarrow v = 1$ (1 higher or equally prioritized level), or $\mathbf{p}' = l_i - 1 \Rightarrow v = l_i$. Consider a UE that has passed the ACB check, has chosen preamble j and has chosen a priority $v - 1$, we obtain its successful BCCR probability:

$$\mathbb{P}\left[x_{i,j} = 1 | \mathbf{p}', \text{ACB passed}\right] = \left(1 - \frac{v}{l_i}\frac{p_i}{M_i}\right)^{n_i - 1}, \tag{4.40}$$

where $\frac{v}{l_i}\frac{p_i}{M_i}$ represents the probability that another UE passes ACB, chooses preamble j and higher or equal priority level. Since the events of choosing any priority levels are a partition of the sample space, we conclude:

$$\begin{aligned}
&\mathbb{P}\left[x_{i,j} = 1 | \text{ACB passed}\right] = \\
&= \sum_{v=1}^{l_i} \mathbb{P}\left[\mathbf{p}' = v - 1\right] \mathbb{P}\left[x_{i,j} = 1 | \mathbf{p}' = v - 1, \text{ACB passed}\right] \\
&= \sum_{v=1}^{l_i} \frac{1}{l_i} \left(1 - \frac{v}{l_i}\frac{p_i}{M_i}\right)^{n_i - 1},
\end{aligned} \tag{4.41}$$

where the summation over $v \in \{1, ..., l_i\}$ considers any possible priority level. By analogy, accounting for the probability of a UE to pass ACB check p_i, choose the preamble j, and that any of n_i could be successful, we obtain a modified expression (4.41):

$$\mathbb{P}[x_{i,j} = 1] = \binom{n_i}{1} \frac{p_i}{l_i M_i} \sum_{v=1}^{l_i} \left(1 - \frac{v}{l_i}\frac{p_i}{M_i}\right)^{n_i - 1}. \tag{4.42}$$

By definition of expectation, we get:

$$\mathbb{E}[x_{i,j}] = \sum_{w=0}^{1} w\, \mathbb{P}[x_{i,j} = w] = \mathbb{P}[x_{i,j} = 1]. \tag{4.43}$$

To obtain (4.20), we recall that $s_i = \sum_{j=1}^{M_i} x_{i,j}$, and use the sum of the expectations rule:

$$S_J = M_i\, \mathbb{E}\left[x_{i,j}\right] = \frac{n_i p_i}{l_i} \sum_{v=1}^{l_i} \left(1 - \frac{v}{l_i}\frac{p_i}{M_i}\right)^{n_i - 1}. \tag{4.44}$$

□

4.B Proof of Lemma 3

First, we prove that the unconstrained problem has only one solution. Consider the objective function as a product of two functions $S_J = f_o(p_i;\bar{k}) \triangleq y_o(p_i)g_o(p_i)$, where $y_o(p_i) \triangleq \frac{\bar{n}}{l_i}p_i$, and $g_o(p_i) \triangleq \sum_{h=1}^{l_i}\left(1-\frac{h}{l_i}\frac{p_i}{M_i}\right)^{\bar{n}-1}$. The first and second order derivatives of these functions are:

$$\frac{dy_o}{dp_i} = \frac{\bar{n}}{l_i}, \quad \frac{d^2y_o}{dp_i^2} = 0, \tag{4.45}$$

$$\frac{dg_o}{dp_i} = -\frac{(\bar{n}-1)}{l_iM_i}\sum_{h=1}^{l_i} h\left(1-\frac{hp_i}{l_iM_i}\right)^{\bar{n}-2}, \tag{4.46}$$

$$\frac{d^2g_o}{dp_i^2} = \frac{(\bar{n}-1)(\bar{n}-2)}{l_i^2M_i^2}\sum_{h=1}^{l_i} h^2\left(1-\frac{hp_i}{l_iM_i}\right)^{\bar{n}-3}. \tag{4.47}$$

Note that since the following holds: $\frac{dg_o}{dp_i} < 0$, $\frac{d^2g_o}{dp_i^2} > 0$, $g_o(p_i)$ is a convex and strictly decreasing function. Now we prove by contradiction that the function $f_o(p_i)$ has a single maximum. Assume that $f_o(p_i)$ has two maximums $p_{i,1}$ and $p_{i,3}$, that implies there is also has a minimum in $p_{i,2}$. Considering that $\frac{d(y_og_o)(p_{i,j})}{dp_i} = 0$, $j \in \{1,2,3\}$, and Eqns. (4.45)–(4.47), we obtain:

$$\frac{dg_o(p_{i,j})}{dp_i} = -\frac{1}{p_{i,j}}g_o(p_{i,j}) \tag{4.48}$$

$$\frac{d^2g_o(p_{i,j})}{dp_i^2} = \frac{g_o(p_{i,j})}{p_{i,j}}\left(\frac{1}{p_{i,j}}-1\right). \tag{4.49}$$

Using Eqns. (4.48), (4.49) we can derive the second derivative of the function $(y_og_o)(p_{i,j})$ as:

$$\frac{d^2(y_og_o)(p_{i,j})}{dp_i^2} = \underbrace{-\frac{\bar{n}}{l_i}}_{<0}g_o(p_{i,j})\underbrace{\left(\frac{1}{p_{i,j}}+1\right)}_{>0}. \tag{4.50}$$

Following our assumption, we have $\frac{d^2(y_og_o)(p_{i,j})}{dp_i^2} < 0$, $j \in \{1,3\}$, and $\frac{d^2(y_og_o)(p_{i,j})}{dp_i^2} > 0$, $j = 2$. This implies that $g_o(p_{i,1}), g_o(p_{i,3}) > 0$, and $g_o(p_{i,2}) < 0$. However, as $g_o(p_i)$ is a decreasing function, and $p_{i,1} < p_{i,2} < p_{i,3}$, we come to a contradiction. Hence, $f_o(p_i)$ has only one maximum, and the unconstrained problem (4.35) has only one solution.

Next, consider the constraint function:

$$R_J\left(\bar{n},\bar{k};p_i\right) = M_ir_I + r_3(1+\bar{k}\delta)\left(M_i - M_i\left(1-\frac{p_i}{M_i}\right)^{\bar{n}_i}\right) \leq \bar{r}. \tag{4.51}$$

We can reformulate it as:

$$p_i \leq p_{\max}^{(J)}, \text{ with } p_{\max}^{(J)} = M_i - M_i \left(1 - \frac{\bar{r} - M_i r_I}{M_i r_3 (1 + \bar{k}\delta)}\right)^{\frac{1}{\bar{n}}}. \tag{4.52}$$

Hence, this constraint is a closed half-plane defined by a constant $p_{\max}^{(J)}$. Clearly, since $f_o(p_i)$ is a function, it has only one interception with (4.52). This implies that the constrained problem (4.29) also has also only one solution.

4.C Proof of Lemma 4

To obtain (4.37), we first apply Tailor series approximation $\left(1 - \frac{hp_i}{l_i M_i}\right)^{n_i - 1} \approx e^{-h\frac{n_i p_i}{M_i l_i}}$ to the objective function given by Eqn. (4.20), then substitute $\phi = \frac{n_i p_i}{M_i l_i}$, and finally simplify the sum as a partial sum of a geometric series, obtaining:

$$S_J = \phi M_i \sum_{h=1}^{l_i} e^{-h\phi} = \phi M_i e^{-\phi} \frac{1 - e^{-\phi l_i}}{1 - e^{-\phi}}. \tag{4.53}$$

Instead of directly optimizing the expression, we apply logarithmic transformation to (4.53), and then obtain its derivative

$$\frac{dS_J}{d\phi} = \frac{d\left(\log \phi M_i + \log e^{-\phi} + \log\left(1 - e^{-\phi l_i}\right) - \log\left(1 - e^{-\phi}\right)\right)}{d\phi}.$$

By simplifying the equation and setting it to 0, we obtain (4.37). Uniqueness of the root follows from Lemma 3.

Chapter 5

From Massive towards Reliable Machine-to-Machine Random Access

In the previous chapters, we targeted massive Machine-to-Machine (mM2M) communications, and optimized the performance in its expectation. In the literature, it is common to characterize mM2M as delay-tolerant applications with relaxed reliability requirements and uniquely large number of end users [Oss+14]. In contrast to it, **ultra reliable Machine-to-Machine (uM2M)** requires low latency communication links and imposes stringent reliability constraints [Pop14]. The distinction between these two classes of applications is sometimes nominal, and the same application might exhibit both massive number of end users and stringent requirements, e.g., consider smart grid applications [Oss+14]. For such applications on the border between mM2M and uM2M [Pop14], e.g., in-cabin communication in an aircraft or large-scale industrial automation [G¨+17b], assessing the average performance is insufficient. For instance, if all the sensors in a factory need to re-connect after an emergency shutdown within a certain time limit [G¨+17b]. In that case, **reliability guarantees for the Random Access CHannel (RACH) performance** are necessary. As a first step towards designing the reliable random access procedures for such scenarios, this chapter aims to answer the question of what the **performance limits** of the existing standardized solutions are.

5.1 Contributions and Structure of the Chapter

In this chapter, we analytically study the probabilistic performance bounds of standardized RACH with Access Class Barring (ACB). We investigate the *burst resolution time*, i.e., the time it takes to connect a burst of Machine-to-Machine (M2M) devices to the base station. Modeling RACH as a queuing system, we approach the analysis by the means of stochastic network calculus [Fid06], which allows, in contrast to conventional queuing theory, to characterize the behavior of the system in probability and not only in expectation [Ciu+14]. We analyze what burst resolution delay can be guaranteed for a given burst size with a certain reliability requirement. We validate the approach using simulations, and illustrate possible applications of the proposed methodology.

The remainder of the chapter is structured as follows. We introduce the problem and relevant concepts in 5.3. The main result, probabilistic reliability analysis of the random

access procedure is presented in 5.4 and numerically verified in 5.5. We conclude the chapter with a summary in 5.6.

The content of the chapter is based on our work published as [Vil+18].

5.2 Related Work

State-of-the-art techniques of overload control and performance improvements, as we have reviewed in detail in Sec. 2.4, have the expected performance (delay, reliability, etc.) as the optimization objective. Similar is the standardized ACB and its derivatives [Dua+16; Jin+17]. Analytically, ACB performance has been also extensively studied in the expectation, i.e., with respect to the *average* burst resolution time and resulting RACH efficiency [WBC15; Che+15; Jin+17; Kos16; Jia+17]. In [WBC15] and the follow-up work [Che+15], the authors devised an analytical framework to assess the expected performance of the standardized ACB and Extended Access Class Barring (EAB) procedures, respectively. Jian *et al.* [Jia+17] have proposed another iterative approach to the ACB analysis, and Koseoglu [Kos16] derived the lower bound on the average random access delay.

We aim at assessing high order statistics of the latency, ideally answering the question of what latency can be achieved with a given reliability level. This problem is far less common in the literature, with only a few recent works. Frameless ALOHA protocol has been analyzed with respect to its latency-reliability performance by Stefanović et al. [SLP17]. Analytical performance assessment of the multi-channel parallel Tree Resolution Algorithms (TRA) has been presented in [GAK17b], alongside with its applications to burst resolution in [GAK17a]. Delay distribution of the single-channel slotted ALOHA protocol has been characterized by the means of z-transform [Tob82], and closed-form expression in [YY03]. In contrast to the reviewed works, we analyze the Random Access Procedure (RAP) with ACB as *multi-channel slotted ALOHA*, and we are interested in the transient performance, i.e., the distribution of the burst resolution time.

5.3 System Model and Preliminaries

5.3.1 System Model

We consider a scenario with a total of N User Equipments (UEs) and one Next Generation Node B (gNB). At time $i < 0$, all UEs are inactive and disconnected from the gNB. An event is occurring at time $i = 0$, triggering all the UEs, and causing them to initiate a connection establishment (random access) procedure towards the gNB. Activation of individual UEs is occurring according to initial arrival process strictly during the time interval $i \in [0, T_a - 1]$, with T_a referred to as the activation time [3GP11].

Upon activation, every UE attempts to connect to the gNB. The connection follows a four step RAP, depicted in Fig. 5.1: (1) A preamble, chosen uniformly random from a set $|\mathcal{M}|= M$, is sent to the gNB in Physical Random Access CHannel (PRACH). (2) The gNB sends a preamble reply for every successfully decoded preamble, containing uplink grants for Radio Resource Control (RRC) connection requests. (3) UE proceeds with sending its connection request containing UE's identity information, on the respective uplink resource. (4) Every correctly decoded connection request is acknowledged by the gNB with a connection reply. If the UEs choose the same preamble in step 1, their connection requests at step 3 are allocated the same uplink resource, which leads to collisions.

Prior to every PRACH attempt i, UEs receive the PRACH location (sub-frame and frequency offset) and contention parameters (number of available preambles M, access probability p_i) from a gNB broadcast. Every UE independently uses access probability as a part of the ACB procedure to decide whether to compete in a given PRACH opportunity i (with probability p_i), or postpone to the next contention round (with probability $1 - p_i$). The access probability could be either static throughout the burst resolution or dynamically adapted for every contention round [Dua+16].

We assume that all four steps occur within contention round, which we define as a multiple integer of the periodicity of the PRACH in the resource grid[1]. The periodicity is determined by the PRACH configuration index, typically ranging from 1 per sub-frame (1 ms) to 1 per frame (10 ms). Random access procedure is modeled as an *M-channel slotted ALOHA protocol*, where a channel corresponds to a PRACH preamble [WBC15]. We further adopt the *collision channel model without capture*, i.e., every preamble m in the contention round i can have one of three states: *idle* (no UE is choosing the preamble), *singleton* (exactly 1 UE), and *collision* ($\geq$ 2 UEs), with a corresponding service $x_{i,m}$ (Random Access Opportunity (RAO)):

$$x_{i,m} = \begin{cases} 1 & \text{if chosen by 1 UE,} \\ 0 & \text{otherwise.} \end{cases} \tag{5.1}$$

Basically, a preamble is serving a UE request with a full channel capacity if no collision occurs and does not provide any service otherwise. Every activated but not yet served UE is denoted as *backlogged*. For further details on the system model and RAP, we refer the reader to Chapter 2.

5.3.2 Problem Statement

Finally, we define the target Quality of Service (QoS) requirement[2] of the system as a tuple $(\bar{b}, \bar{t}, \varepsilon)$. Here, $\bar{b}$ is the maximum tolerated number of unconnected UEs (*target*

[1] See 2.3 and 2.3.1 for more details and illustration.

[2] Typically, QoS is defined per single user/application, and refers to the delay or datarate requirement. In contrast to that, we are analyzing the burst resolution (multiple users), and borrow the term QoS requirement to refer to the backlog and resolution time.

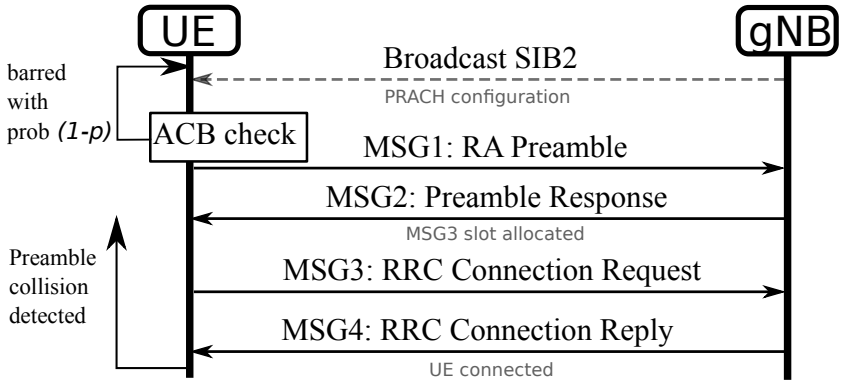

Figure 5.1: Four steps RAP with Access Class Barring.

backlog) by the time $\bar{t}$, which is referred to as the required burst resolution time. The burst is denoted as *resolved* if the number of unconnected UEs is less than or equal than target, $B(\bar{t}) \leq \bar{b}$. The corresponding unreliability ε is the probability that a burst is not resolved by the time $\bar{t}$. The case with $\bar{b} = 0$ corresponds to the full burst resolution, and $\bar{b} > 0$ to the partial burst resolution [PFP04].

The problem we are targeting with the analysis is quantifying how well can the ACB-based random access procedure support a given QoS requirement, i.e., for a given target $\bar{b}$, we would like to compute a bound on the probability ε that a given burst is not resolved within $\bar{t}$ contention rounds.

5.3.3 Analysis Preliminaries

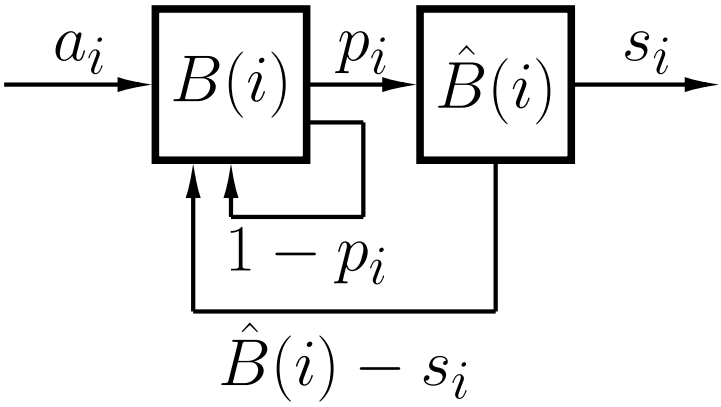

Figure 5.2: Queuing model of an arbitrary i^{th} contention round of RAP. Backlog is evolving according to Lindley's recursion (5.2), with service process determined by ACB with access probability p_i and the collision probability.

Consider the system at an arbitrary contention round $i \geq 0$ with $B(i)$ backlogged UEs. The evolution of the backlog is described by the following recursion:

$$B(i+1) = \max\{0, B(i) + a_i - s_i\}, \tag{5.2}$$

where a_i denotes newly activated UEs, and $s_i = \sum_{m=1}^{M} x_{i,m}$ denotes the total amount of served UEs.

The probability that $s_i = k$ UEs are served during the contention round i, i.e., have successfully connected to the gNB, depends on current backlog $B(i)$ and access probability p_i, limiting the number of admitted for contention UEs $\hat{B}(i)$. Let us first consider that the number of admitted UEs is $\hat{B}(i) = \hat{n}$. In that case, the probability that k out of $\hat{n}$ UEs successfully transmit is [WBC15; Dua+16]:

$$\mathbb{P}\left[s_i = k|\hat{B}(i) = \hat{n}\right] = \binom{\hat{n}}{k}\binom{M}{k}\frac{k!}{M^{\hat{n}}}\sum_{j=1}^{j_{\max}}(-1)^j\binom{M-k}{j}\binom{\hat{n}-k}{j}j!\,(M-k-j)^{\hat{n}-k-j},$$

where $j_{\max} = \min(M-k, \hat{n}-k)$. The number of admitted UEs $\hat{B}(i)$ is binomially distributed with $B(i)$ trials and per trial success probability p_i:

$$\mathbb{P}\left[\hat{B}(i) = \hat{n}|B(i) = n\right] = \binom{n}{\hat{n}}(1-p_i)^{\hat{n}}p_i^{n-\hat{n}}. \tag{5.3}$$

Combining these two equations, we obtain the probability $\mathbb{P}_{k,n} = \mathbb{P}[s_i = k|B(i) = n]$ that k out of n backlogged UEs are successful as:

$$\mathbb{P}_{k,n} = \sum_{\hat{n}=0}^{n}\mathbb{P}\left[\hat{B}(i) = \hat{n}|B(i) = n\right]\mathbb{P}\left[s_i = k|\hat{B}(i) = \hat{n}\right]. \tag{5.4}$$

Eqn. (5.4) already allows a straightforward recursive computation of the burst resolution time. If we consider a state of the system at time i as a tuple $(B(i), a_i)$, where $B(i), a_i \in [0, N]$, then the distribution of the random variable $B(i)$ representing backlog at time i can be computed iteratively starting with contention round 0, using the recursion (5.2). However, this iterative computation requires computing transition matrix from $(N+1)\times(N+1)$ to another $(N+1)\times(N+1)$ dimension state space every time step, and, hence, the complexity is proportional to $(N+1)^2 \times (N+1)^2 \times \bar{t}$. Such computation is only feasible for low total number of UEs N. For large bursts, a different approach is necessary. This motivates an alternative analysis based on the network calculus based analysis, which we present in Sec. 5.4.

5.3.4 Dynamic Access Barring

For large burst arrivals, keeping access probability static is very inefficient. If the probability is too small, burst resolution lasts long due to the medium under-utilization as the backlog decreases. If the access probability is too large, the burst resolution might take even longer due to high preamble collision rates. To optimize the burst resolution times, several works have proposed a dynamic adaptation of the access probability [Dua+16; Jin+17] based on the pseudo-Bayesian broadcast [Riv87]. Consider expected number of

successful UEs in a single contention round as a function of $\hat{B}(i), M$ [WBC15]:

$$\mathbb{E}\left[s_i\right] = \mathbb{E}\left[\sum_{m\in\mathcal{M}} s_{i,m}\right] = \mathbb{E}[\hat{B}(i)]\left(1-\frac{1}{M}\right)^{\mathbb{E}[\hat{B}(i)]-1}. \tag{5.5}$$

It is possible to show that the expectation $\mathbb{E}\left[s_i\right]$ in (5.5) is maximized if the expected number of UEs admitted to contend in a given contention round $\mathbb{E}[\hat{B}(i)] = p_i B(i)$ is equal to the number of preambles M. Hence, the dynamic access barring policy is devised as:

$$p_i^\star \triangleq \arg\max_{p\in(0,1]} \mathbb{E}\left[\sum_{m\in\mathcal{M}} x_{i,m}\right] = \min\left\{1, \frac{M}{B(i)}\right\}. \tag{5.6}$$

We denote $p_i^\star$ defined by (5.6) as *optimal barring policy*.

Remark 9. *In general, the number of UEs contending in a given round $B(i)$ is unknown. However, there exist a number of backlog estimation techniques, producing accurate results [Dua+16; PFP04; Jin+17; LCW16; Zan12]. We study the impact of estimation numerically in Sec. 5.5.*

5.4 Stochastic Performance Bounds Analysis for Burst Resolution Time

In this section, we present the burst resolution time analysis. To introduce the reader to stochastic network calculus, we first provide a brief overview in 5.4.1. Then, we define the queuing model of LTE RACH in 5.4.2, and use it to analyse static (in 5.4.3) and dynamic (in 5.4.4, 5.4.5) ACB policies.

5.4.1 Transient Analysis using Network Calculus

Assuming a fluid-flow, discrete-time queuing system, and given a time interval $[s,t)$, $0 \leq s \leq t$, we define the non-decreasing (in t) bivariate processes $A(s,t), D(s,t)$ and $S(s,t)$ as the cumulative arrival to, departure from, and service offered by the system. We further assume that A, D and S are stationary non-negative random processes with $A(t,t) = D(t,t) = S(t,t) = 0$ for all $t \geq 0$. The cumulative arrival and service processes are given in terms of a_i and s_i as follows

$$A(s,t) = \sum_{i=s}^{t-1} a_i \quad \text{and} \quad S(s,t) = \sum_{i=s}^{t-1} s_i\,, \tag{5.7}$$

for all $0 \leq s \leq t$. We denote by $B(t)$ the backlog (the amount of buffered data) at time t.

Based on this server model, the total backlog can be studied analytically. For a given queuing system with cumulative arrival $A(0,t)$ and departure $D(0,t)$ and for $t \geq 0$, the backlog at time t, $B(t)$ is defined as the amount of traffic remaining in the system by time t. Therefore,

$$B(t) \triangleq A(0,t) - D(0,t)\,. \tag{5.8}$$

While deterministic network calculus [LBT01] can provide worst-case upper bounds on the backlog and the delay if traffic envelopes (an upper bound on the arrival process) as well as a service curve (a lower bound on the service process) are considered, probabilistic performance bounds provide more useful and realistic description of the system performance than deterministic analysis for corresponding systems. Stochastic network calculus has been previously applied to protocol analysis in the context of 802.11 Distributed Coordination Function (DCF) and slotted ALOHA networks [Ciu+14; PC15].

In the probabilistic setting (where the arrival process A and the service process S are stationary random processes), the backlog defined in (5.8) is reformulated in a stochastic sense:

$$\mathbb{P}[B(t) > \bar{b}] \leq \varepsilon, \tag{5.9}$$

where $\bar{b}$ denotes the target probabilistic backlog associated with violation probability ε. This performance bound can be obtained by the distributions of the processes, i.e., in terms of Moment Generating Functions (MGFs) of the arrival and service processes [Fid10]. In general, the MGF-based bounds are obtained by applying Chernoff's bound, that is, given a random variable X, we have

$$\mathbb{P}[X \geq x] \leq e^{-\theta x}\mathbb{E}\left[e^{\theta X}\right] = e^{-\theta x}\mathbb{M}_X(\theta),$$

whenever the expectation exists, where $\mathbb{E}[Y]$ and $\mathbb{M}_Y(\theta)$ denote the expectation and the MGF (or the Laplace transform) of Y, respectively, and θ is an arbitrary non-negative free parameter. Given the stochastic process $X(s,t)$, $t \geq s$, we define the MGF of X for any $\theta \geq 0$ as [Fid06]

$$\mathbb{M}_X(\theta, s, t) \triangleq \mathbb{E}\left[e^{\theta X(s,t)}\right].$$

In a similar way, we define $\overline{\mathbb{M}}_X(\theta, s, t) \triangleq \mathbb{M}_X(-\theta, s, t) = \mathbb{E}\left[e^{-\theta X(s,t)}\right]$.

A number of properties of MGF-based network calculus are summarized in [Fid06]. In this work, we consider a queuing system with an initial backlog for which we are interested in the transient behavior of the backlog itself. In general, the probabilistic backlog bound $\bar{b}(t)$ for a given violation probability ε at time t can be expressed by [Fid06; AZLB13]

$$\bar{b}(t) = \inf_{\theta > 0}\left\{\frac{1}{\theta}\left(\log \mathsf{M}(\theta, t, t) - \log \varepsilon\right)\right\}, \tag{5.10}$$

where $\mathsf{M}(\theta, u, v)$ is given as

$$\mathsf{M}(\theta, u, v) \triangleq \sum_{k=0}^{\min(u,v)} \mathbb{M}_A(\theta, k, v) \cdot \overline{\mathbb{M}}_S(\theta, k, u). \tag{5.11}$$

The consideration of the initial backlog in the system can be finally represented by the choice of an appropriate arrival function, as we discuss in the next section.

5.4.2 Queuing Model of Random Access Procedure

RAP could be viewed as a queuing system, where the incoming UEs are considered as *arrivals* into the queue, and UEs, successfully completing the procedure, as *departures* from the queue. The serving process of such a system is a stochastic process, dependent on the current backlog size, on the advertised access probability p_i, and on the number of available preambles M, as in Eqn. (5.4).

5.4.2.1 Arrival Process

Third Generation Partnership Project (3GPP) offers three burst arrival models [3GP11]: delta (simultaneous, "spike") arrivals with total activation time $T_a = 0$, uniform distribution of arrivals within $[0, T_a - 1]$, or beta distribution $\mathsf{B}(\alpha, \beta)$ within $[0, T_a - 1]$. In this work, we consider only the worst-case scenario with simultaneous activation of all UEs. In that case, the distribution of the activation time t_a of individual UEs and the resulting cumulative arrival process are expressed as:

$$\mathbb{P}[t_a = 0] = 1 \quad \text{and} \quad A(\tau, t) = \begin{cases} N & \tau = 0, \\ 0 & \tau > 0. \end{cases} \tag{5.12}$$

5.4.2.2 Serving Process

Every preamble $m \in \mathcal{M}$ can be considered a server, with the service as defined in (5.1). Hence, cumulative serving process can be expressed as:

$$S(\tau, t) \triangleq \sum_{i=\tau}^{t-1} \sum_{m \in \mathcal{M}} x_{i,m}, \tag{5.13}$$

$$\text{with } s_i \triangleq \sum_{m \in \mathcal{M}} x_{i,m} \sim f_i(k), \tag{5.14}$$

where the probability mass function (PMF) of the service process at the time step i

as $f_i(\cdot)$:

$$\begin{aligned} f_i(k) &= \sum_{j=0}^{N} \mathbb{P}_{k,j}\, \mathbb{P}[B(i-1)=j] \\ &= \sum_{j=0}^{N} \mathbb{P}_{k,j} \sum_{l=0}^{N} \mathbb{P}_{l-j,l}\, \mathbb{P}[B(i-2)=l] \\ &= \sum_{j=0}^{N} \mathbb{P}_{k,j} \sum_{l=0}^{N} \mathbb{P}_{l-j,l} \sum_{m=0}^{N} \mathbb{P}_{m-l,m}\, \mathbb{P}[B(i-3)=m] \\ &= \sum_{j=0}^{N} \mathbb{P}_{k,j} \ldots \sum_{y=0}^{N} \mathbb{P}_{y-x,y}\, \mathbb{P}[B(0)=y] \\ &= \underbrace{\sum_{j=0}^{N} \mathbb{P}_{k,j} \ldots \sum_{y=0}^{N} \mathbb{P}_{y-x,y}}_{i \text{ sums}}\, \mathbb{P}_{N-y,N} \end{aligned} \tag{5.15}$$

Similarly, we derive the distribution of the cumulative service $S(0,i)$, f_{S_i} as

$$\begin{aligned} f_{S_i}(k) &= \mathbb{P}[S(0,i)=k] = \\ &= \sum_{j=\max(0,k-M)}^{k} \mathbb{P}_{k-j,N-j}\, \mathbb{P}[S(0,i-1)=j] = \\ &= \underbrace{\sum_{j=\max(0,k-M)}^{k} \mathbb{P}_{k-j,N-j} \cdots \sum_{w=\max(0,y-M)}^{y} \mathbb{P}_{y-w,N-w}}_{i}\, \mathbb{P}_{w,N}\,. \end{aligned} \tag{5.16}$$

5.4.3 Static Access Barring

Static access barring implies that the access probability p_i does not change over the burst resolution time. While it is inefficient in practice, static barring policy could serve as a baseline for the performance evaluation.

Given the QoS requirement tuple $(\bar{b}, \bar{t}, \varepsilon)$, we are interested in the probability that the burst of size N is still unresolved by the time $\bar{t}$, and what is the backlog remaining at time $\bar{t}$. So, we are looking for the bound on the backlog at the time $\bar{t}$. First, we derive the MGF of the arrival and service process:

$$\mathbb{M}_A(\theta,0,t) = \mathbb{E}\left[e^{\theta A(0,t)}\right] = e^{\theta N}. \tag{5.17}$$

$$\overline{\mathbb{M}}_S(\theta,0,t) = \mathbb{E}\left[e^{-\theta S(0,t)}\right] = \sum_{k=0}^{Mt} e^{-\theta k}\, \mathbb{P}[S(0,t)=k] = \sum_{k=0}^{Mt} e^{-\theta k} f_{S_t}(k). \tag{5.18}$$

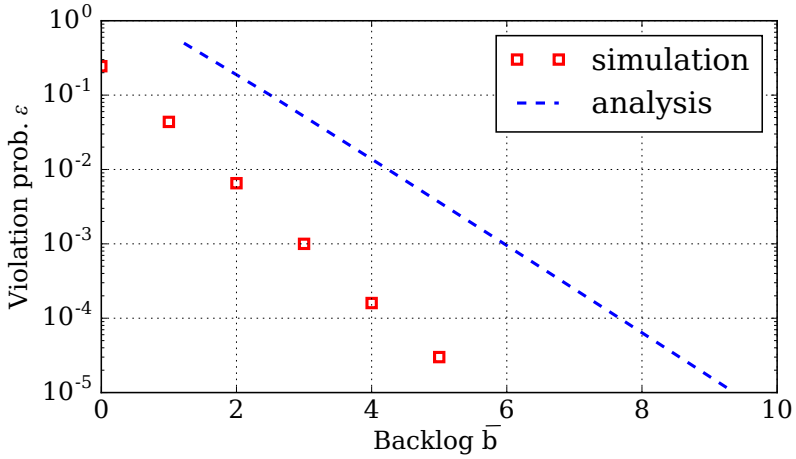

Figure 5.1: Static Access Class Barring: target backlog $\bar{b}$ vs. violation probability ε. Parameters: $M = 30$, $N = 100$, $p = 0.5$.

Now, considering the time horizon $\bar{t} \in [0, \bar{t})$ as a single contention round and substituting Eqns. (5.17) and (5.18) in (5.10), we can numerically compute the bound $\bar{b}(t)$ for a given violation probability ε and resolution time $t = \bar{t}$:

$$\bar{b}(\bar{t}) = \inf_{\theta} \left\{ \frac{1}{\theta} \left(\log(e^{\theta N} \sum_{k=0}^{M\bar{t}} e^{-\theta k} f_{S_{\bar{t}}}(k)) - \log \varepsilon \right) \right\}. \tag{5.19}$$

Alternatively, we can compute the probability ε of violating a backlog bound $\bar{b}$ at a given time $\bar{t}$:

$$\varepsilon = \inf_{\theta} \left\{ e^{-b^{\varepsilon}\theta} \sum_{\tau=0}^{\bar{t}} \mathbb{M}_A(\theta, \tau, \bar{t}) \overline{\mathbb{M}}_S(\theta, \tau, \bar{t}) \right\}. \tag{5.20}$$

We plot a simple numerical example for the static ACB with $p_i = p = 0.5\,\forall i$ in Fig. 5.1. On the x-axis, the backlog bound is plotted, and on the y-axis, the corresponding violation probability for a fixed delay of $\bar{t} = 10$ contention rounds. We observe that the analytical bounds hold, and they are conservative with respect to the simulation with $1 - 2$ orders of magnitude difference.

In the general case of arbitrary static barring factor, computation of performance bounds via MGF calculus requires computing the cumulative service process $f_{S_t}(k)$ via Eqn. (5.16). Hence, computing the cumulative service according to Eqn. (5.16) and then bounding the backlog distribution through network calculus has the same complexity as computing the actual backlog distribution as in 5.3.3. For large N, this becomes computationally infeasible, which motivates finding approximations for the service process, as we do in the remainder of the section.

5.4.4 Dynamic Access Barring

From a practical point of view, dynamic access barring presents a more interesting subject for the analysis, as it maximizes the expected efficiency of the random access procedure. Optimal dynamic barring policy as defined by Eqn. (5.6), is adjusting the access probability in order to maximize the expected number of successful outcomes. In other words, as the expected number of accepted UEs is a function of both access probability, and the backlog, with $\mathbb{E}\left[\hat{B}(i)\right] = p_i B(i)$, dynamic access barring attempts to keep $\mathbb{E}[\hat{B}(i)]$ independent of the backlog.

This leads to an interesting observation about the expected service. The burst resolution time has now two distinct regions: first, where $B(i) \geq M$, for which it holds $p_i = \frac{M}{B(i)}$ and second, where $B(i) < M$ and $p_i = 1$. In any practically relevant case, the first region is dominating the total burst resolution time, since $N \gg M$. Also, for partial burst resolution time, where the target allowed number of non-activated UEs $\bar{b} \geq M$, only the first region is of interest. Here, we first consider the partial burst resolution time with $\bar{b} \geq M$, and then generalize it to full burst resolution in the next subsection.

Given the optimal barring policy, number of UEs admitted to attempt the random access in a given round $\hat{B}(i)$ becomes a binomial random variable:

$$\mathbb{P}[\hat{B}(i) = \hat{n}|B(i)] = \binom{B(i)}{\hat{n}} (p_i^\star)^{\hat{n}} (1 - p_i^\star)^{B(i)-\hat{n}} \tag{5.21}$$

with $\mathbb{E}\left[\hat{B}(i)\right] = p_i^\star B(i) = M$.

The number of admitted UEs and, hence, the service are still dependent on the backlog. To make them independent, we apply Poisson limit theorem to approximate binomial distribution (5.21) by the Poisson with the same mean:

$$\mathbb{P}[\hat{B}(i) = \hat{n}|B(i)] \approx \mathbb{P}[\hat{B}(i) = \hat{n}] = \frac{M^{\hat{n}} e^{-M}}{\hat{n}!}. \tag{5.22}$$

In general, the approximation of (n, p) binomial distribution with a Poisson distribution with mean $\lambda = np$ leads to an underestimation of the probability of getting a value close to the mean, and overestimating the probability of being far from the mean value. For our case, it means that we are actually underestimating the resulting service, hence, we are more conservative. For small n and $p \to 1$, the approximation might become even too conservative.

More importantly, using Eqn. (5.22), and the fact that the approximated service process is independent of the current backlog state $B(i)$, we can express the service MGF via the MGF of the service in a single round $\overline{\mathbb{M}}_S(\theta)$ as:

$$\overline{\mathbb{M}}_S(\theta)^{t-\tau} \triangleq \overline{\mathbb{M}}_S(\theta, \tau, t) = \left(\sum_{k=0}^{M} \sum_{\hat{n}=0}^{N} \frac{M^{\hat{n}} e^{-M}}{\hat{n}!} \mathbb{P}\left[s_i = k | \hat{B}(i) = \hat{n} \right] e^{-\theta k} \right)^{t-\tau}. \tag{5.23}$$

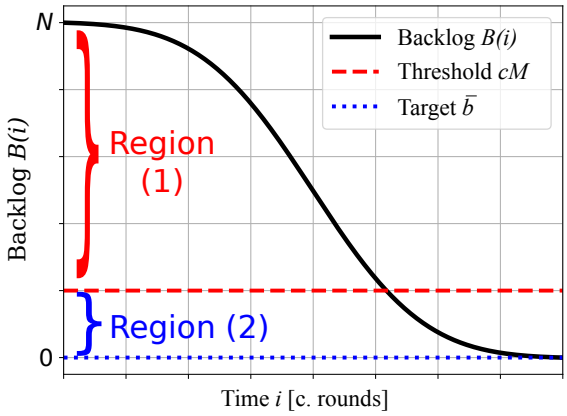

Figure 5.2: Illustration of the full burst resolution. The computation is split into two regions, where backlog is decreasing (1) from $N \geq B(i) \geq cM$, and (2) from $cM > B(i) \geq 0$. $c > 1$ is the parameter to control the conservativeness vs. computation trade-off.

Continuing, we simplify Eqn. (5.20) by using MGF of a service increment:

$$\varepsilon(\bar{b}, \bar{t}) = \inf_{\theta} \left\{ e^{-\bar{b}\theta} \left(e^{\theta N} \overline{\mathbb{M}}_S(\theta)^{\bar{t}} + \overline{\mathbb{M}}_S(\theta) \frac{1 - \overline{\mathbb{M}}_S(\theta)^{\bar{t}-1}}{1 - \overline{\mathbb{M}}_S(\theta)} \right) \right\}. \tag{5.24}$$

Eqns. (5.23) and (5.24) allow us to compute the violation probability for a partial burst resolution with $\bar{b} \geq M$. However, approximation (5.22) becomes very conservative as $\bar{b} \to M$, hence, to provide tighter bounds we need to restrict the use of the approximation to $\bar{b} > M$, and compute the remaining part of the burst resolution iteratively, which we show in the following section.

5.4.5 Full Burst Resolution

First, to control the conservativeness of the bound, we introduce a parameter $c > 1$, such that we refine the split of the total burst resolution time into two regions: (1) $B(i) \geq cM$ and (2) $cM > B(i) \geq 0$. The split is illustrated in Fig. 5.2. Consider the following two random variables: t_1 and t_2, time for the backlog to be reduced from N to cM (region 1), and from cM to 0 (region 2), respectively.

We are interested in the probability that the sum of these partial resolution times, $t_1 + t_2$, is larger than the time of interest $\bar{t}$:

$$\begin{aligned} \mathbb{P}[t_1 + t_2 \geq \bar{t}] &= \sum_{\tau=0}^{\bar{t}} \mathbb{P}[t_2 = \tau] \, \mathbb{P}[t_1 \geq \bar{t} - \tau] \\ &\leq \sum_{\tau=0}^{\bar{t}} \mathbb{P}[t_2 = \tau] \varepsilon(cM, \bar{t} - \tau). \end{aligned} \tag{5.25}$$

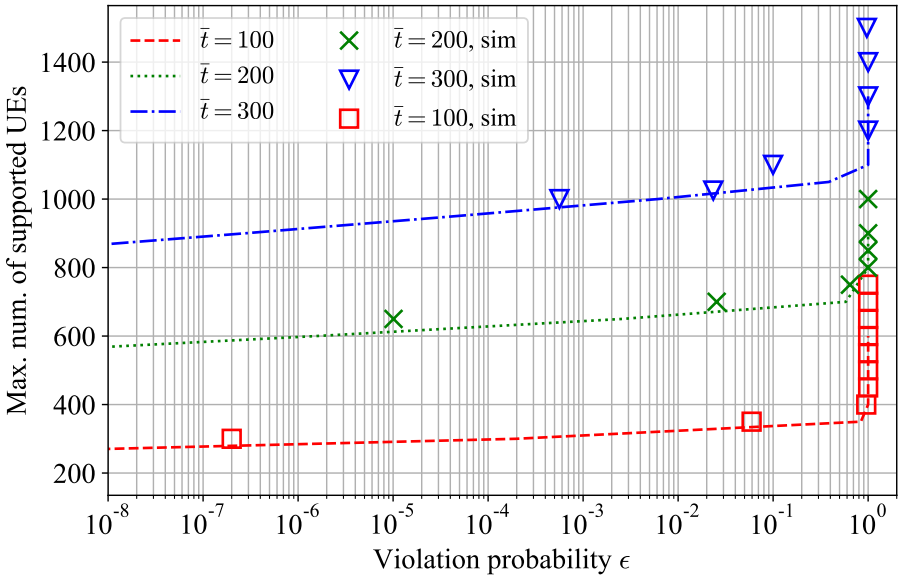

Figure 5.1: Maximum number of supported UEs vs. QoS requirement (total burst resolution $\bar{t}$ violation probability); analysis and simulation; Parameters: $M = 10$, QoS requirement $\bar{t} = \{100, 200, 300\}$, backlog bound $\bar{b} = 0$, simulation with 10^7 samples.

Computing the violation probability for region 1 and t_2 is possible using the previously introduced Eqn. (5.24). Computing the resolution time for the region II can be done either using the methods for static ACB as in 5.4.3, or even directly by iteration using $\mathbb{P}_{k,n}$ as in Eqn. (5.4) and the framework introduced in Sec. 5.3.3. Since $cM \ll N$, the computational complexity is low and proportional to $cM \times \bar{t}$. Parameter $c > 1$ is used to trade off conservativeness and computational complexity.

5.5 Numerical Results

In this section, we provide the numerical performance evaluation and compare the analytical results with the Monte-Carlo simulation based on the Omnet++ [Var+01] framework. We show the results for a simultaneous activation process, where all N nodes are activated at the same time $i = 0$.

We first demonstrate a possible use case of the proposed model for system dimensioning. For a fixed target QoS requirement, $(\bar{b}, \bar{t}, \varepsilon)$, we analytically determine the bound on the maximum number of UEs which could be supported for the requirement. The use case is illustrated in Fig. 5.1, for the full burst resolution requirement $\bar{b}$ and different resolution times $\bar{t} = \{100, 200, 300\}$. We observe that the analytical approach provides tight lower bound for the simulations' results.

To further validate the analytical bounds, Fig. 5.2 depicts the violation probability for

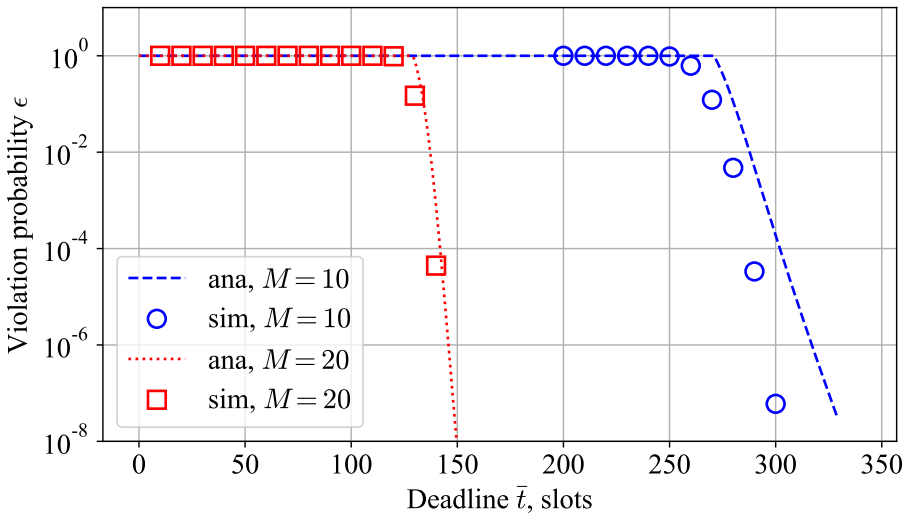

Figure 5.2: Partial burst resolution, backlog bound $\bar{b} = 3M$; minimum violation probability: analysis vs. simulation. Parameters: number of preambles $M \in \{10, 20\}$, number of UEs $N = 1000$. Simulations for 10^8 samples.

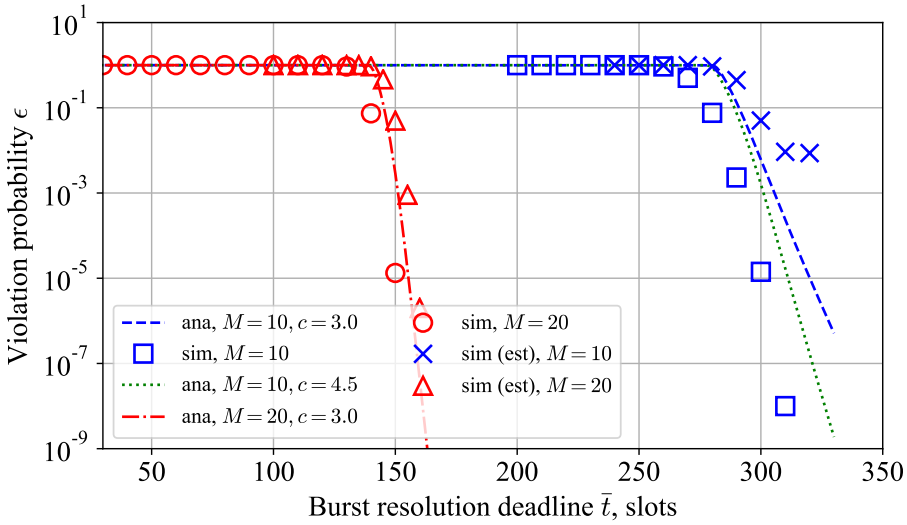

Figure 5.3: Full burst resolution, backlog bound $\hat{b} = 0$. Minimum violation probability: analysis vs. simulation. Simulations were performed with exact backlog knowledge, and with backlog estimation (denoted "sim (est)"). Parameters: $M \in \{10, 20\}$, $c \in \{3, 4.5\}$, $N = 1000$. Simulations for 10^8 samples.

the burst resolution time t for the case of partial burst resolution with the bound $\bar{b} = 3M$ for varying M. We observe that the model provides a conservative bound on violation probability for both cases, and the conservativeness increases with t and decreases with M. For the larger number of preambles $M = 20$, we observe that the slope of the CCDF

is steeper than for $M = 10$, indicating lower variance of the burst resolution times.

Next, Fig. 5.3 illustrates the violation probability dependency on the resolution time for a full resolution scenario. The analytical results merge two computational models using the Eqn. (5.25), at different "splitting points" points cM, with corresponding $c \in \{3.0, 4.5\}$. As expected, increasing c makes the overall model less conservative since the Poisson approximation of the binomial process in (5.22) becomes less conservative, at the expense of slightly longer computation.

5.5.1 Impact of the Backlog Estimation

In some controlled scenarios, where the total number of UEs N and the activation pattern are known to the gNB, deducing the current backlog state $B(i)$ in any time contention round i is possible. However, as we mentioned earlier, in many practical scenarios, backlog remains unknown to the gNB, and, instead, techniques for estimating it have to be used [Dua+16; PFP04; Jin+17; LCW16; Zan12]. To evaluate the impact of estimation, we relax here the assumption about the backlog state knowledge, and simulate same scenarios with the pseudo-bayesian estimation[3] as proposed by Jin *et al.* [Jin+17]. In short, this estimation relies on the maximum-likelihood guess about the backlog $B(i)$ based on the observation of the number of idle and collided preambles in a given round. The guess is adjusted with every new contention round.

The simulation results are also plotted in Fig. 5.3. Comparing to the full state information case, estimation decreases the violation probability by up to almost two orders of magnitude (case $M = 20$), and up to more than three orders of magnitude for the case $M = 10$. The impact of estimation is higher in the second case, because the estimation relies on the observation of the number of idle preambles. When the total number of preambles is low, the estimation becomes inaccurate. Furthermore, we observe that the analytical results do not provide a bound for the case with the estimation, although they correctly capture the slope and are close to the simulation results for $M = 20$. To provide an accurate performance bound for this case, future work should characterize the estimation error, and respectively include it as an offset in the serving process.

5.6 Summary and Discussion

In this chapter, we have proposed a methodology for analyzing the reliability of the Random Access Procedure with Access Class Barring. We have considered a burst arrival scenario, where N UEs are simultaneously trying to connect to the gNB. For a given maximum allowed number of unconnected UEs b^{ε} (target backlog), and resolution time t, we have computed the maximum violation probability ε. For dynamic access barring, we have shown that the partial burst resolution time with target backlog $b^{\varepsilon} > M$, where M

[3] As there is no comparison of the estimation techniques in the current state of the art, we are using here the technique which has performed best in our simulated scenarios.

is the number of available preambles, can be computed by using solely the stochastic network calculus tools. For computing full burst resolution time, we have combined iterative computation and stochastic network calculus to achieve accurate results. The presented analysis can be used for assessing the RAP performance, and integrating random access protocols into the end-to-end system reliability framework. It can also be used in standalone scenarios for system dimensioning, e.g., to decide the maximum number of UEs which could be supported for a given resolution time and reliability requirement.

Finally, as we illustrate numerically, imperfect backlog estimation has significant negative impact on the performance. This motivates further work in incorporating estimation techniques into the random access reliability analysis, as well as developing estimation techniques which can provide reliability guarantees. Additionally, future work in assessing worst-case performance of non-barring based techniques, e.g., tree algorithms [G¨+17b; PFP04; LAAZ14], and extensions of our framework for assessing other burst arrival patterns (Beta or uniform [3GP11]) are necessary.

Chapter 6

Random Access Protocols for Networked Control Systems

In the previous chapters, we have developed methodologies for enhancing performance of the communication protocols in the presence of generic Machine-to-Machine (M2M) applications. The M2M traffic models which we have used are not that of a particular application, but rather representative models of M2M applications in general [3GP11]. The advantage of using the models is their tractability for the analysis. The drawback is, however, that individual applications do not necessary comply to the standard models [PF95; Sha+12; Lan+15], and their precise description becomes increasingly complicated if more application parameters, e.g., mobility patterns or geographic distribution [Gri+17; TMF17; Lan+13], are considered. There is a large number of relevant works dedicated to detailed model- and measurements-based description of M2M applications, e.g., [Gri+17; Lan+15; TMF17; Lan+13; Sha+12; KK13]. A counterpart of the model-based approach are case studies, where a specific application is taken as an example [Li+16; Vil+17a]. Case studies are hardly generalizable, but often provide deeper practical insights than model-based approaches.

In this chapter, we take an intermediate approach between a case study and model-based applications. That is, we direct our attention to a specific class of M2M applications, Networked Control Systems (NCSs). NCS is a system with one or multiple control loops sharing the same communication network. In industrial automation, smart grids, or vehicular communications scenarios, the majority of applications have an underlying control process, therefore, NCS is an important subset of M2M application, both for massive Machine-to-Machine (mM2M) and ultra reliable Machine-to-Machine (uM2M).[1]

The research questions we address in this chapter are essentially part of a cross-layer design problem [GNT+06; LSS06]. We model individual control loops of an NCS as a Linear Time-Invariant (LTI) systems and study the performance of NCS in the presence of shared communication links under different Medium Access Control (MAC) protocols. This approach allows the results to be generalizable, since LTI models are applicable to the majority of control applications. On the same time, it allows us to introduce NCS specific performance metrics such as network-induced error and to address the topic of stability of an NCS under the influence of the MAC protocols.

[1] We refer the reader back to Chapter 2 for a high-level overview of NCS.

6.1 Contributions and Structure of the Chapter

The contributions of the present chapter are split into two main sections.

In the first part, Sec. 6.3, we analyze the behavior an event-triggered NCS consisting of multiple LTI control sub-systems, and sharing the medium using multi-channel slotted ALOHA protocol. First, we describe the local decentralized threshold-based scheduler which determines whether a sub-system is eligible for a transmission attempt. Then, stability of the resulting NCS over the multi-channel slotted ALOHA is discussed in terms of Lyapunov Stability in Probability (LSP). We evaluate the performance of the scheduler and illustrate that there exists an optimal threshold policy minimizing the network-induced error. Based on this observation, we further propose an improved adaptive scheduler. In the new scheduler design, network and control systems are coupled via the knowledge of the network state: Each local scheduler adapts its threshold based on the available network resources. Numerically, we demonstrate that an adaptive choice of the transmission threshold is beneficial compared to the non-adaptive static design.

In the second part, Sec. 6.4, we extend the first part by introducing an approach to dynamically prioritize channel access among multiple sub-systems, employing a binary countdown technique [VRK17; Geh+14], also see Chapter 4. In the proposed approach, priority of the system is determined dynamically based the plant state. Stability of NCSs under the proposed scheme is addressed employing Lyapunov-based concepts of stochastic stability. In addition, numerical analyses illustrate a considerable performance improvement compared to the state-of-the-art decentralized and centralized techniques. It is demonstrated that the proposed scheme can be deployed more efficiently by significantly lowering the collision rate in case of large number of systems utilizing the communication network.

The content of this chapter is based on our previously published works as [Vil+16a; Mam+17], in collaboration with M. H. Mamduhi *et al.* In particular, stability analysis technique used to show stability of the NCSs in 6.3.3 and 6.4.3 has been developed by the co-authors M. H. Mamduhi and S. Hirche.

6.2 Related Work

NCSs is an inherently multidisciplinary research area, with the main contributors from control and communications communities [GC10]. Stability, stabilizability, and performance of control loops are typically studied under resource or energy constraints, packet loss and latency resulting from the underlying communications [WY01].

6.2.1 Event-Triggered NCSs

An important line of works, which we continue in this chapter, have recently proposed event-triggered control and scheduling schemes, in order to utilize the limited communi-

cation and energy resources efficiently [HSVDB08; Tab07; DJ09; TF13; MH13; MH14a]. These aforecited works suggest that it is usually beneficial to transmit the sampled data upon the occurrence of certain events rather than at periodic time instants. This is even more so in case of large-scale networked control systems due to the sheer amount of data that needs to be exchanged. While time-triggered access schemes usually offer lower complexity, event-based rules excel in efficient resource allocation and robustness especially if the communication resources are limited. In the event-based paradigm, events are typically triggered by either deterministic [NT04; WYB02] or stochastic policies [RSJ12; Don+12; TN08; MMH14]. Deterministic event-based policies award the access to the channel to the entity with the highest priority. Try-Once-Discard (TOD) is a basic event-based deterministic protocol that awards the medium access to the system with the largest estimation error and consequently discards the other transmission requests [WYB02]. However, TOD is prone to system noise and can cope with collisions only with a given pre-defined priority order, and hence is not convenient for practical realizations [Chr+14]. Therefore, an efficient event-based policy for dealing with collisions is still an open research topic.

Due to the non-deterministic transmission patterns of event-based control systems and typically long idle periods between consecutive transmissions, it is not possible to reserve radio resources for event-based control applications. Thus, it makes them prone to the notorious problem of existing wireless standards, namely, congestion during the connection establishment phase [HHN13; LAAZ14; Vil+17b; G¨+17b]. As we describe in the previous chapters, the problem has been extensively studied in the context of LTE Random Access Procedure (RAP), where it is commonly modeled as a multi-channel slotted ALOHA system [WBC15]. As we reviewed earlier in Chapter 2, many results exist which propose improvements for the LTE RAP for general class of M2M devices, however, significantly less contributions can be found in coupling the control system properties and efficient network resource allocation. In [BA11a; BA11b] authors compare the event-based and periodic control via single-channel ALOHA for a network of homogeneous integrator sub-systems. Additionally, Cervin *et al.* [CH08] compare different MAC strategies for event-based NCS, however, their assumption about negligible collision resolution time is diminishing the effect of collisions, which is non-negligible for most of the scenarios. In [MH14b], the authors investigate an adaptive price-based scheduling mechanism for multiple loop NCSs with shared communication resource. In their approach, distributed optimization method and adaptive Markov decision process are employed to develop distributed self-regulating event-triggers which are capable of adapting their transmission request rate in order to fulfill a global resource constraint.

6.2.2 State-Aware Communication for NCS

There has been recently an increased attention in joint design of control and communication schemes in NCSs wherein some real-time state informations of the control loop are taken into account in order to increase control performance and communication quality. State-dependent schemes facilitate performance-oriented design such that real-time

behavior of the NCS is monitored and appropriate control or communication decisions are accordingly adopted. It is shown in recent works that state-dependent control and scheduling approaches often outperform static schemes in terms of improving control performance and consuming significantly less of the often costly resources [DFJ12; LL10; RMB09; MH14b].

Within the context of control-aware communication design in NCSs, real-time prioritization can be effectively implemented by awarding scarce communication resources to the systems with the most critical control conditions. Typically, those conditions which determine the necessity of a transmission are formulated as functions of control-related or channel-related states. In deterministic fashion, transmission conditions often appear as event triggers such as threshold policies [WYB02; MH14b]. Dynamic prioritization is alternatively employed in probabilistic fashion through assigning transmission probabilities to each sending station according to state-dependent priority measures [Mam+14; MMH14; TN08; Don+12]. It is shown that probabilistic prioritized channel access mechanisms can be implemented for NCSs with random access protocols resulting in an improved performance compared to non-prioritized counterparts [MKH16]. Earlier mentioned TOD approach is one of the well-known deterministic prioritized resource assignment mechanisms [WYB02]. However, prioritization in the original TOD formulation can be realized at the expense of having a centralized coordination unit, which is not often desired, e.g. in large NCSs, or due to privacy issues. Moreover, centralized coordination of medium access is not possible in many scenarios, such as establishing a network connection, which is performed in a decentralized [GRP16]. Hence, in this chapter, we develop an approach for decentralized state-dependent NCSs priorizitation.

6.3 Adaptive Random Access for Networked Control Systems

In this section, we study a scenario where multiple control loops combined into a NCS are sharing a communication network in a decentralized random access fashion. Sensor units, measuring the state of LTI plants, are transmitting their readings to the respective independent controller units using the multi-channel slotted ALOHA protocol. The section is structured as follows. We start by introducing the problem statement and preliminaries 6.3.1. Stochastic stability of the resulting NCS design in discussed in 6.3.3. Subsection 6.3.4 is dedicated to the numerical performance evaluation and divided into two parts: 6.3.4.1 illustrates the performance of the static scheduler, and in 6.3.4.2 we demonstrate the benefits of using an adaptive scheduler.

6.3.1 Problem statement

We consider an NCS consisting of N physically isolated LTI control sub-systems which are coupled through a shared communication network. A control sub-system i is composed of a linear plant $\mathcal{P}_i$ and a controller $\mathcal{C}_i$. The feedback loop from the plant to the

Table 6.1: Summary of most-used notations in Chapter 6.

x_k^i	system state of a sub-system i at time-step k
e_k^i	error state of a sub-system i at time-step k
w_k^i	system noise of a sub-system i at time-step k
A_i	system matrix of sub-system i
δ_k^i	scheduling variable
θ_k^i	transmission indicator
$\|\cdot\|$	Euclidean norm
$\mathbb{E}[\cdot\|\cdot]$	conditional expectation operator
Λ_i	error threshold for sub-system i
Λ	global error threshold for all sub-systems
N	total number of control sub-systems
M	network state: number of available channels per slot

controller is closed via the shared communication network and the decision of whether to attempt the access to the network is taken by the local scheduler $\mathcal{S}_i$. The plant process is subject to system noise and can be described with the following stochastic difference equation:

$$x_{k+1}^i = A_i x_k^i + B_i u_k^i + w_k^i, \tag{6.1}$$

where $x_k^i \in \mathbb{R}^{n_i}$ denotes the i^{th} system state at time-step k, $u_k^i \in \mathbb{R}^{d_i}$ describes the control input at time-step k. The constant matrices $A_i \in \mathbb{R}^{n_i \times n_i}$, $B_i \in \mathbb{R}^{n_i \times d_i}$ describe system and input matrices, respectively. The noise sequence w_k^i is considered to be an independent and identically distributed (i.i.d) vector distributed according to a zero-mean Gaussian distribution with the covariance matrix W_i. Independent of the noise variables w_k^i, the initial state x_0^i can be considered to be a random variable of any arbitrary symmetric distribution with bounded second moment. At each time-step k, the binary variable $\delta_k^i \in \{0, 1\}$ represents the decision of the **local scheduler**[2] $\mathcal{S}_i$ for sub-system i as follows:

$$\delta_k^i = \begin{cases} 1, & x_k^i \text{ sent through the channel,} \\ 0, & x_k^i \text{ blocked.} \end{cases}$$

Assume that the communication network has M available transmission channels at each time-step (see Fig. 6.2). According to the multi-channel slotted ALOHA protocol, each sub-system which is eligible for transmission selects one of M transmission channels randomly to send its data packet. We denote the number of available channels M as a **network state**.

A collision occurs if two or more sub-systems select the same channel at a certain sample time k. None of the collided sub-systems transmit at k and have to re-try a transmission at $k + 1$. A successful transmission (i.e., the transmitted packet is not

[2]One can think of the local threshold-based scheduler as a variant of a congestion control unit in accordance with the cross-layer model [GNT+06; LSS06].

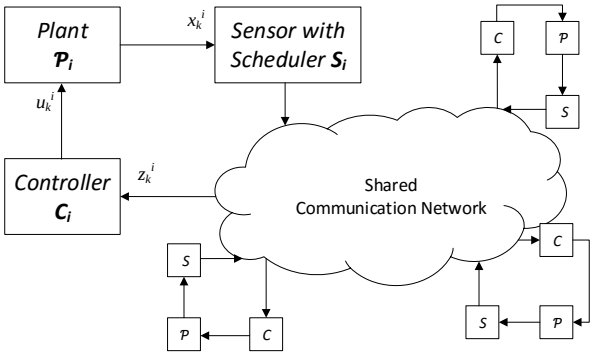

Figure 6.1: A multi-loop NCS with a shared communication medium and local scheduling mechanism.

collided) is denoted by the binary indicator variable $\gamma_k^i \in \{0, 1\}$ as follows:

$$\gamma_k^i = \begin{cases} 1, & x_k^i \text{ successfully received,} \\ 0, & x_k^i \text{ collided.} \end{cases}$$

The reception of data z_k^i at the controller side of the sub-system i is given as a function of scheduling variable δ_k^i and collision indicator γ_k^i as follows:

$$z_k^i = \begin{cases} x_k^i, & \theta_k^i = 1 \\ \varnothing, & \text{otherwise,} \end{cases}$$

where $\theta_k^i = \delta_k^i \gamma_k^i$. Each sub-system is assumed to be controlled by a state-feedback controller which is updated at every time-step k by either the true state values x_k^i (in case the reception is successful, i.e. $\delta_k^i = 1$ and $\gamma_k^i = 1$) or by the state estimates $\mathbb{E}\left[x_k^i\right]$ (in case sub-system i is blocked by its scheduler, i.e. $\delta_k^i = 0$ or a collision occurs, i.e. $\gamma_k^i = 0$).

It is assumed that the sensor and controller of the i^{th} sub-system merely have local knowledge, i.e., of A_i, B_i, W_i and the distribution of x_0^i. Therefore, we assume that the control law ϑ^i is described by measurable and causal mapping of the past observations:

$$u_k^i = \vartheta_k^i(Z_k^i) = -L_i \, \mathbb{E}\left[x_k^i | Z_k^i\right], \tag{6.2}$$

where $Z_k^i = \{z_0^i, \ldots, z_k^i\}$ is the i^{th} controller observation history, and L_i is an arbitrary stabilizing feedback gain. Basically, in accordance with emulation-based approaches, we assume that each loop is stabilized if the data is not received successfully. In case a transmission fails, either due to a blocking by the local scheduler (i.e. $\delta_k^i = 0$) or collision (i.e. $\gamma_k^i = 0$), the estimate of system state x_k^i is computed by a model-based estimator as follows:

$$\mathbb{E}\left[x_k^i | Z_k^i\right] = (A_i - B_i L_i) \, \mathbb{E}\left[x_{k-1}^i | Z_{k-1}^i\right], \tag{6.3}$$

with the initial condition $\mathbb{E}\left[x_0^i | Z_0^i\right] = 0$. The estimate (6.3) is well-behaved only if a stabilizing gain L_i exists to ensure that the closed-loop matrix $(A_i - B_i L_i)$ is Hurwitz.

Accordingly, the network-induced estimation error $e_k^i \in \mathbb{R}^{n_i}$ is defined as the difference between the actual and estimated values of the system state, i.e.

$$e_k^i := x_k^i - \mathbb{E}\left[x_k^i | Z_k^i\right]. \tag{6.4}$$

Having the definition (6.4) and employing (6.1)-(6.3), we can derive the dynamics of the networked-induced error state e_k^i. Assume that a sub-system i successfully transmits at time-step k, i.e. $\theta_k^i = 1$. Therefore, $z_k^i = x_k^i$ and subsequently $u_k^i = -L_i x_k^i$. Thus, the error at the next time-step can be calculated as:

$$\begin{aligned} e_{k+1}^i &= x_{k+1}^i - \mathbb{E}\left[x_{k+1}^i | Z_{k+1}^i\right] \\ &= A_i x_k^i - B_i L_i x_k^i + w_k^i - \mathbb{E}\left[A_i x_k^i - B_i L_i x_k^i + w_k^i | x_k^i\right] \\ &= (A_i - B_i L_i) x_k^i + w_k^i - (A_i - B_i L_i) x_k^i \\ &= w_k^i. \end{aligned}$$

On the other hand, if the sub-system i does not successfully transmit at time-step k, i.e. $\theta_k^i = 0$, then the controller is updated by the estimated value of x_k^i, i.e. $u_k^i = -L_i \mathbb{E}[x_k^i | Z_k^i]$. In this case, we have

$$\begin{aligned} e_{k+1}^i &= A_i x_k^i - B_i L_i \mathbb{E}\left[x_k^i | Z_k^i\right] + w_k^i \\ &- \mathbb{E}\left[A_i x_k^i - B_i L_i \mathbb{E}\left[x_k^i | Z_k^i\right] + w_k^i | Z_k^i\right] \\ &= A_i x_k^i - B_i L_i \mathbb{E}\left[x_k^i | Z_k^i\right] + w_k^i - A_i \mathbb{E}\left[x_k^i | Z_k^i\right] + B_i L_i \mathbb{E}\left[x_k^i | Z_k^i\right] \\ &= A_i (x_k^i - \mathbb{E}\left[x_k^i | Z_k^i\right]) + w_k^i \\ &= A_i e_k^i + w_k^i. \end{aligned}$$

Rewriting the error dynamics for the general θ_k^i, we obtain the following form:

$$e_{k+1}^i = (1 - \theta_k^i) A_i e_k^i + w_k^i. \tag{6.5}$$

Remark 10. *In this chapter, we assume that true state information x_k^i is sent through the communication channel to update the controllers, i.e. the output matrices are considered to be unity and measurement noise does not exist. This assumption is merely considered for the ease of derivations. However, noisy measurements can also be considered under the observability assumption within each sub-system. Then, Kalman filters should be integrated in control units to estimate the system state when new sensor measurements arrive. The full analysis of the output feedback problem is out of scope of this work, however it has been studied in [MKH16].*

The decision whether to attempt a transmission or not is taken by the scheduler $\mathcal{S}_i$ described in the next Sec. 6.3.2.

6.3.2 Local threshold-based scheduler

The local scheduler situated at each local control loop decides to access the medium at every time-step k only if the following threshold inequality holds:

$$\|e_k^i\| > \Lambda_i, \tag{6.6}$$

where Λ_i is the local error threshold for sub-system i. Therefore, if (6.6) is satisfied at some time-step k, then the corresponding sub-system is eligible for transmission at the next time-step $k+1$. Otherwise, it is deterministically excluded from the channel access, i.e.

$$\mathbb{P}[\delta^i_{k+1} = 1 | e^i_k] = \begin{cases} 0, & \text{if } \|e^i_k\| \leq \Lambda_i \\ 1, & \text{otherwise.} \end{cases} \tag{6.7}$$

Note that the deployed scheduling policy (6.7) is not explicitly dependent on whether the transmission has been successful or it has collided, therefore, channel sensing or acknowledgments are not necessary for its implementation.

The communication network model is restricted to the MAC layer and is represented by a multi-channel slotted ALOHA protocol [RS90], see Fig. 6.2. As the most common practical example, we can refer to LTE-based system and its Random Access CHannel (RACH) [LAAZ14], whereas mappings to different single-hop wireless or even bus systems can also be imagined. In every time slot, there are several non-overlapping transmission channels available. We denote the number of available channels as M. As we investigate the multi-channel model in this section, we assume $M \geq 2$. The information about the available number of channels is assumed to be known for all sub-systems in the beginning of each time slot.

For the sake of simplicity, we assume that the communication time slots are equal in duration to the control sampling periods, and that all sub-systems' control periods are synchronized. Thus, in every control period we have M available transmission channels:

$$\sum_{i=1}^{N} \theta^i_k \leq M. \tag{6.8}$$

According to the slotted ALOHA protocol, if a packet is scheduled for transmission, it is sent through one of M channels, randomly chosen. Thus, if we denote a set of sub-systems which are eligible for transmission at time-step k as $\mathcal{G}_k$, then the probability of

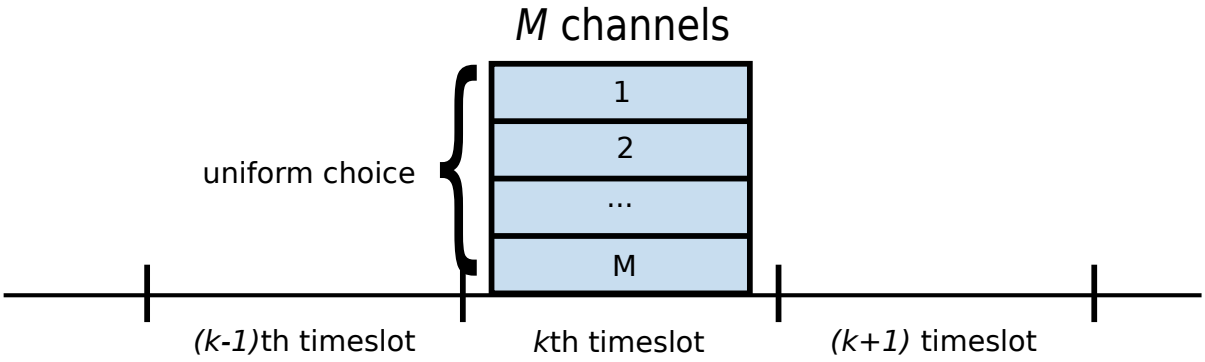

Figure 6.2: Communication system model: multi-channel slotted ALOHA. One time slot is assumed equal to a control period of any sub-system. A channel can represent a frequency, code [Tya+15] or time domain transmission opportunity, depending on the communication technology in use.

successful transmission for a given eligible sub-system in the k time-step is calculated as:

$$\mathbb{P}[\gamma_{k+1}^i = 1 | \|e_k^i\| > \Lambda_i] = \left(\frac{M-1}{M}\right)^{g_k}, \tag{6.9}$$

where g_k is the cardinality of the set $\mathcal{G}_k$.

The transmission threshold Λ_i is directly influencing both the error of the sub-system, and the arrival rate of the requests for network access. Since the network is modeled by slotted ALOHA mechanism, too high arrival rate of requests results in a high collision rate and consequently significantly degrades the performance of the overall networked system. Following this observation, our hypothesis is that adapting Λ_i to network state, can be beneficial for the control performance.

6.3.3 Stability Analysis

In this subsection, we study stability of multiple-loop NCSs with shared multi-channel communication networks subject to the constraint (6.8), and the introduced threshold-based decentralized scheduling policy (6.7). We show stochastic stability of the overall networked system by the notion of Lyapunov Stability in Probability. Before introducing the notion of LSP, we describe the overall network state at some time-step k by the aggregation of the system states x_k^i from all sub-systems $i \in \{1, \ldots, N\}$ and error states e_k^i from all sub-systems $i \in \{1, \ldots, N\}$, i.e. $\left[x_k^\mathsf{T}, e_k^\mathsf{T}\right]^\mathsf{T}$, where $x_k = \left[x_k^{1\mathsf{T}}, \ldots, x_k^{N\mathsf{T}}\right]^\mathsf{T}$ and $e_k = \left[e_k^{1\mathsf{T}}, \ldots, e_k^{N\mathsf{T}}\right]^\mathsf{T}$. From (6.1)-(6.3), together with the definition of the estimation error e_k^i in (6.4), it is straightforward to see that the individual aggregate networked state $\left[x_k^{i\mathsf{T}}, e_k^{i\mathsf{T}}\right]^\mathsf{T}$ within each sub-system i has triangular dynamics as follows:

$$\begin{bmatrix} x_{k+1}^i \\ e_{k+1}^i \end{bmatrix} = \begin{bmatrix} A_i - B_i L_i & (1-\theta_k^i) B_i L_i \\ 0 & (1-\theta_k^i) A_i \end{bmatrix} \begin{bmatrix} x_k^i \\ e_k^i \end{bmatrix} + \begin{bmatrix} w_k^i \\ w_k^i \end{bmatrix}. \tag{6.10}$$

This implies that the evolution of the error state e_k^i is in fact independent of the system state x_k^i. We employ an emulation-based control design to stabilize the control sub-systems in case their corresponding loops are closed, i.e. the controllers are updated with their own true state values. Thus, assuming each pair (A_i, B_i) is stabilizable, there exists stabilizing feedback gain L_i such that the closed-loop matrix $(A_i - B_i L_i)$ is Hurwitz, and consequently the system state x_k^i is asymptotically stable. It should however be noted that existence of stabilizing control inputs u_k^i's does not guarantee the stability of overall networked system with the introduced networked state $\left[x_k^{i\mathsf{T}}, e_k^{i\mathsf{T}}\right]^\mathsf{T}$, since the evolution of the error state is independent of the control laws. This statement is clear from (6.10), which illustrates that if a sub-system does not transmit at a certain time-step, stabilizing gain L_i guarantees the stability only if error state e_k^i is stable. Now we are ready to introduce the concept of stability considered in this chapter, i.e. LSP:

Definition 5. *(Lyapunov Stability in Probability, [Koz69]) A linear system with state vector x_k possesses LSP if given $\varepsilon, \varepsilon' > 0$, exists $\rho(\varepsilon, \varepsilon') > 0$ such that $|x_0| < \rho$ implies*

$$\lim_{k\to\infty} \sup \mathbb{P}\left[x_k^T x_k \geq \varepsilon'\right] \leq \varepsilon. \tag{6.11}$$

The following lemma shows that LSP is achievable by solely considering the aggregated error state e_k.

Lemma 5. *For an NCS described by (6.1)-(6.5), the condition in (6.11) is equivalent to*

$$\lim_{k\to\infty} \sup \mathbb{P}\left[e_k^T e_k \geq \xi'\right] \leq \xi, \tag{6.12}$$

where $\xi' > 0$ and the constant ξ fulfills $0 \leq \xi \leq \varepsilon$.

Proof. As follows from (6.2)-(6.4), the system state x_k^i for each control loop i evolves as

$$x_{k+1}^i = (A_i - B_i L_i) x_k^i + (1 - \theta_k^i) B_i L_i e_k^i + w_k^i. \tag{6.13}$$

As already discussed, the evolution of the error e_k^i is independent of the system state x_k^i within each individual control loop. Furthermore, by assuming the emulative control law (6.2), the closed-loop matrix $(A_i - B_i L_i)$ is ensured to be Hurwitz. Together with the assumption that x_0^i has a symmetric bounded variance distribution, it follows that the system state x_k^i is converging with any stabilizing feedback gain L_i. In addition, the disturbance process w_k^i is i.i.d. according to $\mathcal{N}(0, I)$, and is bounded in probability. Thus, showing $\lim_{k\to\infty} \sup \mathbb{P}\left[e_k^{i\mathsf{T}} e_k^i \geq \xi_i'\right] \leq \xi_i$ ensures existence of constants ε_i and $\varepsilon_i' > 0$ such that $\lim_{k\to\infty} \sup \mathbb{P}\left[x_k^{i\mathsf{T}} x_k^i \geq \varepsilon_i'\right] \leq \varepsilon_i$, where $\xi_i \leq \varepsilon_i$. As individual loops operate independently, we take the aggregate NCS state (x_k, e_k). Then, the existence of ξ and $\xi' > 0$ such that $\lim_{k\to\infty} \sup \mathbb{P}\left[e_k^T e_k \geq \xi'\right] \leq \xi$, implies existence of ε and $\varepsilon' > 0$ such that $\lim_{k\to\infty} \sup \mathbb{P}\left[x_k^T x_k \geq \varepsilon'\right] \leq \varepsilon$ for $\xi \leq \varepsilon$, and the proof readily follows. □

This lemma enables us to study stability of the overall networked system only by looking at the error state e_k, considering that stabilizing feedback gains L_i are designed.

Using Markov's inequality, we employ the following inequality for $\xi' > 0$ as

$$\mathbb{P}\left[e_k^T e_k \geq \xi'\right] \leq \frac{\mathbb{E}\left[e_k^T e_k\right]}{\xi'}. \tag{6.14}$$

This confirms that showing that the error is uniformly bounded in expectation ensures finding appropriate ξ and $\xi' > 0$ such that (6.12) is satisfied for arbitrary $\rho(\xi', \xi)$. Therefore, we focus on deriving an upper bound for the expectation of quadratic error norm, i.e.

$$\mathbb{E}\left[e_k^T e_k\right] = \sum_{i=1}^{N} \mathbb{E}\left[e_k^{i\mathsf{T}} e_k^i\right] = \sum_{i=1}^{N} \mathbb{E}\left[\|e_k^i\|^2\right] \tag{6.15}$$

This modifies the condition (6.12) as follows:

$$\lim_{k\to\infty} \sup \mathbb{P}\left[e_k^T e_k \geq \bar{\xi}'\right] \leq \bar{\xi}. \tag{6.16}$$

Due to the nature of the multi-channel communication network with capacity constraint (6.8) and threshold-based scheduler policy (6.7), the boundedness of (6.15) cannot always be shown over one time-step transition, i.e. $k \to k+1$. This observation is discussed in the following illustrative example:

Illustrative example. Consider an NCS consisting of three identical scalar unstable sub-systems with systems matrices $A_1 = A_2 = A_3 = A > 1$, competing for two available transmission channels at each time slot over a shared multi-channel communication network. For simplicity, assume $\Lambda_1 = \Lambda_2 = \Lambda_3 = \bar{\Lambda}$, and $e_k^1 = e_k^2 = e_k^3 = \bar{e}_k$. In addition, consider that the condition (6.6) is fulfilled, i.e. all three sub-systems are eligible for channel access at time-step $k+1$. Each sub-system selects each of the two available transmission channels by probability of $\frac{1}{2}$. Two scenarios are possible: 1) One successful transmission occurs from one of the sub-systems, and the other two inevitably collide. It is straightforward to calculate that this scenario happens with the probability of $\frac{3}{4}$; 2) All three sub-systems choose the same transmission channel and consequently all three will collide, which means no successful transmission is occurred, where this scenario occurs with probability of $\frac{1}{4}$. As the sub-systems are identical, and for the sake of illustrative purposes, assume a realization for the first scenario that e.g. sub-system 1 transmits and sub-systems 2 and 3 are collided. Employing (6.5), we calculate the error expectation in (6.15) for one step transition, as follows:

$$
\begin{aligned}
&\sum\nolimits_{i=1}^{3} \mathbb{E}\left[\|e_{k+1}^i\|^2 | e_k\right] = \sum\nolimits_{i=1}^{3} \mathbb{E}\left[\|\left(1-\theta_k^i\right) A e_k^i + w_k^i\|^2\right] \\
&= \frac{1}{4} \sum\nolimits_{i=1}^{3} \mathbb{E}[\|A\bar{e}_k + w_k^i | \bar{e}_k\|^2] \\
&+ \frac{3}{4} \left(\mathbb{E}[\|A\bar{e}_k + w_k^2 | \bar{e}_k\|^2] + \mathbb{E}[\|A\bar{e}_k + w_k^3 | \bar{e}_k\|^2] + \mathbb{E}[\|w_k^1\|^2]\right) \\
&= \frac{1}{4} \sum\nolimits_{i=1}^{3} \|A\bar{e}_k\|^2 + \mathbb{E}[\|w_k^i\|^2] \\
&+ \frac{3}{4} \left(2\|A\bar{e}_k\|^2 + \sum\nolimits_{i=1}^{3} \mathbb{E}[\|w_k^i\|^2]\right) \\
&= \frac{1}{4} \left(3\|A\bar{e}_k\|^2 + 3\right) + \frac{3}{4} \left(2\|A\bar{e}_k\|^2 + 3\right) \\
&= 3 + 2.25\|A\bar{e}_k\|^2,
\end{aligned}
$$

which is not uniformly bounded for arbitrary $\bar{e}_k$ and system matrix A. Intuitively, between two consecutive transmissions of each sub-system, they operate in open loop. Hence, in general, the respective local errors are expected to grow. Thus, to obtain boundedness of error state, we need to look at an interval of time-steps rather than only one transition step such that, given the constraint (6.8), there is a non-zero probability for all sub-systems to transmit successfully. Therefore, one can infer that an interval of length $\lceil \frac{N}{M-1} \rceil$ provides enough transmission opportunities[3] for an NCS of N sub-systems with M available transmission channels per time-step. It should be reminded that the linearity of our sub-systems guarantees the boundedness over any finite longer horizons.

[3]For $g_k \leq M$, period of $\lceil \frac{N}{M} \rceil$ is sufficient. However, for $g_k > M$ at least one collision occurs in the first time step.

For the tractability of the stability analysis, we assume the worst case scenario by considering the minimum number of available transmission channels, i.e. only $M = 2$ at each time-step. This yields that the minimum length of the interval over which LSP is investigated equals N.

Theorem 3. *Consider an NCS with N heterogeneous LTI control sub-systems, with the plants given by (6.1), sharing a multi-channel communication network with two available transmission channels per time-step. Given the control law (6.2) and threshold policy (6.7), the NCS of interest is Lyapunov stable in probability if the MAC employs slotted ALOHA protocol.*

Proof. See our previously published work [Vil+16a] for the proof of the theorem. □

Remark 11. *The notion of stability considered in this chapter, i.e., LSP, determines the probability that the overall NCS state remain bounded. This probability is not one due to the fact that there exists a non-zero probability, though might be very close-to-zero, such that at all time-steps the NCS is operating all the transmissions fail due to successive collisions. This is the structural property of the decentralized MAC we are considering in this chapter and in case such a scenario occurs, it means all control loops, either stable or unstable, operate in open-loop which consequently lead to instability of the overall NCS due to the presence of unstable plants.*

6.3.4 Performance evaluation

In this subsection, we evaluate the performance of a threshold-based scheduler over multi-channel slotted ALOHA. Both communication and control-related performance metrics are investigated.

For the simulations, we consider a setup as follows. An NCS in consideration is composed of two heterogeneous classes of scalar control loops: class one including multiple homogeneous stable plants with the system matrix $A_1 = 0.75$ and class two consisting of open-loop unstable plants with $A_2 = 1.25$. The plants within each group are homogeneous, and all sub-systems are influenced by the i.i.d. noise processes randomly chosen from the standard normal distribution, i.e. $w_k^i \sim \mathcal{N}(0, 1)$ for all time-steps k. The input matrices for both groups are $B_1 = B_2 = 1$. For the plants' stabilization, deadbeat control law $L_i = A_i$ is employed. We consider the total amount of sub-systems to be N, while each group of control loops has $N/2$ sub-systems. The number of transmission channels in each time step, unless stated otherwise, is considered to be $M = 10$. It is worth mentioning that not only stability or instability of a plant determines the urge of a transmission, but also system noise influences the threshold-based policy. Therefore, it is not guaranteed that if a plant is stable, then it is asymptotically stable even if no transmission is associated with that sub-system. Due to presence of noise, a sub-system with stable plant might become in more urgent situation for transmission than a sub-system with an open-loop unstable plant.

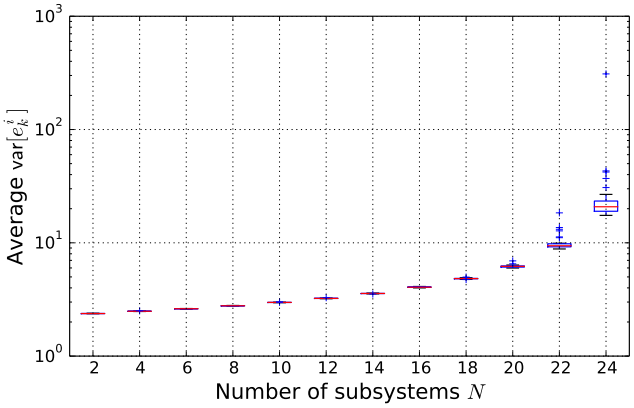

Figure 6.3: Average error variance Σ vs. number of sub-systems N (30 runs). Parameters: $M = 10$, $\Lambda = 2$.

For a control performance evaluation, we study the average error variance among N sub-systems:

$$\Sigma = \frac{1}{N} \sum_{i=1}^{N} \mathsf{var}[e_k^i] \tag{6.17}$$

For the communication performance, we use two metrics. First one is average channel utilization, commonly known as normalized throughput T, defined as:

$$T = \frac{\mathbb{E}[s]}{M}, \tag{6.18}$$

where $\mathbb{E}[s]$ is expected number of successful transmissions s per slot. Ratio of collided packets is used as the second performance metric. It is defined as:

$$r_{coll} = \mathbb{E}\left[\frac{c}{c+s}\right], \tag{6.19}$$

where c is the number of collided transmissions per slot.

The transmission threshold Λ_i is considered homogeneous for all N sub-systems throughout the simulation:

$$\Lambda_i = \Lambda_j = \Lambda, \quad \forall i, j \in N. \tag{6.20}$$

6.3.4.1 Static Threshold Scheduler

For the first setup, we consider a scheduler where the transmission threshold is chosen arbitrary and is independent of the number of transmission channels M.

Fig. 6.3 demonstrates the evolution of the average error variance with the increasing number of sub-systems. We observe a non-linear growth of the error variance, and, on the same time, higher variation of the resulting variance over multiple runs. The growth

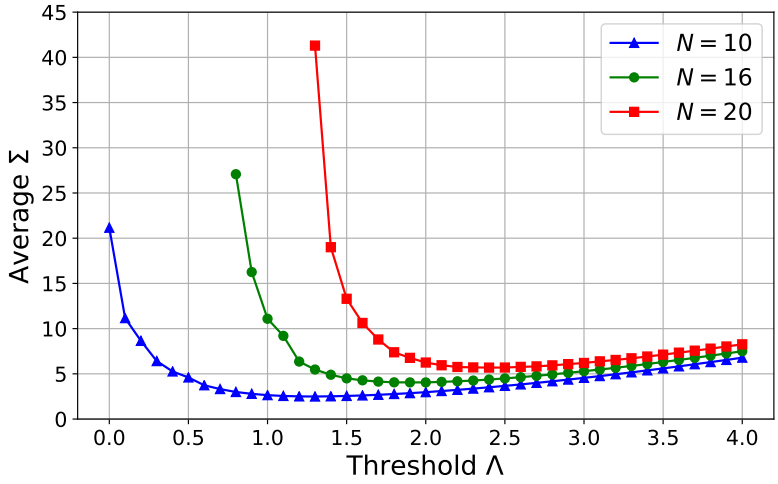

Figure 6.4: Average error variance Σ vs. Λ. Parameters: $M = 10$.

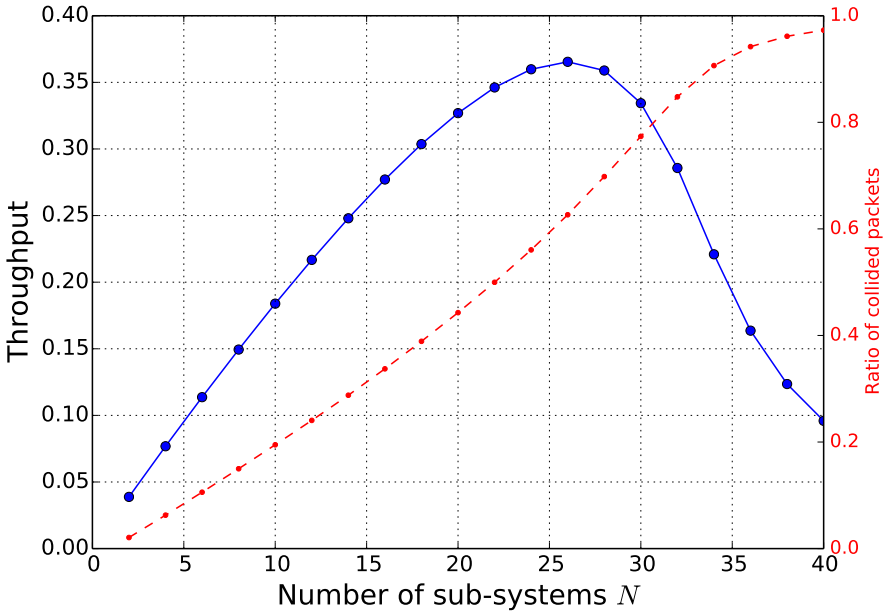

Figure 6.5: Average normalized throughput T and collision rate vs. number of sub-systems N. Parameters: $M = 10$.

of the error variance can be explained by looking at Fig. 6.5: With the increasing number of sub-systems, we observe an increase in collision rate. Since the error accumulates exponentially with every collision for the unstable systems, linear increase in collisions results in a non-linear increase in the variance of the error.

In Fig. 6.5, we observe that the shape of the plot for throughput corresponds to the commonly known dependency for multi- and single-channel slotted ALOHA with Poisson distribution arrival rate [Tya+15]. The highest value $T \approx 1/e \approx 0.368$ is achieved at $N = 26$.

Fig. 6.4 shows how the error variance depends on the transmission threshold Λ. As we observe, and it is inline with the hypothesis we have stated in the Subsection 6.3.1, the dependency is a convex function. With the values of Λ close to 0, the transmission is attempted every time, thus, causing many collisions and shifting the throughput S operating region as in 6.5 to the right. The collisions, in turn, further increase the $\|e_k^i\|$ for all unstable systems with $A_i > 1$, thus, further increasing the amount of access attempts. As expected, the error variance among all sub-systems grows. If, however, the Λ is chosen too high, the increase in the error variance is caused by the underutilized communication medium (throughput T low). Thus, it is observed that there exists an optimal value for $\Lambda = \Lambda^\star$ in a given NCS scenario defined by N, M.

6.3.4.2 Scheduler with Threshold Adaptation

Following the observation about the existence of an optimal $\Lambda = \Lambda^\star$, we propose a simple illustrative improvement to the threshold design defined in (6.21). Namely, we use a knowledge about the current network state M and the number of present sub-systems N, in order to choose the Λ for the optimal performance in terms of the average error variance:

$$\Lambda^\star = \arg\max_{\Lambda} \Sigma\left(N, M\right), \tag{6.21}$$

where higher number of channels results in a higher Λ. Numerically obtained values for $\Lambda^\star$ for M and N choices we use for evaluation are summarized in Table 6.2.

The benefits of this approach can be seen for the case of the varying number of available channels M. For simplicity, we model the number of channels as a random variable with two possible values $M \in \{M_1, M_2\}$, $M_1 < M_2$, with:

$$\alpha \triangleq 1 - \mathbb{P}[M = M_1] = \mathbb{P}[M = M_2]. \tag{6.22}$$

Table 6.2: Optimal event-trigger threshold $\Lambda^\star$.

	N						
	4	6	8	10	12	14	16
$M = 5$	1.0	1.5	2.0	2.4	3.5	5.2	8.1
$M = 10$	0.6	0.8	1.0	1.2	1.4	1.6	1.8

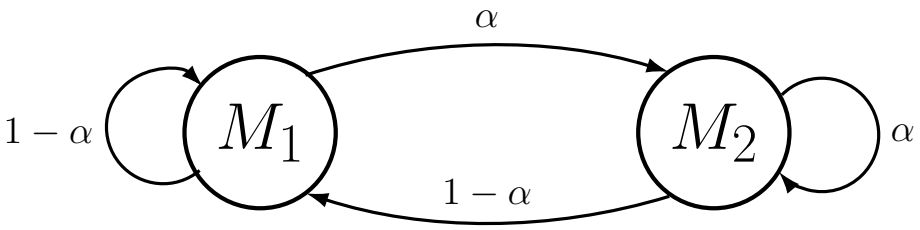

Figure 6.6: Model of number of channels M variations.

The model is depicted in Fig. 6.6. These two states can represent presence or absence of a background traffic with reserved channels, for example, as described in [KKA13; LKY11]. Although we consider only two states for M, it has to be noted that the proposed scheduler design is extendable for a more general case of multiple states. In the evaluation scenario $M_1 = 5$ and $M_2 = 10$, and $\alpha = 0.5$ are assumed.

For comparison, we consider two choices of Λ for static scheduler: **(A)** first, where Λ is statically set to minimize the error variance for $M = M_1$, and **(B)** second to minimize the error variance for $M = M_2$ for a given number of sub-systems in the simulation N. The comparison results are presented in Fig. 6.7. It is observed, that the error variance with the adaptive scheduler is always lower or equal than for non-adaptive. It is further observed, that the first static scheduler **(A)**, optimizing the threshold for the lower number of channels M_2 is performing noticeably better than the scheduler **(B)**, optimizing the threshold for the higher number of channels M_1. The effect is supported by the observations from Fig. 6.4 that the slope on the left from the optimal point is much higher than on the right from it, thus, over-utilization is more harmful for the error variance than underutilization.

To evaluate how the probability of a network state change α influences the performance gain of the adaptive scheduler, we use the reduction of the average error variance as a metric for adaptation gain:

$$G_{adap} = \frac{\Sigma_{na} - \Sigma_a}{\Sigma_{na}}, \tag{6.23}$$

where Σ_{na} and Σ_a represent the average error variances for static (non-adaptive) and adaptive schedulers, respectively.

The resulting dependency is depicted in Fig. 6.8. The parameter α is in this case a

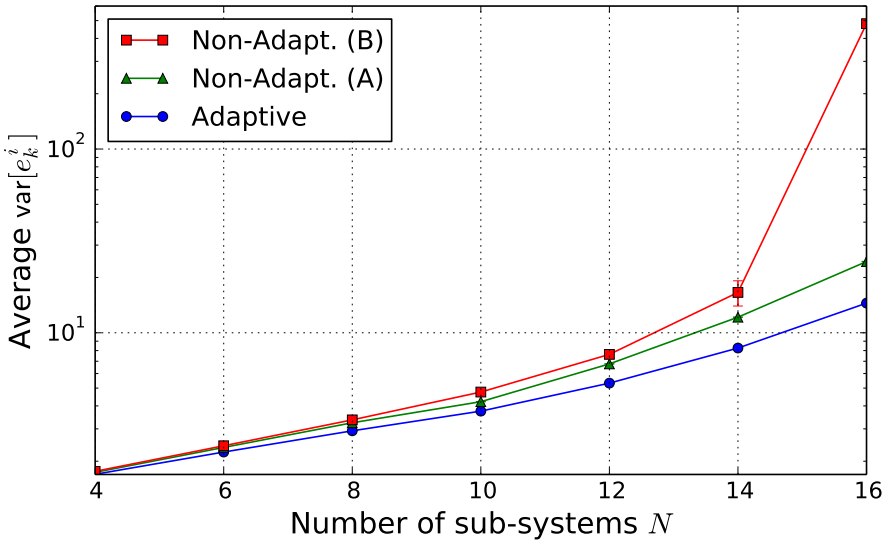

Figure 6.7: Average error variance vs. number of sub-systems N for three cases: Adaptive $\Lambda^\star$, Non-Adaptive (A) (Λ optimal for M_1 channels), Non-Adaptive (B) (Λ optimal for M_2 channels). Parameters: $M_1 = 5$, $M_2 = 10$, $\alpha = 0.5$.

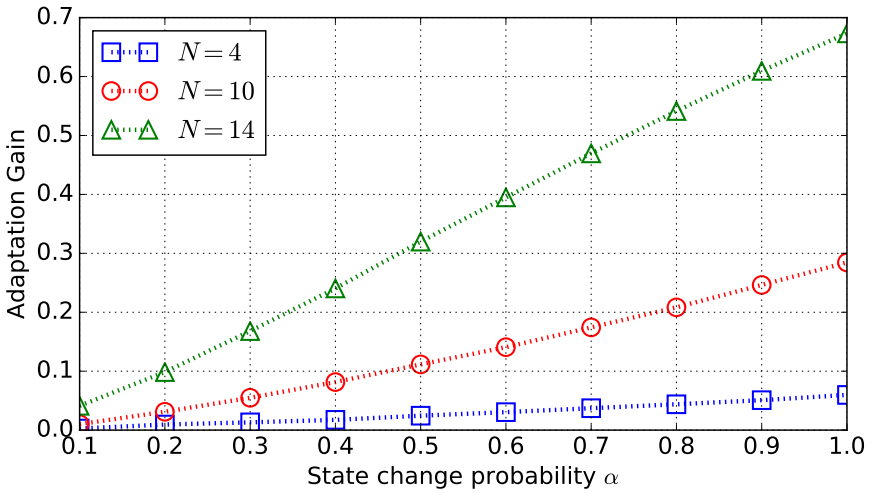

Figure 6.8: Adaptation gain G_{adap} vs. probability of the "good" channel α for $N \in \{4, 10, 14\}$. Parameters: $M_1 = 5$, $M_2 = 10$.

measure of how frequently the network state is changing. For $\alpha = 0.1$ almost no changes are there, hence, both schedulers are close to optimal. On the other hand, for $\alpha = 1$, although also no changes are present, the default state of the channel is $M = M_2$, thus, the static scheduler is not optimal in any time-slot. For the network state changing every second time, the adaptive scheduler is able to reduce the error variance by up to 30%.

6.4 Binary Countdown for Prioritization in Networked Control Systems

In the previous section, we have observed the significant impact which collisions have on the networked-induced error of an NCS. We have also observed the coupling between the event-triggering policy and the collision rate via the threshold, and how a proper adjustment of the threshold can improve the performance and lower the networked-induced error. We have considered only a uniform threshold setting for all control sub-systems, reactively adjusted according to the network state, i.e. available network resources. It is intuitively clear that the globally optimal threshold policy should also reflect the internal dynamics of the sub-system, in addition to the network state. To address this limitations, in this section, we develop an approach for a state-dependent contention resolution, where the access priority during the contention is determined dynamically for every sub-system based on its plant current state, measured by the sensor. For that, we adapt Binary Countdown Contention Resolution (BCCR) approach introduced earlier in Chapter 4.

The section structured as follows. We present the problem formulation and the NCS model in Subsec. 6.4.1. The deterministic priority assignment mechanism for random

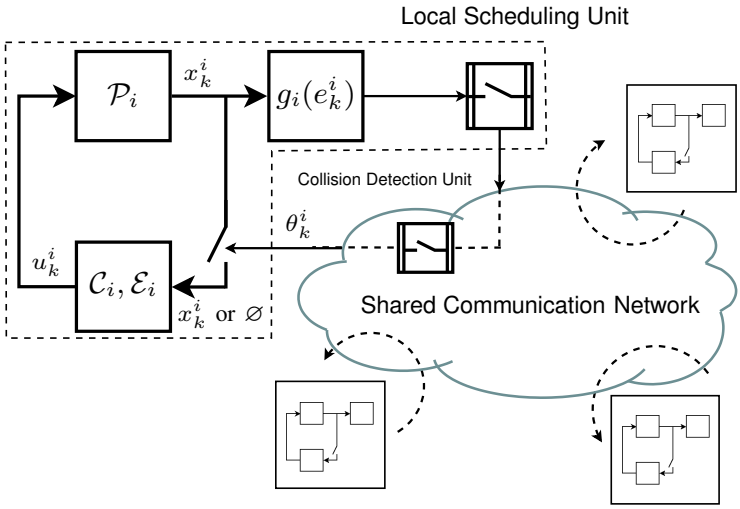

Figure 6.1: Schematic of a shared resource multi-loop NCS with local scheduling unit.

access channels is introduced in Subsec. 6.4.2. Stability properties of the described NCS under the proposed channel arbitration is discussed in Subsec. 6.4.3. Simulation results and discussions are presented in Subsec. 6.4.4.

6.4.1 Problem Statement

Throughout this section, the system model is largely similar to the previous 6.3.1. The main difference are that we target slotted ALOHA with a single channel, and the focus of the work is instead on the dynamic plant state dependent prioritization technique. Additionally, no local threshold based policy is explicitly assumed here, but the priority levels described later in Sec. 6.4.2 serve as implicit thresholding policy. To allow consistent reading, we repeat main assumptions of the problem statement here.

We consider NCS consisting of N heterogeneous linear controlled sub-systems that share a common communication medium subject to capacity limitations. Each local sub-system $i \in \{1, \ldots, N\}$ consists of an LTI stochastic process $\mathcal{P}_i$, and a control unit including a state estimator $\mathcal{E}_i$ and a feedback controller $\mathcal{C}_i$, see Fig. 6.1. Again, we model each sub-system i in discrete-time by the following stochastic state-space equation

$$x_{k+1}^i = A_i x_k^i + B_i u_k^i + w_k^i, \tag{6.24}$$

where vectors $x_k^i \in \mathbb{R}^{n_i}$ and $u_k^i \in \mathbb{R}^{m_i}$ denote local system state, and control input of sub-system i, respectively, with $A_i \in \mathbb{R}^{n_i \times n_i}$ and $B_i \in \mathbb{R}^{n_i \times m_i}$ describing system matrix, and input matrix. In addition, the pair (A_i, B_i) is assumed to be controllable. The system noise is also assumed to be a random sequence where the sample realizations $w_k^i \sim \mathcal{N}(0, W_i)$ are independent and identically distributed (i.i.d.). The initial state x_0^i for all $i \in \{1, \ldots, N\}$ are randomly chosen from an arbitrary distribution with bounded second moment.

The communication channel is such that only a limited number of sub-systems can transmit simultaneously. To determine whether a transmission is feasible at each sample time, decentralized scheduling units $\mathcal{S}_i$ are integrated within each local sub-system. At every time-step k the decision about either "transmit" or "back-off" is denoted by the binary variable $\delta_k^i \in \{0,1\}$ as follows

$$\delta_k^i = \begin{cases} 1, & x_k^i \text{ sent through the channel} \\ 0, & x_k^i \text{ blocked.} \end{cases}$$

Without loss of generality, we assume that if data packets do not collide in the channel, they are received and successfully decoded by their corresponding control units. Successful transmission of a data packet is acknowledged via an error-free link to every station. We assume that if a collision occurs, then involved sub-systems are blocked and data packets are dropped. We define the binary variable γ_k^i as the collision indicator at a time-step k as follows

$$\gamma_k^i = \begin{cases} 1, & x_k^i \text{ successfully received,} \\ 0, & x_k^i \text{ collided.} \end{cases}$$

Upon receiving new state information at the control unit, the control input is computed by a local feedback controller. In case of a blocked transmission, a model-based estimation of the system state is utilized. Defining the history of received information *and observed channel state* at a control side $\mathcal{C}_i$ as $\mathcal{I}_k^i = \{\theta_0^i, x_0^i, \ldots, \theta_{k-1}^i, x_{k-1}^i\}$, where $\theta_k^i \triangleq \delta_k^i \gamma_k^i$, we have

$$\mathbb{E}[x_k^i|\mathcal{I}_k^i] = \begin{cases} x_k^i, & \theta_k^i = 1, \\ (A_i - B_i L_i)\,\mathbb{E}[x_{k-1}^i|\mathcal{I}_{k-1}^i], & \theta_k^i = 0. \end{cases} \tag{6.25}$$

The control input u_k^i is then computed according to the following linear state-feedback law described by a measurable and causal mapping of $\mathcal{I}_k^i$

$$u_k^i = -L_i\,\mathbb{E}[x_k^i|\mathcal{I}_k^i], \tag{6.26}$$

where, $L_i \in \mathbb{R}^{m_i \times n_i}$ is a stabilizing feedback control gain. The estimate (6.25) is well-behaved since the control gain L_i is stabilizing, and the pair (A_i, B_i) is stabilizable.

We introduce the network-induced error e_k^i for each sub-system $i \in \{1, \ldots, N\}$, at every time-step k as follows

$$e_k^i \triangleq x_k^i - \mathbb{E}[x_k^i|\mathcal{I}_k^i]. \tag{6.27}$$

Concatenating the system state x_k^i and the error state e_k^i in one vector, i.e. $[x_k^{i\mathsf{T}}\ e_k^{i\mathsf{T}}]^\mathsf{T}$, as the aggregate state of sub-system i in the NCS, it is straightforward to derive the following local dynamics according to (6.24)-(6.27), analogous to (6.10).

$$x_{k+1}^i = (A_i - B_i L_i)x_k^i + B_i L_i e_k^i + w_k^i, \tag{6.28}$$

$$e_{k+1}^i = \left(1 - \theta_k^i\right) A_i e_k^i + w_k^i. \tag{6.29}$$

Table 6.1: Summary of the communication model notations

$\mathcal{K}$	set of transmission priorities
n	number of contention resolution slots
$\hat{m}_k^i$	priority of the node i at time-step k
$(M_k^i)^n$	priority of the node i as a binary sequence of length n
$m_k^{i,j}$	jth element of the sequence m_k^i; $m_k^{i,j} \in \{0,1\}$
T_s	duration of the data transmission
T_{CR}	duration of a contention resolution slot
p_b	back-off probability in a given transmission slot

It can be seen from (6.29) that evolution of error state e_k^i is again independent from the system state x_k^i within every local sub-system. This enables us to take an emulation-based control approach and choose a stabilizing controller as in (6.26). The control inputs are realized according to this law with true state value x_k^i if transmission is successful. Otherwise, the control inputs are computed with model-based estimate $\mathbb{E}[x_k^i|\mathcal{I}_k^i]$. Note that stabilizability of pair (A_i, B_i) ensures that the closed-loop matrix $(A_i - B_i L_i)$ is stable. Hence, it follows from (6.28) a sub-system i with the aggregate state $[x_k^{i\mathsf{T}} \; e_k^{i\mathsf{T}}]^\mathsf{T}$ is stable if the error state e_k^i is convergent.

6.4.2 Priority-based Contention Resolution MAC

We assume that the communication operates with a time slotted medium access (i.e., there is a basic level of coarse synchronization between the nodes, achieved with periodic broadcast signals [YYH03]), where the timeline is divided into *transmission slots* of equal duration T, where T is equal to the sampling period of control systems. Conventional slotted ALOHA protocols assume that all communicating nodes send the data directly in the closest available transmission slot, so if another node sends in the same slot, transmissions collide and data packets are lost. The communication channel is assumed to be under the following per-time-step constraint

$$\sum\nolimits_{i=1}^{N} \theta_k^i \leq 1, \tag{6.30}$$

which specifies that at most one node can successfully transmit at each time-step.

Here, we further consider a variation of slotted ALOHA with *access reservation phase*, where the access reservation is performed via priority-based contention resolution [YYH03]. The modified transmission slot structure consists of a contention resolution period of duration nT_{CR}, and a data transmission period of duration T_s (i.e., instead of only data transmission period in classic slotted ALOHA). This scenario is schematically shown in Fig. 6.2. Contention resolution period consists of n contention resolution slots, each of duration T_{CR}. The proposed contention resolution protocol operates as fol-

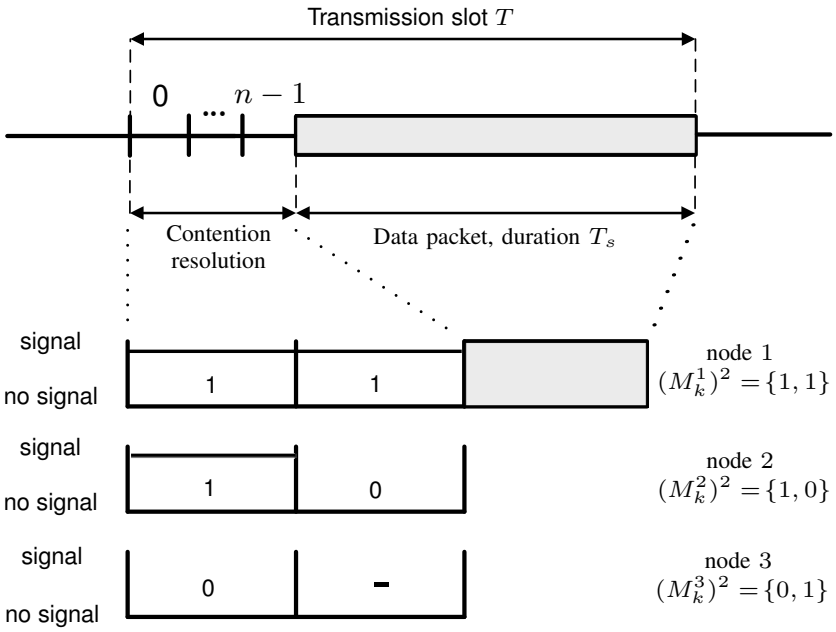

Figure 6.2: Transmission slot structure and priority resolution example.

lows. Prior to transmission at a time-step k, node i is assigned a priority level $\hat{m}_k^i \in \mathbb{N}_0$ from the common set of priorities $\mathcal{K}$. The value of $\hat{m}_k^i$ is mapped onto a binary sequence $(M_k^i)^n = \{m_k^{i,0}, \ldots, m_k^{i,n-1}\}$ of length n (padded with zeros until n, if necessary) such that

$$\begin{aligned} \hat{m}_k^i &= m_k^{i,0} \cdot 2^{n-1} + m_k^{i,1} \cdot 2^{n-2} + \ldots \\ &+ m_k^{i,j} \cdot 2^{n-j-1} + \ldots m_k^{i,n-1} \cdot 2^0. \end{aligned} \tag{6.31}$$

This sequence is then used to decide whether the node i sends a presence signal in a given contention resolution slot or not. If $m_k^{i,j} = 1$, the node is sending in the j^{th} slot, otherwise, the node keeps listening to the medium. In the latter case, if the node i detects the presence of a signal in the j^{th} slot, it means that there exits a contending node p with priority $\hat{m}_k^p > \hat{m}_k^i$. Hence, node i does not proceed with contention resolution and does not send data in the current data transmission period.

Illustrative example: Three sub-systems (nodes) with IDs $i \in \{1,2,3\}$ are assigned with the priorities $\hat{m}_k^1 = 2$, $\hat{m}_k^2 = 1$ and $\hat{m}_k^3 = 0$, respectively, where higher value implies higher priority, see Fig. 6.2. Given $\mathcal{K} = \{0,1,2,3\}$, there are $n = \log_2|\mathcal{K}| = 2$ contention resolution slots. Thus, the binary priority sequences are $(M_k^1)^2 = \{1,1\}$, $(M_k^2)^2 = \{1,0\}$, and $(M_k^3)^2 = \{0,1\}$, respectively. Therefore, only node 1 and node 2 are sending signals in slot 1 (i.e., $M_k^{1,1} = M_k^{2,1} = 1$), while node 3 is only listening to the medium ($m_k^{3,1} = 0$). Since node 3 detects a non-empty signal in the resolution slot 1, it concludes that nodes with higher priorities exist, and hence backs off. At the resolution slot 2, only node 1 sends the signal, and node 2 is detecting it while listening. Hence, node 2 does not proceed with transmission and as a result, only node 1 continues with sending its data packet.

Note that the proposed MAC protocol is not capable of fully resolving the contentions. In case there exist at least two nodes on the highest priority in a given slot, they collide. Therefore, to reduce the probability of successive collisions with randomization, we introduce a barring factor $p_b \in [0, 1)$ which denotes the probability that each node decides whether to attempt a transmission at all, i.e. $\mathbb{P}[\delta_k^i = 1] = 1 - p_b$. If two nodes are assigned with the highest priority $m_k^p = m_k^i = m_{\max}$, the probability of collision is given by $(1 - p_b)^2$, otherwise, if $p_b = 0$, collision probability is 1.

Taking into account both the decision to transmit in a given time-step k, and the priority level $\hat{m}_k^i$ defined by (6.33), the probability that a given node successfully transmits is

$$\mathbb{P}[\theta_k^i = 1] = (1 - p_b)\, \mathbb{P}[\hat{m}_k^i > \hat{m}_k^j, \forall j \in \mathcal{N}_a \setminus i], \tag{6.32}$$

where $\mathcal{N}_a \subseteq \{1, \ldots, N\}$ is a subset of sub-systems eligible for transmission (non-barred).

6.4.2.1 State-dependent Priority Measure

If a node decides to attempt a transmission, the question of choosing the set of priority levels and determining the priority of a given node or data packet is left open in the current protocol. Usually, it is assumed to be static, i.e., time-invariant and pre-determined for a given node. In contrast to the state-of-the-art, we propose to choose the priority levels dynamically at every time-step k for every sub-system i with the following deterministic law

$$\hat{m}_k^i = \begin{cases} 0 & \text{if } g\left(e_k^i\right) < \lambda_i, \\ \lceil g\left(e_k^i\right) \rceil & \text{if } \lambda_i \leq g\left(e_k^i\right) \leq m_{\max}, \\ m_{\max} & \text{otherwise,} \end{cases} \tag{6.33}$$

where function $g(\cdot) : \mathbb{R}^{n_i} \to \mathbb{R}^+$ is assumed to be continuous and strictly increasing with the increase of the absolute value of the vector e_k^i element-wise. The exact form of the law $g\left(e_k^i\right)$ can be determined empirically and adjusted according the target set-up. It can also be derived as the optimal solution for a given cost function. The following results however are generic for any function g with the mentioned properties. Here, $m_{\max}$ is a static parameter, which can be dimensioned according to the number of sub-systems in the network. It is straightforward to conclude that the required number of contention resolution slots for up to $m_{\max}$ priorities equals $n = \lceil \log_2(m_{\max} + 1) \rceil$.

According to the priority assignment law (6.33) we can compute the probability that a forwarded transmission fails to be successfully delivered due to collision. Considering $p_b = 0$, i.e. the worst case collision scenario, a collision occurs if at least two sub-systems i and j are assigned with identical highest priority orders. Thus, the probability that at

a time-step k, a collision occurs equals

$$\begin{aligned}\mathbb{P}\left[\theta_k^i = 0, \forall i\right] &= \mathbb{P}\left[m_k^i = 0, \forall i \in \{1, \ldots . N\}\right] \\ &+ \mathbb{P}[\underbrace{m_k^1 = \ldots = m_k^p}_{\cup p \in \{2,3,\ldots,N\}} = \bar{m}_k > m_k^l, l \neq p] \\ &+ \mathbb{P}[\underbrace{m_k^1 = \ldots = m_k^p}_{\cup p \in \{2,3,\ldots,N\}} = m_{\max}],\end{aligned} \tag{6.34}$$

where $\bar{m}_k$ is an arbitrary priority assignment from the set of priorities $\mathcal{K}$, except 0 and $m_{\max}$. The notation $\cup p \in \{2, 3, \ldots, N\}$ denotes the union of probabilities that p sub-systems are assigned with identical priorities, where p can be any set of two, three, till N sub-systems. Note that a collision corresponding to a sub-system with assigned probability 0, i.e. if $g(e_k^i) \leq \lambda_i$, can only occur if there exists no other sub-system with higher priority. Recalling that $g(\cdot)$ is a continuous function and e_k^i is a continuous Gaussian random variable for all $i \in \{1, \ldots, N\}$, it follows that $g(e_k^i)$ is also a continuous random variable. Therefore, it is possible to compute the cumulative distribution function (CDF) and expected value of $g(e_k^i)$, which are respectively denote by $F_{g(e_k^i)}$, and $\bar{\mu}_{g(e_k^i)}$. In addition, $\lceil g(e_k^i) \rceil$ is a discrete random variable, with the probability mass function denoted by $f_{\lceil g(e_k^i) \rceil}$. Employing the inclusion-exclusion principle and Markov's inequality, we can find an upper-bound for the probability of collision, at one time-step k, as follows:

$$\begin{aligned}\mathbb{P}\left[\theta_k^i = 0, \forall i\right] &\leq \prod_{i=1}^{N} F_{g(e_k^i)}(\lambda_i) \\ &+ \sum_{p=2}^{N} (-1)^p \binom{N}{p} \left[\prod_{q=1}^{p} f_{\lceil g(e_k^q) \rceil}(\bar{m}_k)\right] \prod_{l=1}^{N-p} F_{g(e_k^l)}(\bar{m}_k) \\ &+ \sum_{p=2}^{N} (-1)^p \binom{N}{p} \frac{\prod_{q=1}^{p} \bar{\mu}_{g(e_k^q)}}{m_{\max}^p},\end{aligned} \tag{6.35}$$

where, $m_{\max} = 2^n - 1$, and $\binom{N}{p}$ represents the p-combination of the set of N sub-systems. Expression (6.35) clarifies that the number of resolution slots plays the crucial role in the collision rate. This is an expected observation as decreasing n decreases the number of distinct priority indexes, which consequently leads to higher collisions.

6.4.3 Stability Analysis

In this section, we study stability properties of the described multiple-loop NCSs under random access scheduling with the prioritization given in (6.33). Due to the existence of additive stochastic noise to system dynamics, we employ concepts of stochastic stability to investigate the asymptotic properties of the NCS. In addition, the decentralized nature of the medium access control implies that there exists non-zero probability of successive collisions at all sampling times. This means, it is theoretically possible that no

transmission is successful and all sub-systems operate in open-loop. This scenario leads to instability of the overall NCS in terms of Lyapunov m^{th}-mean if only one sub-system is unstable in open-loop. To that end, we employ the concept of Lyapunov Stability in Probability as previously defined in (6.11).

Recall that the aggregate state of sub-system i in presence of the communication constraint is $[x_k^{i\mathsf{T}}\ e_k^{i\mathsf{T}}]^{\mathsf{T}}$. In addition, we discussed that emulative controllers guarantees stability of each individual sub-system in the absence of the communication constraint, i.e. stabilizing gains L_i exist such that the closed-loop matrix $(A_i{-}B_iL_i)$ is stable. Therefore, having the independence of the error state e_k^i from the system state x_k^i, we solely investigate the convergence properties of the network-induced state e_k^i in order to show stability of sub-system i. This is summarized in the following lemma:

Lemma 6. *For a control loop i with state vector $[x_k^{i\mathsf{T}}\ e_k^{i\mathsf{T}}]^{\mathsf{T}}$, described in (6.28)-(6.29), the LSP condition (6.11) is equivalently satisfied if $\xi_i' > 0$ and $0 \leq \xi_i \leq \varepsilon_i$ exists such that*

$$\lim_{k\to\infty} \sup \mathbb{P}\Big[e_k^{i\mathsf{T}} e_k^i \geq \xi_i'\Big] \leq \xi_i. \tag{6.36}$$

Proof. The proof is analogous to that of Lemma 5. □

We define the overall NCS state as $[x_k^{\mathsf{T}}\ e_k^{\mathsf{T}}]^{\mathsf{T}}$, in which x_k and e_k contain local system states x_k^i and local error states e_k^i from all sub-systems $i \in \{1,\dots,N\}$, respectively. Note that within a local sub-system i, the control and scheduling laws generate the input signals for the local system state x_k^i and the local error state e_k^i independently form the other sub-systems $j \in \{1,\dots,N\} \setminus i$. Therefore, local stability of all sub-systems in the NCS guarantees stability of the overall NCS with the augmented state $[x_k^{\mathsf{T}}\ e_k^{\mathsf{T}}]^{\mathsf{T}}$. In terms of stochastic concept of Lyapunov stability in probability, it translates to existence of $\xi, \xi' > 0$ such that if $\lim_{k\to\infty} \sup \mathbb{P}\big[e_k^{\mathsf{T}} e_k \geq \xi'\big] \leq \xi$ holds, then exist $\varepsilon, \varepsilon' > 0$ such that $\lim_{k\to\infty} \sup \mathbb{P}\big[x_{k+1}^{\mathsf{T}} x_{k+1} \geq \varepsilon'\big] \leq \varepsilon$ holds. Knowing that the existence of stabilizing control gains L_i's guarantee that the augmented system state x_k is converging, the LSP condition for the overall networked system becomes

$$\lim_{k\to\infty} \sup \mathbb{P}\big[e_k^{\mathsf{T}} e_k \geq \xi'\big] \leq \xi. \tag{6.37}$$

Employing the Markov's inequality for the non-negative random variable $e_k^{\mathsf{T}} e_k$, we have

$$\mathbb{P}\left[e_k^{\mathsf{T}} e_k \geq \xi'\right] \leq \frac{\mathbb{E}\left[e_k^{\mathsf{T}} e_k\right]}{\xi'} = \frac{\sum_{i=1}^{N} \mathbb{E}\left[\|e_k^i\|^2\right]}{\xi'}. \tag{6.38}$$

In the following theorem, we show that the boundedness of $\mathbb{E}\left[e_k^{\mathsf{T}} e_k\right]$ guarantees that the LSP condition (6.37) holds.

Theorem 4. *Assume a multiple-loop NCS consists of N heterogeneous LTI stochastic sub-systems modeled by (6.24), sharing a communication channel subject to the constraint (6.30). Under the control, estimation laws (6.26) and (6.25), and random access prioritization law (6.33), the overall NCS with augmented state $[x_k^{\mathsf{T}}, e_k^{\mathsf{T}}]^{\mathsf{T}}$ is LSP for any positive λ_i's, continuous and strictly increasing function g, and $n \geq 2$.*

Proof. We address stability of the described NCS by the concept of LSP (see Definition 5). Due to the independence of error dynamics from the system state (see (6.28) and (6.29)), together with knowing that the system state is converging in the absence of communication constraints, we investigate convergence properties of the error state. According to the LSP condition (6.37), the probability that the NCS might not be stable equals the probability that successive collisions occur, such that a finite-length time interval cannot be found over which N transmissions successfully take place. Note that, due to the dynamic prioritization mechanism, it is not guaranteed that over such an interval all N sub-systems transmits exactly once. In addition, it is worth mentioning that showing the expression (6.37) holds over a finite-length time interval provides only a sufficient stability condition. The detailed proof is found in our previously published work [Mam+17]. □

Remark 12. *In Theorem 4, we consider the worst case situation by assuming that the backoff probability $p_b = 0$. As this probability is constant, the results can be similarly derived for the case $1 > p_b > 0$. It is also the case for computation of the collision probability in (6.34) and its upper-bound (6.35). Expectedly, taking into account non-zero backoff probability p_b reduces the probability of collision, and consequently tightens the stability margins, whereas limiting the performance in some cases due to additionally introduced delays.*

6.4.4 Numerical Results

For the evaluations, we consider an NCS comprised of N independent stochastic sub-systems, divided into two homogeneous classes I (unstable class) and II (stable class), with the following system matrices:

$$A_I = \begin{bmatrix} 1.25 & 0 \\ 0 & 1.1 \end{bmatrix}, \quad A_{II} = \begin{bmatrix} 0.9 & 0 \\ 0 & 0.9 \end{bmatrix}.$$

The input matrices are assumed to be identity, i.e., $B_I = B_{II} = I_{2\times 2}$. Each sub-system i is affected by the additive random Gaussian noise with covariance matrix $W_I = W_{II} = I_{2\times 2}$. Moreover, every sub-systems is controlled by a dead-beat control law $L_i = \frac{A_i}{B_i}$. We perform Monte Carlo simulations and plot the averages over 30 runs with 95% confidence intervals. The number of sub-systems is varied within $N \in [2, 60]$, where each class I and II contains $\frac{N}{2}$ sub-systems. We define the positive, continuous, and strictly increasing function $g(e_k^i)$, introduced in (6.33), as

$$g\left(e_k^i\right) = \|e_k^i\|. \tag{6.39}$$

The number of resolution slots is set to $n = 12$ (according to common values for power-line communication [Geh+14]), hence, the number of priority levels equals $m_{\max} = 4095$.

6.4.4.1 Performance Evaluation

We compare the performance of our proposed protocol (denoted as PRIO in the figures) with some of the common scheduling schemes, such as Time Division Multiple Access (TDMA), TOD, and slotted ALOHA with the optimal channel access probability $p_b = \frac{1}{N}$ [Riv87] (denoted RA). We consider two variations of TDMA: (1) full round-robin scheme wherein every node transmits without contention every N^{th} time-step, denoted by TDMA in the figures, and (2) reduced round-robin wherein only open-loop unstable sub-systems (class I) transmit every $N/2$ step, which is denoted by TDMA(U). To simulate centralized TOD scheme, we assume that, at each time-step, only the sub-system with the highest error norm $\|e_k^i\|$ transmits. For PRIO we assume no access barring, i.e. $p_b = 0$. Note that TOD approach has to be implemented in centralized fashion, hence, it requires additional communication resources to communicate with the centralized scheduler. If such a channel is not present, TOD cannot be implemented at all. To evaluate the efficiency of each scheme, we define two performance metrics: the average error norm $\bar{E}$ defined as

$$\bar{E} \;=\; \frac{1}{N t_{\max}} \sum_{k=0}^{t_{\max}-1} \sum_{i=1}^{N} \|e_k^i\|, \tag{6.40}$$

and the average collision rate p_{coll} defined as the ratio of time-steps where no transmission occurred due to collision to all time-steps.

Figs. 6.3(a) and (b) illustrate the performance comparisons. From (a), we observe that for low number of sub-systems, up to $N = 4$, our prioritized protocol achieves comparable error norm as with TDMA, TDMA(U) and TOD. For higher number of sub-systems though, our proposed protocol outperforms TDMA, achieving up to 50 times lower average error norm. Expectedly, TDMA(U) performs better than TDMA, however PRIO starts outperforming TDMA(U) for $N > 20$, and the performance gap increases by increasing N. The centralized TOD approach is depicted here as the lower bound, achieving the best performance. Classical random access, even with optimal Bayesian back-off scheme, only delivers acceptable performance for $N = 2$.

From Fig. 6.3(b), we observe that the collision rate for the PRIO is higher than in the classical random access as long as $N < 14$. We also observe that collision rates for RA saturate at ≈ 0.26 for high N. This saturation is due to the optimal back-off choice $p_b = 1/N$. Interestingly, collision rates for PRIO start to decline for $N > 4$. This effect is explained by higher errors for larger number of sub-systems, and, hence, higher variations in the priority levels. TOD and TDMA are contention-free protocols and are not depicted in Fig. 6.3(b).

6.4.4.2 Impact of Protocol Overhead

In order to determine which protocol is more beneficial to be employed, we suggest a joint control/communication-related metric J, which incorporates the error of each local

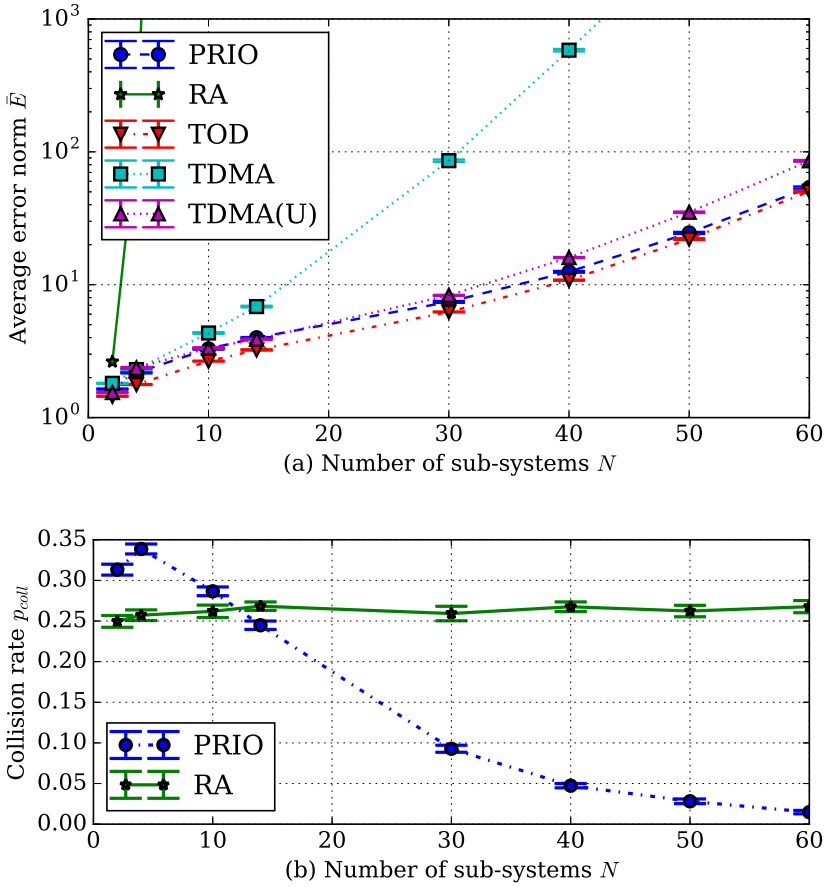

Figure 6.3: (a) Average error norm $\bar{E}$ and (b) collision rate p_{coll} vs. number of sub-systems N for random access with back-offs (RA), proposed prioritized contention resolution (PRIO), try-once-discard (TOD) and schedule-based access (TDMA). Average values over 30 runs with 95% confidence intervals.

control system, and the so called "cost of the protocol"

$$J = \frac{1}{N t_{\max}} \sum_{k=0}^{t_{\max}-1} \sum_{i=1}^{N} \|e_k^i\|(1+\alpha_k^i), \tag{6.41}$$

where α_k^i denotes the relative overhead ("cost") of the protocol. For TOD, it is defined as the ratio of resources needed to implement the centralized decision, i.e., to deliver the error information from every sub-system i at every time-step k towards the central coordinator. Assuming that the transmission of $\|e_k^i\|$ takes time T_o^i, we define $\alpha_k^i = T_o^i/T_s$. Similarly, for the proposed contention resolution protocol, overhead is defined by the ratio of resources consumed by the contention resolution slots $\alpha_k^i = (nT_{\text{CR}})/(NT_s)\ \forall i,k$. We divide here by N, because the slots are used by all sub-systems equally. For TDMA and random access with optimal back-off, $\alpha_k^i = 0$ (we assume that the back-off dimensioning and TDMA schedule allocation is done off-line, and neglect its overhead). Consequently, the term $\|e_k^i\|(1+\alpha_k^i)$, intuitively, denotes the error *weighted by the cost of the protocol* at

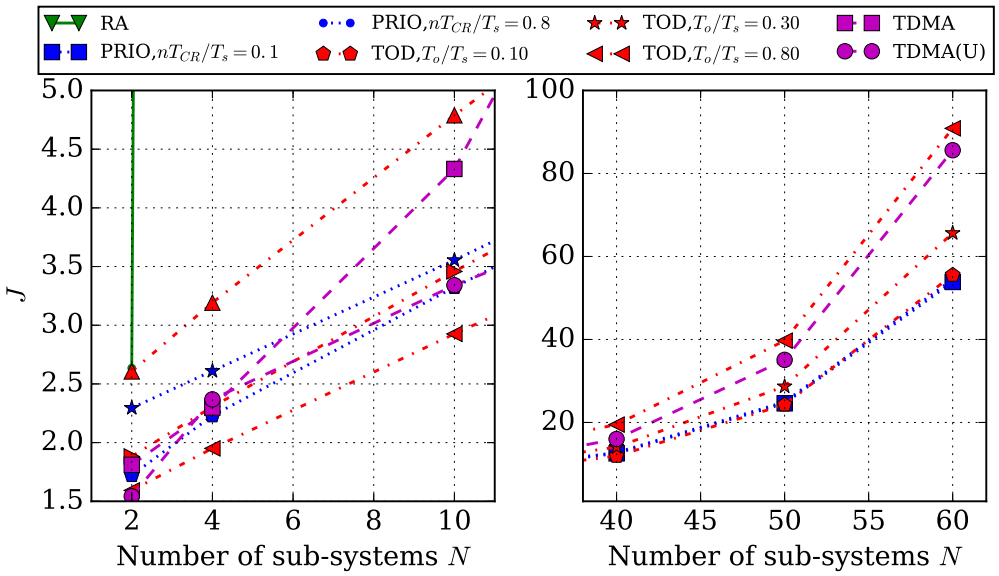

Figure 6.4: J vs. number of sub-systems N for random access with back-offs (RA), proposed prioritized contention resolution (PRIO), try-once-discard (TOD) and schedule-based access (TDMA).

a given time-step. Having α_k^i defined and assuming that TOD overhead is homogeneous among sub-systems, i.e., $T_o^i = T_o \ \forall i \in [1, N]$, we obtain:

$$\begin{aligned} J^{\mathrm{RA}} = J^{\mathrm{TDMA}} &= \frac{1}{N t_{\max}} \sum_{k=0}^{t_{\max}-1} \sum_{i=1}^{N} \|e_k^i\|, \\ J^{\mathrm{TOD}} &= \frac{T_o}{N T_s t_{\max}} \sum_{k=0}^{t_{\max}-1} \sum_{i=1}^{N} \|e_k^i\|, \\ J^{\mathrm{PCR}} &= \frac{n T_{\mathrm{CR}}}{N^2 T_s t_{\max}} \sum_{k=0}^{t_{\max}-1} \sum_{i=1}^{N} \|e_k^i\|. \end{aligned}$$

Fig. 6.4 depicts the exemplary values of J for $N \in \{2, 4, 10, 14, 20\}$, and exemplary overhead values for TOD, $T_o/T_s \in \{0.1, 0.3, 0.8\}$ and priority-based protocol $nT_{\mathrm{CR}}/T_s \in \{0.1, 0.8\}$. We first observe that, despite no overhead, TDMA only performs well until $N \leq 10$, and RA is only suitable for $N = 2$. TDMA(U) is preferable than TOD with $T_o/T_s = 0.8$ even for large number of sub-systems. The overhead of PRIO has a distinct effect on J only for low number of sub-systems, and it scales gently with increasing N. Intuitively, adding more sub-systems does not increase the number of contention resolution slots n, hence, the overhead per sub-system decreases. On the contrary, overhead of TOD is increasing linearly with N. As a consequence, for $T_o/T_s \geq 0.3$ (30% overhead for centralized decision taking) and $N > 10$, PRIO results in lower joint cost than TOD, hence, is preferable to use.

6.5 Summary and Discussion

In this chapter, we have considered a detailed model of Networked Control Systems, an important class of M2M applications. We have studied the performance of an NCS with individual sensor to controller communications taking place over a shared medium. Multi-channel ALOHA and binary countdown control resolution protocols have been considered alongside with event-based local control strategy in Secs. 6.3 and 6.4, respectively.

In Sec. 6.3, we have demonstrated that there exists a global threshold value minimizing the average error variance. We have further introduced a local resource-aware scheduler design with an adaptive choice of the error threshold based on knowledge of the network state, and numerically demonstrated that by deploying it instead of the static threshold we can increase the control performance. Future work in this direction must aim at finding the analytical relation between the network state and the optimal transmission threshold or its close approximation.

In Sec. 6.4, we have proposed a practical state-dependent contention resolution mechanism for multi-loop NCSs with random access medium such as wireless, bus, or powerline networks. According to an error dependent measure, the priorities are deterministically assigned to each sub-system at every time-step, and the highest priority sub-system sends its data packet. It is shown that the proposed state-dependent prioritization preserves stability of the overall NCS in terms of Lyapunov stability in probability. Simulation results validate stability claim and illustrate achieved performance enhancement by the proposed error-dependent approach compared to the related protocols. Moreover, our approach performs closely to the centralized TOD approach in terms of average error variance. A joint control-communication metric is also introduced which can be used to select the appropriate protocol depending on the size of an NCS.

The results of the chapter should be viewed as an exemplary approach to the cross-layer design problem. We have presented a joint analysis of the system with NCS and medium access protocols. Furthermore, we have shown the potential benefits of a cross-layer approach, by designing adaptive decentralized scheduler in Sec. 6.3, and adaptive state-dependent prioritization in Sec. 6.4. Future works in the area should target development of a systematic co-design framework, accounting for various MAC protocols and communication medium specifics.

Chapter 7

Conclusions and Outlook

In this thesis, we have studied random access protocols for Machine-to-Machine (M2M) communications. We have proposed several approaches for modeling, performance analysis and optimization of the protocols to accommodate massive Machine-to-Machine (mM2M) and ultra reliable Machine-to-Machine (uM2M) applications in 5G wireless networks. With this chapter, we conclude the thesis by summarizing main results and presenting an outlook for future work.

7.1 Summary and Discussion

In Chapter 3, we have targeted the improvement of the steady-state performance of Random Access Procedure (RAP). We have analyzed how throughput, request drop ratio and delay depend on the number of allocated resources, and analytically found the optimal resource allocation maximizing the throughput. On the example of two Quality of Service (QoS) classes, we have studied how separation of the resources impacts the performance of User Equipments (UEs) from both classes. Based on the analytical insights, we have developed Load-Adaptive Throughput MAximizing Preamble Allocation (LATMAPA), an algorithm to prioritize QoS classes using the resource allocation based on their current load. We compared LATMAPA to prioritization approaches from the state of the art, and demonstrated its superior performance with respect to the achieved throughput and request drop ratio. In the second part of the chapter, we have studied aggregation as a technique to enhance RAP performance in the high load regimen. We have analytically modeled the system where the connection requests from UEs are aggregated on the intermediate nodes before being forwarded to the Next Generation Node B (gNB). From the analytical model, we have derived the steady-state throughput, delay, and drop ratio, and demonstrated the delay trade-off of the aggregation process.

In Chapter 4, we have targeted the improvement of the transient performance of RAP. We have considered a scenario of burst arrival of connection request from large amount of UEs, causing long lasting overload in the Random Access CHannel (RACH). First, we have defined the uplink resource consumption of the four-way handshake random access, and demonstrated the difference to the two-way handshake in the classical multi-channel slotted ALOHA protocol. Then, we have compared two ways of resource-aware throughput optimization of random access: by considering resource efficiency as a composite met-

ric, or by considering Pareto-optimal solution. We have concluded that Pareto-optimal solution is superior, since it is solving a bi-objective optimization problem, maximizing throughput and minimizing resource consumption at the same time. We have further devised a jointly optimal channel allocation – access barring algorithm to deliver solutions from the Pareto set. Finally, we have proposed to aid RAP with an additional stage of Binary Countdown Contention Resolution (BCCR) to reduce collision rates. Based on the proposal and previously introduced bi-objective optimization framework, we have developed Dynamic Binary Countdown - Access barring (DBCA) algorithm, to reduce the burst resolution delay and improve efficiency of RAP.

In Chapter 5, in contrast to the previous chapters where mM2M applications are the primary target, we have focused on the reliability of the random access protocols for uM2M applications. For the burst arrival scenario as in Chapter 4, we have proposed a methodology to analyze the performance of the RAP with respect to latency–reliability profile. The methodology answers the question: What latency can be achieved for a given reliability requirement? In the proposed methodology, RAP is modeled as a queuing system, and its probabilistic performance bounds are derived using stochastic network calculus. We have demonstrated how the methodology is applied on the example of static and dynamic Access Class Barring algorithm for burst resolution.

In Chapter 6, we have considered Networked Control Systems (NCSs) as an exemplary class of M2M applications. We have studied the performance of NCSs in terms of the estimation error, where individual control sub-systems are sharing wireless communication medium. We have assumed that the access to the medium is regulated by single- or multi-channel slotted ALOHA protocols. We have illustrated that, due to stochastic nature of the protocol, the performance of event-triggered sub-systems is coupled, and there exists a globally optimal event-triggering policy. We use this insight to propose a simple threshold policy adaptive to the amount of resources offered by the network. In the second part of the chapter, we additionally utilize BCCR, and develop a state-dependent priority assignment for BCCR, to reduce collision rates and improve NCS estimation error on the same time.

7.2 Directions for Future Work

The results of the thesis point out multiple possible future work opportunities. To address the needs of mM2M devices, other scalability bottlenecks besides the connection establishment procedure must be studied. Some functionality of the core network becomes a bottleneck in the presence of the massive amount of devices. Authentication procedures, mobility management, or bearer establishment are associated with significant overhead, and hence become increasingly inefficient in the case of mM2M small data transmissions. Approaches to circumvent or simplify these procedures will help in creating scalable networks for mM2M. Similarly, instead of improving the connection establishment procedure, one can imagine an alternative approach, where the focus is on simplifying the maintenance of already captured resources, thus reducing the need

and frequency of connection establishment. There are currently few notable research initiatives in this direction, in particular connection-less communication and grant-free access [3GP17a].

Addressing the needs of uM2M devices requires major modifications throughout the protocol stack, not only the medium access protocols. Ensuring millisecond-range latencies in 5G networks requires in the first place changes in the frame structure of LTE to allow smaller transmission slots (mini-slots) and flexible frame structure [Ped+16]. On the other hand, high reliability requires the introduction of redundancy and retransmissions, via time, frequency or interface diversity. This trade-off of latency and reliability is complex, and its careful evaluation requires novel approaches modeling and analysis. Moreover, at the moment, reliable communication over the wireless link is a trending research topic [Pop+17], whereas the question of how to ensure end-to-end reliability is rarely addressed in the literature [Zop+18] and remains largely open. Full provisioning of QoS in terms of latency and reliability requires careful study of the interplay of the different parts in the end-to-end chain, including both wireless and wired network domains. Developing methodologies for analysis and optimization of end-to-end performance is an open and promising research topic to be addressed on the way to uM2M support in 5G.

Bibliography

Publications by the author

Book chapters

[G¨+17a] H. M. Gürsu, M. Vilgelm, E. Fazli, and W. Kellerer. "A Medium-access Approach to Wireless Technologies for Reliable Communication in Aircraft". In: *Wireless Sensor Systems for Extreme Environments Wireless Sensor Systems for Extreme Environments: Space, Underwater, Underground and Industrial*. Wiley Online Library, 2017, pp. 431–452. DOI: 10.1002/9781119126492.ch20.

Journal publications

[G¨+17b] H. M. Gürsu, M. Vilgelm, W. Kellerer, and M. Reisslein. "Hybrid Collision Avoidance-Tree Resolution for M2M Random Access". In: *IEEE Transactions on Aerospace and Electronic Systems* 53.4 (Mar. 2017), pp. 1974–1987. ISSN: 0018-9251. DOI: 10.1109/TAES.2017.2677839.

[Gür+19a] H. M. Gürsu, M. Vilgelm, A. Martinez Alba, M. Berioli, and W. Kellerer. "Admission Control Based Traffic-Agnostic Delay-Constrained Random Access (AC/DC-RA) for M2M Communication". In: *IEEE Transactions on Wireless Communications* 18.5 (May 2019), pp. 2858–2871. ISSN: 1536-1276. DOI: 10.1109/TWC.2019.2908914.

[Vil+17b] M. Vilgelm, H. M. Gürsu, W. Kellerer, and M. Reisslein. "LATMAPA: Load-Adaptive Throughput-MAximizing Preamble Allocation for Prioritization in 5G Random Access". In: *IEEE Access* 5 (Jan. 2017), pp. 1103–1116. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2017.2651170.

[VRK19a] M. Vilgelm, S. Rueda Liñares, and W. Kellerer. "Dynamic Binary Countdown for Massive M2M Random Access in Dense 5G Networks". In: *IEEE Internet of Things Journal* 6.4 (Apr. 2019), pp. 6896–6908. DOI: 10.1109/JIOT.2019.2912424.

[VRK19b] M. Vilgelm, S. Rueda Liñares, and W. Kellerer. "On the Resource Consumption of M2M Random Access: Efficiency and Pareto Optimality". In: *IEEE Wireless Communications Letters* 8.3 (Dec. 2019), pp. 709–712. ISSN: 2162-2337. DOI: 10.1109/LWC.2018.2886892.

[Zop+18] S. Zoppi, A. Van Bemten, H. M. Gürsu, M. Vilgelm, J. Guck, et al. "Achieving Hybrid Wired/Wireless Industrial Networks with WDetServ: Reliability-Based Scheduling for Delay Guarantees". In: *IEEE Transactions on Industrial Informatics* 14.5 (Feb. 2018), pp. 2307–2319. DOI: 10.1109/TII.2018.2803122.

Conference publications

[Aya+19] O. Ayan, M. Vilgelm, M. Klügel, S. Hirche, and W. Kellerer. "Age-of-Information vs. Value-of-Information Scheduling for Cellular Networked Control Systems". In: *Proc. ACM/IEEE International Conference on Cyber-Physical Systems*. ACM/IEEE, Apr. 2019. DOI: 10.1145/3302509.3311050.

[G¨16] H. M. Gürsu, M. Vilgelm, S. Zoppi, and W. Kellerer. "Reliable Coexistence of 802.15.4e TSCH-based WSN and Wi-Fi in an Aircraft Cabin". In: *IEEE ICC2016-Workshops: W13-Second Workshop on Advanced PHY and MAC Technology for Super Dense Wireless Networks (CROWD-Net)*. May 2016. DOI: 10.1109/ICCW.2016.7503863.

[Gri+17] E. Grigoreva, M. Laurer, M. Vilgelm, T. Gehrsitz, and W. Kellerer. "Coupled Markovian Arrival Process for Automotive Machine Type Communication traffic modeling". In: *2017 IEEE International Conference on Communications (ICC)*. May 2017, pp. 1–6. DOI: 10.1109/ICC.2017.7996498.

[Gür+19b] H. M. Gürsu, C. Moroglu, M. Vilgelm, F. Clazzer, and W. Kellerer. "System Level Integration of Irregular Repetition Slotted ALOHA for Industrial IoT in 5G New Radio". In: *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. Oct. 2019. DOI: 10.1109/PIMRC.2019.8904385.

[JVK19] A. Jacquelin, M. Vilgelm, and W. Kellerer. "Grant-Free Access with Multipacket Reception: Analysis and Reinforcement Learning Optimization". In: *Proc. IEEE/IFIP Wireless On-demand Network systems and Services Conference (WONS)*. IEEE, Jan. 2019, pp. 1–8. DOI: 10.23919/WONS.2019.8795459.

[Mam+17] M. Mamduhi, M. Vilgelm, W. Kellerer, and S. Hirche. "Prioritized Contention Resolution for Random Access Networked Control Systems". In: *Proc. IEEE Conference on Decision and Control (CDC)*. Dec. 2017. DOI: 10.1109/CDC.2017.8264667.

[Sol+18] T. Soleymani, S. Zoppi, M. Vilgelm, S. Hirche, W. Kellerer, et al. "Covariance-Based Transmission Power Control for Estimation over Wireless Sensor Networks". In: *European Control Conference (ECC)*. June 2018. DOI: 10.23919/ECC.2018.8550129.

[SVK17] V. Schrader, M. Vilgelm, and W. Kellerer. "On Random Access Channel Performance and M2M Support in Standalone LTE Unlicensed". In: *Proc. IEEE Global Communications Conference (GLOBECOM)*. Dec. 2017. DOI: 10.1109/GLOCOM.2017.8254689.

[Vil+16a] M. Vilgelm, M. H. Mamduhi, W. Kellerer, and S. Hirche. "Adaptive Decentralized MAC for Event-Triggered Networked Control Systems". In: *Proc. ACM International Conference on Hybrid Systems: Computation and Control (HSCC)*. HSCC '16. Vienna, Austria: ACM, 2016, pp. 165–174. ISBN: 978-1-4503-3955-1. DOI: 10.1145/2883817.2883829.

[Vil+16b] M. Vilgelm, H. M. Gürsu, S. Zoppi, and W. Kellerer. "Time Slotted Channel Hopping for Smart Metering: Measurements and Analysis of Medium Access". In: *Proc. IEEE International Conference on Smart Grid Communications (SmartGridComm)*. Nov. 2016, pp. 109–115. DOI: 10.1109/SmartGridComm.2016.7778747.

[Vil+17a] M. Vilgelm, O. Ayan, S. Zoppi, and W. Kellerer. "Control-aware Uplink Resource Allocation for Cyber-Physical Systems in Wireless Networks". In: *European Wireless 2017; 23th European Wireless Conference*. May 2017.

[Vil+18] M. Vilgelm, S. Schiessl, H. Al-Zubaidy, W. Kellerer, and J. Gross. "On the Reliability of LTE Random Access: Performance Bounds for Machine-to-Machine Burst Resolution Time ". In: *Proc. IEEE International Conference on Communications (ICC)*. May 2018. DOI: 10.1109/ICC.2018.8422323.

[VK17a] M. Vilgelm and W. Kellerer. "Binary Contention Resolution for M2M Random Access Prioritization in LTE-A and 5G". In: *Proc. IFIP Networking Conference (Poster and Demo Session)*. June 2017. DOI: 10.23919/IFIPNetworking.2017.8264872.

[VK17b] M. Vilgelm and W. Kellerer. "Impact of Request Aggregation on Machine Type Connection Establishment in LTE-Advanced". In: *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*. Mar. 2017, pp. 1–6. DOI: 10.1109/WCNC.2017.7925664.

[VRK17] M. Vilgelm, S. Rueda Liñares, and W. Kellerer. "Enhancing Cellular M2M Random Access with Binary Countdown Contention Resolution". In: *Proc. IEEE International Symposium on Personal, Indoor and Mobile Radio Communications (PIMRC)*. Oct. 2017. DOI: 10.1109/PIMRC.2017.8292435.

[Zop+17] S. Zoppi, H. M. Gürsu, M. Vilgelm, and W. Kellerer. "Reliable Hopping Sequence Design for Highly Interfered Wireless Sensor Networks". In: *2017 IEEE International Symposium on Local and Metropolitan Area Networks (LANMAN)*. June 2017, pp. 1–7. DOI: 10.1109/LANMAN.2017.7972164.

General publications

[3GP11] 3GPP. *Technical Specification Group Radio Access Network; Study on RAN Improvements for Machine-type Communications*. TR 37.868. 3rd Generation Partnership Project (3GPP), Sept. 2011.

[3GP12] 3GPP. *Technical Specification Group Services and System Aspects; System improvements for Machine-Type Communications (MTC) (Release 11) V11.0.0*. TR 23.888. 3rd Generation Partnership Project (3GPP), Sept. 2012.

[3GP17a] 3GPP. *Technical Specification Group Radio Access Network; Study on New Radio Access Technology Physical Layer Aspects (Release 14 V14.0.1*. TR 38.802. 3rd Generation Partnership Project (3GPP), Aug. 2017.

[3GP17b] 3GPP. *Technical Specification Group Services and System Aspects; Service requirements for Machine-Type Communications (MTC); Stage 1 (Release 14 V14.0.1*. TR 23.368. 3rd Generation Partnership Project (3GPP), Aug. 2017.

[3GP18a] 3GPP. *Technical Specification Group Radio Access Network; New Radio; Physical channels and modulation (Release 15);* TS 38.211. 3rd Generation Partnership Project (3GPP), 2018.

[3GP18b] 3GPP. *Technical Specification Group Services and System Aspects; Study on Communication for Automation in Vertical Domains (Release 16) V16.0.0*. TR 22.804. 3rd Generation Partnership Project (3GPP), June 2018.

[AAF16] T. P. de Andrade, C. A. Astudillo, and N. L. da Fonseca. "Allocation of Control Resources for Machine-to-Machine and Human-to-Human Communications Over LTE/LTE-A Networks". In: *IEEE Internet of Things Journal* 3.3 (2016), pp. 366–377.

[Abr70] N. Abramson. "The ALOHA System: another alternative for computer communications". In: *Proc. Fall Joint Computer Conference (AFIPS)*. ACM. 1970, pp. 281–285.

[AG17] Z. Alavikia and A. Ghasemi. "Overload control in the network domain of LTE/LTE-A based machine type communications". In: *Wireless Networks, in print* (2017), pp. 1–16.

[AHK17] M. S. Ali, E. Hossain, and D. I. Kim. "LTE/LTE-A random access for massive machine-type communications in smart cities". In: *IEEE Communications Magazine* 55.1 (2017), pp. 76–83.

[Aij+17] A. Aijaz, M. Dohler, A. H. Aghvami, V. Friderikos, and M. Frodigh. "Realizing the tactile internet: Haptic communications over next generation 5G cellular networks". In: *IEEE Wireless Communications* 24.2 (2017), pp. 82–89.

[AK16] O. Arouk and A. Ksentini. “General model for RACH procedure performance analysis”. In: *IEEE Communications Letters* 20.2 (2016), pp. 372–375.

[And+15] S. Andreev, O. Galinina, A. Pyattaev, M. Gerasimenko, T. Tirronen, et al. “Understanding the IoT connectivity landscape: a contemporary M2M radio technology roadmap”. In: *IEEE Communications Magazine* 53.9 (2015), pp. 32–40.

[Ara+13] G. Araniti, C. Campolo, M. Condoluci, A. Iera, and A. Molinaro. “LTE for vehicular networking: a survey”. In: *IEEE Communication Magazine* 51.5 (May 2013), pp. 148–157. ISSN: 0163-6804.

[Aug+16] A. Augustin, J. Yi, T. Clausen, and W. M. Townsley. “A study of LoRa: Long range & low power networks for the Internet of Things”. In: *Sensors* 16.9 (2016), p. 1466.

[AZLB13] H. Al-Zubaidy, J. Liebeherr, and A. Burchard. “A $(\min, \times)$ network calculus for multi-hop fading channels”. In: *Proc. IEEE International Conference on Computer Communications (INFOCOM)*. IEEE. 2013, pp. 1833–1841.

[BA11a] R. Blind and F. Allgöwer. “Analysis of Networked Event-Based Control with a Shared Communication Medium: Part I - Pure ALOHA”. In: *Proc. IFAC World Congress*. 2011.

[BA11b] R. Blind and F. Allgöwer. “Analysis of networked event-based control with a shared communication medium: Part II-Slotted ALOHA”. In: *Proc. IFAC World Congress*. 2011, pp. 8830–8835.

[Bai+17] A. Baiocchi, I. Tinnirello, D. Garlisi, and A. Lo Valvo. “Random Access with Repeated Contentions for Emerging Wireless Technologies”. In: *Proc. IEEE International Conference on Computer Communications (INFOCOM)*. IEEE. 2017, pp. 1–9.

[BGH87] D. P. Bertsekas, R. G. Gallager, and P. Humblet. *Data networks*. Vol. 2. Prentice-hall Englewood Cliffs, NJ, 1987.

[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.

[BJR17] Y. D. Beyene, R. Jäntti, and K. Ruttik. “Random Access Scheme for Sporadic Users in 5G”. In: *IEEE Transactions on Wireless Communications* 16.3 (Mar. 2017), pp. 1823–1833. ISSN: 1536-1276.

[Boc+14] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski. “Five disruptive technology directions for 5G”. In: *IEEE Communications Magazine* 52.2 (Feb. 2014), pp. 74–80. ISSN: 0163-6804.

[Boc+16] C. Bockelmann, N. Pratas, H. Nikopour, K. Au, T. Svensson, et al. "Massive machine-type communications in 5G: physical and MAC-layer solutions". In: *IEEE Communications Magazine* 54.9 (Sept. 2016), pp. 59–65. ISSN: 0163-6804.

[Cap77] J. I. Capetanakis. "The multiple access broadcast channel: protocol and capacity considerations." PhD thesis. Massachusetts Institute of Technology, 1977.

[Cap79] J. Capetanakis. "Tree algorithms for packet broadcast channels". In: *IEEE Transactions on Information Theory* 25.5 (1979), pp. 505–515.

[Cen+16] M. Centenaro, L. Vangelista, A. Zanella, and M. Zorzi. "Long-Range Communications in Unlicensed Bands: the Rising Stars in the IoT and Smart City Scenarios". In: *IEEE Wireless Communications* 23 (Oct. 2016).

[CGH07] E. Casini, R. D. Gaudenzi, and O. Herrero. "Contention resolution diversity slotted ALOHA (CRDSA): An enhanced random access schemefor satellite access packet networks". In: *IEEE Transactions on Wireless Communications* 6.4 (2007).

[CH08] A. Cervin and T. Henningsson. "Scheduling of event-triggered controllers on a shared network". In: *Proc. IEEE Conference on Decision and Control*. Dec. 2008, pp. 3601–3606.

[Che+15] R.-G. Cheng, J. Chen, D.-W. Chen, and C.-H. Wei. "Modeling and analysis of an extended access barring algorithm for machine-type communications in LTE-A networks". In: *IEEE Transactions on Wireless Communications* 14.6 (2015), pp. 2956–2968.

[Cho+11] S. Choi, W. Lee, D. Kim, K.-J. Park, S. Choi, et al. "Automatic configuration of random access channel parameters in LTE systems". In: *Proc. IFIP Wireless Days*. Oct. 2011, pp. 1–6.

[Chr+14] D. Christmann, R. Gotzhein, S. Siegmund, and F. Wirth. "Realization of Try-Once-Discard in Wireless Multihop Networks". In: *IEEE Transactions on Industrial Informatics* 10.1 (2014), pp. 17–26.

[Chu+15] Y.-Y. Chu, R. Harwahyu, R.-G. Cheng, and C.-H. Wei. "Study of generalized resource allocation scheme for multichannel slotted ALOHA systems". In: *Proc. IEEE Int. Symp. on Personal, Indoor and Mobile Radio Communications (PIMRC)*. 2015, pp. 1702–1706.

[Ciu+14] F. Ciucu, R. Khalili, Y. Jiang, L. Yang, and Y. Cui. "Towards a system theoretic approach to wireless network capacity in finite time and space". In: *Proc. IEEE International Conference on Computer Communications (INFOCOM)*. IEEE. 2014, pp. 2391–2399.

[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 Global Communications Conference (GLOBECOM) Workshops*. Dec. 2011, pp. 368–372.

[Con+16a] M. Condoluci, G. Araniti, T. Mahmoodi, and M. Dohler. "Enabling the IoT machine age with 5G: Machine-type multicast services for innovative real-time applications". In: *IEEE Access* 4 (2016), pp. 5555–5569.

[Con+16b] M Condoluci, G Araniti, M Dohler, A Iera, and A Molinaro. "Virtual code resource allocation for energy-aware MTC access over 5G systems". In: *Ad Hoc Networks* 43 (2016), pp. 3–15.

[Con+16c] M. Condoluci, M. Dohler, G. Araniti, A. Molinaro, and J. Sachs. "Enhanced radio access and data transmission procedures facilitating industry-compliant machine-type communications over LTE-based 5G networks". In: *IEEE Wireless Communications* 23.1 (2016), pp. 56–63.

[Cox12] C. Cox. *An introduction to LTE: LTE, LTE-advanced, SAE and 4G mobile communications*. John Wiley & Sons, 2012.

[CS88] I. Cidon and M. Sidi. "Conflict multiplicity estimation and batch resolution algorithms". In: *IEEE Transactions on Information Theory* 34.1 (Jan. 1988), pp. 101–110. ISSN: 0018-9448.

[CY+14] C. Chun-Yuan, Y.-H. Chen, Y.-X. Zheng, and F. Yu-Chuan. *Prioritized random access method*. US Patent 8,705,352. Apr. 2014.

[CYZ03] M. C. Chuah, O.-C. Yue, and Q. Zhang. *Methods and apparatus for random backoff based access priority in a communications system*. US Patent 6,594,240. July 2003.

[Dai+15] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, et al. "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends". In: *IEEE Communications Magazine* 53.9 (2015), pp. 74–81.

[DFJ12] D. Dimarogonas, E. Frazzoli, and K. Johansson. "Distributed Event-Triggered Control for Multi-Agent Systems". In: *IEEE Transactions on Automatic Control* 57.5 (2012), pp. 1291 –1297.

[DJ09] D. Dimarogonas and K. Johansson. "Event-triggered control for multi-agent systems". In: *Prof. IEEE Conference on Decision and Control, jointly with Chinese Control Conference (CDC/CCC)*. Dec. 2009, pp. 7131–7136.

[DKP16] G. Durisi, T. Koch, and P. Popovski. "Toward massive, reliable, and low-latency wireless communication with short packets". In: *Proceedings of the IEEE* 104.9 (2016), pp. 1711–1726.

[DOK15] D. Drajić, N. Ognjanović, and S. Krčo. "Architecture and Standards for M2M Communications". In: *Machine-to-Machine (M2M) Communications: Architecture, Technology, Standards, and Applications*. Ed. by V. Mišić and J. Mišić. Taylor & Francis Group, 2015, pp. 31–56.

[Don+12] M. Donkers, W. Heemels, D Bernardini, A Bemporad, and V Shneer. "Stability analysis of stochastic networked control systems". In: *Automatica* 48.5 (2012), pp. 917–925.

[DP+17] E. De Poorter, J. Hoebeke, M. Strobbe, I. Moerman, S. Latré, et al. "Sub-GHz LPWAN network coexistence, management and virtualization: an overview and open research challenges". In: *Wireless Personal Communications* 95.1 (2017), pp. 187–213.

[DPS18] E. Dahlman, S. Parkvall, and J. Skold. *5G NR: The next generation wireless access technology*. Academic Press, 2018.

[DS+15] M. De Sanctis, E. Cianca, G. Araniti, I. Bisio, and R. Prasad. "Satellite communications supporting internet of remote things". In: *IEEE Internet of Things Journal* 3.1 (2015), pp. 113–123.

[Du+16] Q. Du, W. Li, L. Liu, P. Ren, Y. Wang, et al. "Dynamic RACH Partition for Massive Access of Differentiated M2M Services". In: *Sensors* 16.4 (2016), p. 455.

[Dua+16] S. Duan, V. Shah-Mansouri, Z. Wang, and V. W. Wong. "D-ACB: Adaptive congestion control algorithm for bursty M2M traffic in LTE networks". In: *IEEE Transactions on Vehicular Technology* 65.12 (2016), pp. 9847–9861.

[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* 83.2 (2015), pp. 1251–1272.

[ETS11] ETSI. *Technical Specification; Machine-to-Machine communications; Functional architecture; V1.1.1*. TS 102.690. European Telecommunications Standards Institute (ETSI), Oct. 2011.

[Fid06] M. Fidler. "An end-to-end probabilistic network calculus with moment generating functions". In: *Proc. IEEE Int. Workshop on Quality of Service (IWQoS)*. IEEE. 2006, pp. 261–270.

[Fid10] M. Fidler. "Survey of deterministic and stochastic service curve models in the network calculus". In: *IEEE Communications Surveys & Tutorials* 12.1 (Jan. 2010), pp. 59–86. ISSN: 1553-877X.

[Fod+16] G. Foddis, R. G. Garroppo, S. Giordano, G. Procissi, S. Roma, et al. "On RACH preambles separation between human and machine type communication". In: *Proc. IEEE International Conference on Commmunications (ICC)*. 2016, pp. 1–6.

[GAK17a] H. M. Gürsu, A. M. Alba, and W. Kellerer. "Slotted ALOHA Filtered Tree (SAFT) for a Reliable LTE RACH". In: *Proc. European Wireless Conference*. May 2017, pp. 1–7.

[GAK17b] H. M. Gürsu, A. M. Alba, and W. Kellerer. "Delay Analysis of Multichannel Parallel Contention Tree Algorithms (MP-CTA)". In: *Computer Research Repository (CoRR)* abs/1707.09754 (2017). arXiv: 1707.09754.

[Gaz17] V. Gazis. "A Survey of Standards for Machine-to-Machine and the Internet of Things". In: *IEEE Communications Surveys & Tutorials* 19.1 (2017), pp. 482–511. ISSN: 1553-877X.

[GC10] R. A. Gupta and M. Y. Chow. "Networked Control System: Overview and Research Trends". In: *IEEE Transactions on Industrial Electronics* 57.7 (2010), pp. 2527–2535. ISSN: 0278-0046.

[Geh+14] T. Gehrsitz, R. Durner, H. Kellermann, H.-T. Lim, and W. Kellerer. "Priority-based energy-efficient MAC protocols for the in-car power line communication". In: *Proc. IEEE Vehicular Networking Conference (VNC)*. IEEE. 2014, pp. 61–68.

[Gha+16] M. Gharbieh, H. El-Sawy, A. Bader, and M.-S. Alouini. "Tractable stochastic geometry model for IoT access in LTE networks". In: *Proc. IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2016, pp. 1–7.

[GJ15] A. Gupta and R. K. Jha. "A Survey of 5G Network: Architecture and Emerging Technologies". In: *IEEE Access* 3 (2015), pp. 1206–1232.

[GNT+06] L. Georgiadis, M. J. Neely, L. Tassiulas, et al. "Resource allocation and cross-layer control in wireless networks". In: *Foundations and Trends® in Networking* 1.1 (2006), pp. 1–144.

[GRP16] K. Gatsis, A. Ribeiro, and G. J. Pappas. "Control-aware random access communication". In: *Proc. ACM/IEEE 7th International Conference on Cyber-Physical Systems (ICCPS)*. IEEE. 2016, pp. 1–9.

[Gun+11] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, et al. "Smart grid technologies: communication technologies and standards". In: *IEEE Transactions on Industrial Informatics* 7.4 (2011), pp. 529–539.

[Gun+13] V. C. Gungor, D. Sahin, T. Kocak, S. Ergut, C. Buccella, et al. "A Survey on Smart Grid Potential Applications and Communication Requirements". In: *IEEE Transactions on Industrial Informatics* 9.1 (Feb. 2013), pp. 28–42. ISSN: 1551-3203.

[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.

[GXK16] E. Grigoreva, J. Xu, and W. Kellerer. "M2M Wake-ups over Cellular Networks: Over-the-top SIP". In: *Proc. Workshop on All Things Cellular: Operations, Applications and Challenges*. ACM, 2016, pp. 37–42. ISBN: 978-1-4503-4249-0.

[HH10] K. Hashiura and H. Habuchi. "Performance evaluation of the vehicular ad-hoc network using the modified binary countdown scheme". In: *Proc. Asia-Pacific Conference on Communications (APCC)*. IEEE. 2010, pp. 352–356.

[HHN13] M. Hasan, E. Hossain, and D. Niyato. "Random access for machine-to-machine communication in LTE-advanced networks: issues and approaches". In: *IEEE Communications Magazine* 51.6 (2013), pp. 86–93.

[HIW14] C. Hägerling, C. Ide, and C. Wietfeld. "Coverage and capacity analysis of wireless M2M technologies for smart distribution grid services". In: *Prof. IEEE Int. Conference on Smart Grid Communications (SmartGridComm)*. IEEE. 2014, pp. 368–373.

[HLR11] N. Hu, X.-l. Li, and Q.-n. Ren. "Random access preamble assignment algorithm of TD-LTE". In: *Advances in Computer, Communication, Control and Automation*. Springer, 2011, pp. 701–708.

[HSVDB08] W. P.M. H. Heemels, J. H. Sandee, and P. P. J. Van Den Bosch. "Analysis of event-driven controllers for linear systems". In: *International Journal of Control* 81.4 (2008), pp. 571–590.

[HWT14] Y.-H. Hsu, K. Wang, and Y.-C. Tseng. "Efficient cooperative access class barring with load balancing and traffic adaptive radio resource management for M2M communications over LTE-A". In: *Computer Networks* 73 (2014), pp. 268–281.

[IEE12] IEEE. "IEEE Standard for Local and metropolitan area networks – Part 15.4: Low-Rate Wireless Personal Area Networks (LR-WPANs) Amendment 3: Physical Layer (PHY) Specifications for Low-Data-Rate, Wireless, Smart Metering Utility Networks". In: *IEEE 802.15* (2012).

[Jan+14] H. S. Jang, S. M. Kim, K. S. Ko, J. Cha, and D. K. Sung. "Spatial group based random access for M2M communications". In: *IEEE Communications Letters* 18.6 (2014), pp. 961–964.

[Jia+17] X. Jian, Y. Liu, Y. Wei, X. Zeng, and X. Tan. "Random Access Delay Distribution of Multichannel Slotted ALOHA With Its Applications for Machine Type Communications". In: *IEEE Internet of Things Journal* 4.1 (Feb. 2017), pp. 21–28. ISSN: 2327-4662.

[Jin+17] H. Jin, W. Toor, B. C. Jung, and J. B. Seo. "Recursive Pseudo-Bayesian Access Class Barring for M2M Communications in LTE Systems". In: *IEEE Transactions on Vehicular Technology* 66.9 (Sept. 2017), pp. 8595–8599. ISSN: 0018-9545.

[JOP+01] E. Jones, T. Oliphant, P. Peterson, et al. *SciPy: Open source scientific tools for Python*. http://www.scipy.org/. [Online; accessed 2016-10-12]. 2001–.

[JPS17] H. S. Jang, H.-S. Park, and D. K. Sung. "A non-orthogonal resource allocation scheme in spatial group based random access for cellular M2M communications". In: *IEEE Transactions on Vehicular Technology* 66.5 (2017), pp. 4496–4500.

[KCR10] D. K. Klair, K.-W. Chin, and R. Raad. "A survey and tutorial of RFID anti-collision protocols". In: *IEEE Communications Surveys & Tutorials* 12.3 (2010), pp. 400–421.

[KJS15] T. Kim, H. S. Jang, and D. K. Sung. "An enhanced random access scheme with spatial group based reusable preamble allocation in cellular M2M networks". In: *IEEE Communications Letters* 19.10 (2015), pp. 1714–1717.

[KK13] R. H. Khan and J. Y. Khan. "A comprehensive review of the application characteristics and traffic requirements of a smart grid communications network". In: *Computer Networks* 57.3 (2013), pp. 825 –845. ISSN: 1389-1286.

[KKA13] D. Kim, W. Kim, and S. An. "Adaptive random access preamble split in LTE". In: *Proc. International Conference on Wireless Commun. and Mobile Computing (IWCMC)*. July 2013, pp. 814–819.

[Ko+12] K. S. Ko, M. J. Kim, K. Y. Bae, D. K. Sung, J. H. Kim, 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. ISSN: 1089-7798.

[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.

[Koz69] F. Kozin. "A Survey of Stability of Stochastic Systems". In: *Automatica* 5.1 (Jan. 1969), pp. 95–112.

[KPR14] M. Kuzlu, M. Pipattanasomporn, and S. Rahman. "Communication network requirements for major smart grid applications in HAN, NAN and WAN". In: *Computer Networks* 67 (2014), pp. 74–88.

[KRR16] A. A. Khan, M. H. Rehmani, and M. Reisslein. "Cognitive radio for smart grids: Survey of architectures, spectrum sensing mechanisms, and networking protocols". In: *IEEE Communications Surveys & Tutorials* 18.1 (2016), pp. 860–898.

[KVGAZ16] C. Kalalas, F. Vazquez-Gallego, and J. Alonso-Zarate. "Handling Mission-Critical Communication in Smart Grid Distribution Automation Services through LTE". In: *Proc. IEEE Int. Conf. on Smart Grid Communications*. 2016.

[LAAZ14] 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.

[Lan+13] M. Laner, P. Svoboda, N. Nikaein, and M. Rupp. "Traffic Models for Machine Type Communications". In: *Prof. Int. Symp. on Wireless Communication Systems*. 2013, pp. 1–5.

[Lan+15] M. Laner, N. Nikaein, D. Drajic, P. Svoboda, M. Popovic, et al. "Traffic models for machine-to-machine (M2M) communications: types and applications". In: *Machine-to-machine (M2M) communications*. Ed. by C. Anton-Haro and M. Dohler. Vol. 69. Woodhead Publishing series in electronic and optical materials. Amsterdam: Elsevier/Woodhead Publ, 2015.

[LBT01] J.-Y. Le Boudec and P. Thiran. *Network calculus: a theory of deterministic queuing systems for the Internet*. Vol. 2050. Springer Science & Business Media, 2001.

[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.

[LCW16] G. Y. Lin, S. R. Chang, and H. Y. Wei. "Estimation and Adaptation for Bursty LTE Random Access". In: *IEEE Transactions on Vehicular Technology* 65.4 (Apr. 2016), pp. 2560–2577. ISSN: 0018-9545.

[Lee+12] K.-D. Lee, M. Reisslein, K. Ryu, and S. Kim. "Handling randomness of multi-class random access loads in LTE-Advanced network supporting small data applications". In: *Proc. 2012 IEEE Globecom Workshops*. 2012, pp. 436–440.

[Li+16] B. Li, Y. Ma, T. Westenbroek, C. Wu, H. Gonzalez, et al. "Wireless Routing and Control: a Cyber-Physical Case Study". In: *Proc. ACM/IEEE Int. Conference on Cyber-Physical Systems (ICCPS)*. IEEE. 2016, pp. 1–10.

[Lie+12] S.-Y. Lien, T.-H. Liau, C.-Y. Kao, and K.-C. Chen. "Cooperative Access Class Barring for Machine-to-Machine Communications". In: *IEEE Trans. Wireless Commun.* 11.1 (Jan. 2012), pp. 27–32. ISSN: 1536-1276.

[Lin+14] T.-M. Lin, C.-H. Lee, J.-P. Cheng, and W.-T. Chen. "PRADA: prioritized random access with dynamic access barring for MTC in 3GPP LTE-A networks". In: *IEEE Transactions on Vehicular Technology* 63.5 (2014), pp. 2467–2472.

[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.

[LKY11] K.-D. Lee, S. Kim, and B. Yi. "Throughput comparison of random access methods for M2M service over LTE networks". In: *Proc. IEEE GLOBECOM Workshops (GC Wkshps)*. Dec. 2011, pp. 373–377.

[LL10] J. Lunze and D. Lehmann. "A state-feedback approach to event-based control". In: *Automatica* 46.1 (2010), pp. 211–215. ISSN: 0005-1098.

[LM+17] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner. "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure and the Access Class Barring Scheme". In: *IEEE Transactions on Wireless Communications* 16.12 (Dec. 2017), pp. 7785–7799. ISSN: 1536-1276.

[Lo+11] A. Lo, Y. W. Law, M. Jacobsson, and M. Kucharzak. "Enhanced LTE-advanced random-access mechanism for massive machine-to-machine communications". In: *Proc. 27th WWRF Meeting*. 2011, pp. 1–5.

[Lora] *LoRa Alliance*. Accessed: 2018-06-16. URL: https://lora-alliance.org/.

[Lorb] *The Things Network: LoRaWAN distance record*. Accessed: 2018-08-04. URL: https://www.thethingsnetwork.org/article/ground-breaking-world-record-lorawan-packet-received-at-702-km-436-miles-distance.

[LSS06] X. Lin, N. B. Shroff, and R. Srikant. "A tutorial on cross-layer optimization in wireless networks". In: *IEEE Journal on Selected areas in Communications* 24.8 (2006), pp. 1452–1463.

[Mad+15] G. C. Madueno, N. K. Pratas, Č. Stefanović, and P. Popovski. "Massive M2M access with reliability guarantees in LTE systems". In: *Proc. IEEE ICC*. 2015, pp. 2997–3002.

[Mad+16] G. C. Madueño, J. J. Nielsen, D. M. Kim, N. K. Pratas, Č. Stefanović, 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 (2016), pp. 675–688.

[Mag+18] D. Magrin, C. Pielli, Č. Stefanović, and M. Zorzi. "Enabling LTE RACH Collision Multiplicity Detection via Machine Learning". In: *CoRR* pp, abs/1805.11482 (2018). arXiv: 1805.11482.

[Mam+14] M. Mamduhi, D. Tolic, A. Molin, and S. Hirche. "Event-triggered Scheduling for Stochastic Multi-loop Networked Control Systems with Packet Dropouts". In: *Proc. IEEE Conference on Decision and Control*. 2014, pp. 2776–2782.

[MDH15] M. H. Mamduhi, F. Deroo, and S. Hirche. "Event-based data scheduling for a class of interconnected networked control systems". In: *Proc. IEEE Conference on Decision and Control*. Dec. 2015, pp. 4183–4189.

[Meh+15] Y. Mehmood, S. Nawaz Khan Marwat, C. Görg, Y. Zaki, and A. Timm-Giel. "Evaluation of M2M Data Traffic Aggregation in LTE-A Uplink". In: *Proc. ITG/VDE Mobile Communication Conference* (2015), pp. 24–29.

[MG16] F. Morvari and A. Ghasemi. "Two-Stage Resource Allocation for Random Access M2M Communications in LTE Network". In: *IEEE Commun. Letters* 20.5 (2016), pp. 982–985.

[MH13] A. Molin and S. Hirche. "On the Optimality of Certainty Equivalence for Event-Triggered Control Systems". In: *IEEE Transactions on Automatic Control* 58.2 (2013), pp. 470 –474.

[MH14a] A. Molin and S. Hirche. "A bi-level approach for the design of event-triggered control systems over a shared network". In: *Discrete Event Dynamic Systems* 24.2 (2014), pp. 153–171.

[MH14b] A. Molin and S. Hirche. "Price-based Adaptive Scheduling in Multi-Loop Control Systems with Resource Constraints". In: *IEEE Transactions on Automatic Control* (2014), pp. 3282 –3295.

[Mie08] K. Miettinen. "Introduction to multiobjective optimization: Noninteractive approaches". In: *Multiobjective optimization*. Springer, 2008, pp. 1–26.

[Mik79] V. Mikhailov. "Methods of random multiple access". In: *Candidate Eng. Thesis, Moscow Institute of Physics and Technology* (1979).

[MKH16] M. H. Mamduhi, M. Kneissl, and S. Hirche. "Decentralized event-triggered medium access control for networked control systems". In: *Proc. IEEE Conference on Decision and Control (CDC)*. 2016, pp. 513–519.

[MM16] J. Mišić and V. B. Mišić. "To Shout or Not to Shout: Performance of Power Ramping during Random Access in LTE/LTE-A". In: *Proc. IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2016, pp. 1–6.

[MM18] J. Mišić and V. B. Mišić. "Efficiency of power ramping during random access in LTE". In: *IEEE Transactions on Vehicular Technology* 67.2 (2018), pp. 1698–1712.

[MMA17] J. Mišić, V. B. Mišić, and M. Z. Ali. "Explicit Power Ramping during Random Access in LTE/LTE-A". In: *Proc. IEEE Wireless Communications and Networking Conference (WCNC)*. 2017, pp. 1–6. DOI: 10.1109/WCNC.2017.7925667.

[MMH14] M. Mamduhi, A. Molin, and S. Hirche. "Event-based Scheduling of Multi-loop Stochastic Systems over Shared Communication Channels". In: *Proc. Int. Symposium on Mathematical Theory of Networks and Systems (MTNS)*. 2014, pp. 266–273.

[MPH16] K. Mikhaylov, J. Petaejaejaervi, and T. Haenninen. “Analysis of capacity and scalability of the LoRa low power wide area network technology”. In: *Proc. European Wireless Conference*. VDE. 2016, pp. 1–6.

[MSP14] G. Madueno, S. Stefanović, and P. Popovski. “Efficient LTE access with collision resolution for massive M2M communications”. In: *Proc. IEEE GLOBECOM Workshops*. Dec. 2014, pp. 1433–1438.

[Muk+16] A. Mukherjee, J. F. Cheng, S. Falahati, H. Koorapaty, D. H. Kang, et al. “Licensed-Assisted Access LTE: coexistence with IEEE 802.11 and the evolution toward 5G”. In: *IEEE Communications Magazine* 54.6 (June 2016), pp. 50–57. ISSN: 0163-6804.

[Mul17] MulteFire. *MulteFire Release 1.0 Technical Paper: A New Way to Wireless*. Tech. rep. 2017.

[NT04] D. Nesic and A. Teel. “Input-output stability properties of networked control systems”. In: *IEEE Transactions on Automatic Control* 49.10 (2004), pp. 1650–1667.

[NWK14] D. Niyato, P. Wang, and D. I. Kim. “Performance Modeling and Analysis of Heterogeneous Machine Type Communications”. In: *IEEE Transactions on Wireless Communications* 13.5 (May 2014), pp. 2836–2849. ISSN: 1536-1276.

[Oss+14] A. Osseiran, F. Boccardi, V. Braun, K. Kusume, P. Marsch, et al. “Scenarios for 5G mobile and wireless communications: the vision of the METIS project”. In: *IEEE Communications Magazine* 52.5 (2014), pp. 26–35.

[Par+17] P. Park, S. C. Ergen, C. Fischione, C. Lu, and K. H. Johansson. “Wireless Network Design for Control Systems: A Survey”. In: *IEEE Communications Surveys & Tutorials* (2017).

[PC15] F. Poloczek and F. Ciucu. “Service-martingales: Theory and applications to the delay analysis of random access protocols”. In: *Proc. IEEE International Conference on Computer Communications (INFOCOM)*. IEEE. 2015, pp. 945–953.

[Ped+16] K. I. Pedersen, G. Berardinelli, F. Frederiksen, P. Mogensen, and A. Szufarska. “A flexible 5G frame structure design for frequency-division duplex cases”. In: *IEEE Communications Magazine* 54.3 (2016), pp. 53–59.

[PF95] V. Paxson and S. Floyd. “Wide area traffic: the failure of Poisson modeling”. In: *IEEE/ACM Transactions on Networking* 3.3 (1995), pp. 226–244. ISSN: 1063-6692.

[PFP04] P. Popovski, F. H. Fitzek, and R. Prasad. “Batch conflict resolution algorithm with progressively accurate multiplicity estimation”. In: *Proc. Workshop on Foundations of Mobile Computing*. ACM. 2004, pp. 31–40.

[PL16] J. Park and Y. Lim. “Adaptive Access Class Barring Method for Machine Generated Communications”. In: *Mobile Information Systems* (2016), pp. 1–6.

[Pop+17] P. Popovski, J. J. Nielsen, C. Stefanovic, E. de Carvalho, E. G. Ström, et al. "Wireless Access for Ultra-Reliable Low-Latency Communication (URLLC): Principles and Building Blocks". In: *Computer Research Repository (CoRR* abs/1708.07862 (2017). arXiv: 1708.07862.

[Pop14] P. Popovski. "Ultra-reliable communication in 5G wireless systems". In: *Proc. Int. Conference on 5G for Ubiquitous Connectivity (5GU)*. IEEE. 2014, pp. 146–151.

[Pra+12] N. K. Pratas, H. Thomsen, Č. Stefanović, and P. Popovski. "Code-expanded random access for machine-type communications". In: *Proc. IEEE Global Communications Conference (GLOBECOM) Workshops*. 2012, pp. 1681–1686.

[Pra+16] N. K. Pratas, C. Stefanovic, G. C. Madueño, and P. Popovski. "Random access for machine-type communication based on Bloom filtering". In: *Proc. IEEE Global Communications Conference (GLOBECOM)*. IEEE. 2016, pp. 1–7.

[RALRCP09] M. E. Rivero-Angeles, D. Lara-Rodríguez, and F. A. Cruz-Pérez. "Differentiated backoff strategies for prioritized random access delay in multiservice cellular networks". In: *IEEE Trans. Vehicular Techn.* 58.1 (2009), pp. 381–397.

[Riv87] R Rivest. "Network control by Bayesian broadcast". In: *IEEE Transactions on Information Theory* 33.3 (1987), pp. 323–328.

[RMB09] M. Rabi, G. Moustakides, and J. Baras. "Adaptive sampling for linear state estimation". In: *SIAM journal on control and optimization* (2009).

[RS90] R. Rom and M. Sidi. *Multiple Access Protocols: Performance and Analysis*. New York, NY, USA: Springer-Verlag New York, Inc., 1990. ISBN: 0-387-97253-6.

[RSJ12] C. Ramesh, H. Sandberg, and K. Johansson. "Stability analysis of multiple state-based schedulers with CSMA". In: *Decision and Control (CDC), 2012 IEEE 51st Annual Conference on*. 2012, pp. 7205–7211.

[Sai+13] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, et al. "Non-orthogonal multiple access (NOMA) for cellular future radio access". In: *Proc. IEEE Vehicular Technology Conference (VTC Spring)*. IEEE. 2013, pp. 1–5.

[SH00] J. H. Sarker and S. J. Halme. "An optimum retransmission cut-off scheme for slotted ALOHA". In: *Wireless Personal Commun.* 13.1-2 (2000), pp. 185–202.

[Sha+12] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, and J. Wang. "A first look at cellular machine-to-machine traffic: large scale measurement and characterization". In: *ACM SIGMETRICS Performance Evaluation Review* 40.1 (2012), pp. 65–76.

[Sha+15] H. Shariatmadari, P. Osti, S. Iraji, and R. Jäntti. "Data Aggregation in Capillary Networks for Machine-to-Machine Communications". In: *Proc. IEEE Personal, Indoor and Mobile Radio Communications (PIMRC) - Workshop on M2M Communications: Challenges, Solutions and Applications* (2015), pp. 1100–1105.

[Shi+06] D.-H. Shih, P.-L. Sun, D. C. Yen, and S.-M. Huang. "Taxonomy and survey of RFID anti-collision protocols". In: *Computer communications* 29.11 (2006), pp. 2150–2166.

[Sig] *SigFox*. Accessed: 2018-06-16. URL: www.sigfox.com.

[Sin+04] B. Sinopoli, L. Schenato, M. Franceschetti, K. Poolla, M. I. Jordan, et al. "Kalman filtering with intermittent observations". In: *IEEE Transactions on Automatic Control* 49.9 (2004), pp. 1453–1464.

[SL11] J.-B. Seo and V. C. Leung. "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems". In: *IEEE Trans. Vehicular Techn.* 60.8 (2011), pp. 3975–3989.

[SLP17] C. Stefanović, F. Lazaro, and P. Popovski. "Frameless ALOHA with Reliability-Latency Guarantees". In: *Proc. IEEE Global Communications Conference (GLOBECOM)*. Dec. 2017, pp. 1–6.

[Son+17] L. Song, W. Zhou, Y. Hou, and M. Gao. "Load-aware ACB Scheme for M2M Traffic in LTE-A Networks". In: *Advances on Broad-Band Wireless Computing, Communication and Applications: Proceedings of the 11th International Conference On Broad-Band Wireless Computing, Communication and Applications (BWCCA–2016) November 5–7, 2016, Korea*. Ed. by L. Barolli, F. Xhafa, and K. Yim. Springer International Publishing, 2017, pp. 69–80. ISBN: 978-3-319-49106-6.

[SRCN11] S. Sen, R. Roy Choudhury, and S. Nelakuditi. "No time to countdown: Migrating backoff to the frequency domain". In: *Proc. ACM Int. Conference on Mobile Computing and Networking (MobiCom)*. ACM. 2011, pp. 241–252.

[Tab07] P. Tabuada. "Event-Triggered Real-Time Scheduling of Stabilizing Control Tasks". In: *IEEE Transactions on Automatic Control* 52.9 (2007), pp. 1680 –1685.

[Tan02] A. Tanenbaum. *Computer Networks*. 4th. Prentice Hall Professional Technical Reference, 2002. ISBN: 0130661023.

[TF13] D. Tolić and R. Fierro. "Decentralized Output Synchronization of Heterogeneous Linear Systems with Fixed and Switching Topology via Self-Triggered Communication". In: *American Control Conference*. 2013, pp. 4655–4660.

[TM78] B. S. Tsybakov and V. A. Mikhailov. "Free synchronous packet access in a broadcast channel with feedback". In: *Problemy Peredachi Informatsii* 14.4 (1978), pp. 32–59.

[TMF17] H. Thomsen, C. N. Manchon, and B. H. Fleury. "A traffic model for machine-type communications using spatial point processes". In: *Proc. IEEE Int. Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*. 2017, pp. 1–6.

[TN08] M. Tabbara and D. Nesic. "Input–output stability of networked control systems with stochastic protocols and channels". In: *IEEE Transactions on Automatic Control* 53.5 (2008), pp. 1160–1175.

[Tob82] F. A. Tobagi. "Distributions of packet delay and interdeparture time in slotted ALOHA and carrier sense multiple access". In: *Journal of the ACM (JACM)* 29.4 (1982), pp. 907–927.

[TPM16] D. Tsolkas, N. Passas, and L. Merakos. "Device discovery in LTE networks: A radio access perspective". In: *Computer Networks* 106 (2016), pp. 245–259.

[TST12] S.-Y. Tsai, S.-I. Sou, and M.-H. Tsai. "Effect of Data Aggregation in M2M Networks". In: *Proc. Int. Symposium on Wireless Personal Multimedia Communications (WPMC)* (2012), pp. 95–99. ISSN: 13476890.

[Tya+15] R. Tyagi, F. Aurzada, K.-D. Lee, S. Kim, and M. Reisslein. "Impact of Retransmission Limit on Preamble Contention in LTE-Advanced Network". In: *IEEE Systems Journal* 9.3 (Sept. 2015), pp. 752–765. ISSN: 1932-8184.

[Tya+17] R. Tyagi, F. Aurzada, K.-D. Lee, and M. Reisslein. "Connection Establishment in LTE-A Networks: Justification of Poisson Process Modeling". In: *IEEE Systems Journal, in print* PP.99 (2017), pp. 1–12. ISSN: 1932-8184.

[Var+01] A. Varga et al. "The OMNeT++ discrete event simulation system". In: *Proc. European Simulation Multiconference*. Vol. 9. sn. 2001, p. 65.

[Wan+13] S. H. Wang, H. J. Su, H. Y. Hsieh, S. P. Yeh, and M. Ho. "Random access design for clustered wireless machine to machine networks". In: *Proc. IEEE Int. Black Sea Conference on Communications and Networking (BlackSeaCom)* (2013), pp. 107–111.

[Wat+01] R. Wattenhofer, L. Li, P. Bahl, and Y.-M. Wang. "Distributed topology control for power efficient operation in multihop wireless ad hoc networks". In: *Proc. IEEE International Conference on Computer Communications (INFOCOM)*. Vol. 3. IEEE. 2001, pp. 1388–1397.

[WBC15] 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.

[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 (2015), pp. 33–46.

[Wu+11] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson. “M2M: From mobile to embedded Internet”. In: *IEEE Communications Magazine* 49.4 (Apr. 2011), pp. 36–43. ISSN: 0163-6804.

[WY01] G. C. Walsh and H. Ye. “Scheduling of networked control systems”. In: *Control Systems, IEEE* 21.1 (2001), pp. 57–65.

[WYB02] G. C. Walsh, H. Ye, and L. G. Bushnell. “Stability analysis of networked control systems”. In: *IEEE Transactions on Control Systems Technology* 10.3 (2002), pp. 438–446.

[XHL14] L. D. Xu, W. He, and S. Li. “Internet of Things in Industries: A Survey”. In: *IEEE Transactions on Industrial Informatics* 10.4 (Nov. 2014), pp. 2233–2243. ISSN: 1551-3203.

[Xia05] Y. Xiao. “Performance analysis of priority schemes for IEEE 802.11 and IEEE 802.11e wireless LANs”. In: *IEEE Trans. Wireless Commun.* 4.4 (2005), pp. 1506–1515.

[YFE12] X. Yang, A. Fapojuwo, and E. Egbogah. “Performance Analysis and Parameter Optimization of Random Access Backoff Algorithm in LTE”. In: *Proc. Vehicular Techn. Conf. (VTC Fall)*. Sept. 2012, pp. 1–5.

[YG07] Y. Yu and G. B. Giannakis. “High-throughput random access using successive interference cancellation in a tree algorithm”. In: *IEEE Transactions on Information Theory* 53.12 (2007), pp. 4628–4639.

[YHH11] O. N. Yilmaz, J. Hämäläinen, and S. Hämäläinen. “Self-optimization of random access channel in 3rd Generation Partnership Project Long Term Evolution”. In: *Wireless Communications and Mobile Computing* 11.12 (2011), pp. 1507–1517.

[Yun12] J.-H. Yun. “Cross-layer analysis of the random access mechanism in Universal Terrestrial Radio Access”. In: *Computer Networks* 56.1 (2012), pp. 315–328.

[YY03] Y. Yang and T.-S. P. Yum. “Delay distributions of slotted ALOHA and CSMA”. In: *IEEE Transactions on Communications* 51.11 (2003), pp. 1846–1857.

[YYH03] T. You, C.-H. Yeh, and H. Hassanein. “A new class of collision prevention MAC protocols for wireless ad hoc networks”. In: *Proc. IEEE International Conference on Communications*. Vol. 2. IEEE. 2003, pp. 1135–1140.

[Zan12] A. Zanella. “Estimating collision set size in framed slotted aloha wireless networks and RFID systems”. In: *IEEE Communications Letters* 16.3 (2012), pp. 300–303.

[ZGA16] N. Zangar, S. Gharbi, and M. Abdennebi. "Service differentiation strategy based on MACB factor for M2M Communications in LTE-A Networks". In: *Proc. IEEE Consumer Commun. & Netw. Conf. (CCNC)*. 2016, pp. 693–698.

[Zha+12] Y. Zhang, R. Yu, M. Nekovee, Y. Liu, S. Xie, et al. "Cognitive machine-to-machine communications: Visions and potentials for the smart grid". In: *IEEE Network* 26.3 (2012), pp. 6–13.

[ZZF14] X. Zhao, J. Zhai, and G. Fang. "An Access Priority Level Based Random Access Scheme for QoS Guarantee in TD-LTE-A Systems". In: *Proc. IEEE Vehicular Techn. Conf. (VTC2014-Fall)*. 2014, pp. 1–5.

[3GP12] 3GPP. *Technical Report 36.822; LTE Radio Access Network (RAN) enhancements for diverse data applications*. Tech. rep. 3rd Generation Partnership Project, Sept. 2012.

[3GP15a] 3GPP. *Technical Specification 23.203; Policy and charging control architecture*. Tech. rep. 3rd Generation Partnership Project, June 2015.

[3GP15b] 3GPP. *Technical Specification 36.211; Evolved Terrestrial Radio Access (E-UTRA); Physical channels and modulation (Release 12)*. Tech. rep. 3rd Generation Partnership Project, Aug. 2015.

[3GP16] 3GPP. *Technical Specification 36.331; Evolved Terrestrial Radio Access (E-UTRA); Radio Resource Control (RRC); Protocol specification*. Tech. rep. 3rd Generation Partnership Project, June 2016.

[3GP18a] 3GPP. *Technical Report 38.913; Technical Specification Group Radio Access Network; Study on Scenarios and Requirements for Next Generation Access Technologies; (Release 15)*. Tech. rep. 3rd Generation Partnership Project, June 2018.

[3GP18b] 3GPP. *Technical Specification 36.321; Evolved Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification; Version 15.1.0*. Tech. rep. 3rd Generation Partnership Project, Apr. 2018.

[Hog+18] A. Hoglund, D. P. Van, T. Tirronen, O. Liberg, Y. Sui, et al. "3GPP Release 15 Early Data Transmission". In: *IEEE Communications Standards Magazine* 2.2 (June 2018), pp. 90–96. DOI: 10.1109/MCOMSTD.2018.1800002.

[ICT15] ICT-317669 METIS. *Deliverable 6.6 Version 1 "Final report on the METIS 5G system concept and technology roadmap"*. Tech. rep. 2015.

[Nok15] Nokia Solutions and Networks. *LTE-M - Optimizing LTE for the Internet of Things, Whitepaper*. Tech. rep. 2015.

[Roh16] Rohde und Schwarz. *Narrowband Internet of Things, Whitepaper*. Tech. rep. 2016.

List of Figures

List of Tables

Acronyms

3GPP Third Generation Partnership Project

ACB Access Class Barring

AMF Access and Mobility Management Function

BCCR Binary Countdown Contention Resolution

gNB Next Generation Node B

CAN Controller Area Network

CRS Contention Resolution Slot

CSMA Carrier Sense Multipe Access

CSMA/CA Carrier Sense Multipe Access with Collision Avoidance

DBCA Dynamic Binary Countdown - Access barring

DCF Distributed Coordination Function

EAB Extended Access Class Barring

ETSI European Telecommunications Standards Institute

FDMA Frequency Division Multiple Access

H2H Human-to-Human

HVAC Heating Ventilation and Air Conditioning

LATMAPA Load-Adaptive Throughput MAximizing Preamble Allocation

LBT Listen Before Talk

LoRa Long Range

LoRaWAN Long Range Wide-Area Network

LPWAN Low-Power Wide-Area Network

LSP Lyapunov Stability in Probability

LTI Linear Time-Invariant

M2M Machine-to-Machine

MAC Medium Access Control

MGF Moment Generating Function

mM2M massive Machine-to-Machine

MME Mobility Management Entity

MTC Machine Type Communications

NB-IoT Narrowband Internet of Things

NCS Networked Control System

NOMA Non-Orthogonal Multiple Access

OFDM Orthogonal Frequency Division Multiplexing

OFDMA Orthogonal Frequency Division Multiple Access

PRACH Physical Random Access CHannel

PUSCH Physical Uplink Shared CHannel

QCI Quality of Service Class Indicator

QoS Quality of Service

RACH Random Access CHannel

RAN Radio Access Network

RAO Random Access Opportunity

RAP Random Access Procedure

RB Resource Block

RFID Radio Frequency IDendintification

RRC Radio Resource Control

SIC Successive Interference Cancellation

SINR Signal to Interference to Noise Ratio

SNR Signal to Noise Ratio

TA Timing Advance

TDMA Time Division Multiple Access

TRA Tree Resolution Algorithms

TSCH Time Slotted Channel Hopping

UE User Equipment

uM2M ultra reliable Machine-to-Machine

URLLC Ultra-Reliable Low Latency Communication