Università degli Studi di Torino
Department of Computer Science
Doctoral School in Sciences and Innovative Technologies

RESEARCH DOCTORATE IN COMPUTER SCIENCE
XXX Cycle

Performance evaluation of massive communications in future wireless access networks

Doctoral Dissertation of:
Paolo Castagno

Advisor:
Prof. Matteo Sereno

Supervisor of the Doctoral Program:
Prof. Marco Grangetto

Academic Year: **2017/2018**

Scientific Disciplinary Sector: **INF/01**

Acknowledgments

I would like to thank all the people I met during this three years of PhD, who made me grow up as a person and a researcher.

First of all, I would like to express my gratitude to three people who make the difference for my academic career. Thanks to professor Matteo Sereno, for having taught me how to do research. I very thank him for our long discussions, for his patience, his availability and for our coffee breaks of "doing research". Thanks to professor Vincenzo Mancuso, who gave me the opportunity to learn from its versatile and countless talents. Thanks to professor Marco Ajmone Marsan, his extraordinary abilities and foresight will always be the model I aim for.

Thanks to all the people of the (not anymore) "l'Acquario" and all the people from ImDEA Networks. Our coffee breaks have been of fundamental importance for my research activity. Thanks to all my friends for all the time spent together.

Special thanks to my family, for their unconditioned support and love.

Finally, thanks to Enrica. There are so many things to be grateful for, but the only that matters is that you are every day at my side.

Contents

List of Figures

List of Tables

List of acronyms

eNB	Evolved node B
1G	First Cellular Generation
2G	Second Cellular Generation
3G	Third Cellular Generation
3GPP	3rd Generation Partnership Project
4G	Fourth Cellular Generation
5G	Fifth generation of cellular system
5GPPP	5th generation Public and Private Partnership
ACB	Access Class Barring
AS	Access Stratum
BS	Base Station
C-RNTI	Cell Radio Network Temporal Identifier
CN	Core Network
D2D	Device to Device
DRX	Discontinuous Reception
E-UTRAN	Evolved Universal Terrestrial Radio Access Network
E2E	End to End
EAB	Extended Access Barring
EPC	Evolved Packet Core
EPS	Evolved Packet System
RRC	Radio Resource Control
FoF	Factory of the Future
GSM	Global System for Mobile communications
HetNet	Hetereogeneous Networks
HSS	Home Subscriber Server
HTC	Human Type communications

ICT	Information and Communications Technology
IoT	Internet of Things
IP	Internet Protocol
KPI	Key Parameter Indicator
LTE	Long-Term evolution
LTE-A	Long Term Evolution Advanced
M&E	Media and Enterteinmnet
M2M	Machine to Machine
MAC	Medium Access Control
MME	Mobility Management Entity
MTC	Machine Type Communications
MTD	Machine Type Device
MVA	Mean Value Analysis
NAS	Non-Access Stratum
NB-IoT	Narrow-Band IoT
NFV	Network Function Virtualisation
OFDMA	Orthogonal Frequency Division Multiple Access
P-GW	Packet Gateway
PCEF	Policy Control Enforcement Function
PCRF	Policy Control Rules Function
PDCCH	Physical Downlinl Control Channel
PDCP	Packet Dara Convergence Protocol
PDN	Packet Data Network
PS	Preamble Signature
QAM	Quadrature Amplitude Modulation
QoE	Quality of Experience
QoS	Quality of Service
RA	Random Access
RACH	RandomAccess Channel
RAO	Random Access Opportunity
RAR	Random Access Response
RB	Resource Block
RLC	Radio Link Control
S-GW	Serving Gateway
S1-MME	S1 control-plane
S1-U	S1 user-plane
SAE	System Architecture Evolution
SDN	Software Defined Networks
SF	Smart Factory
SNR	Signal to Noise ratio
TC-RNTI	Temporal C-RNTI
UE	User Equipment
ULSCH	Upload Shared Channel

Abstract

This thesis contributes to the state of the art providing, both for human-type communications (HTC) and machine type communications (MTC), detailed models of network access in 4G and future 5G networks. The focus, in the case of HTC, is on types of scenarios where a huge crowd gathers in a relatively small place, such as a stadium during a music concert. In such case, on one side the metric of interest is the quality perceived by the users while on the other is the number of users to which the connectivity is guaranteed. Therefore, the presented models allow to analytically identify networks operational regions relating the number of devices and the QoS they perceive. In the case of MTC, given the specific requirements of autonomous communications, guaranteeing strict latency constraints and high reliability is of primary importance. Hence, this thesis describes a stochastic model of the behavior of autonomous devices in a Smart Factory scenario. The model allows to evaluate operational conditions and to derive the distribution of latencies experienced by network access requests.

In 4G networks, and in the next generation 5G networks, access to resources is the result of a two-step operation: in the first place mobile devices notify to the network the attempt to join the network, through the RA procedure, and only in case of networks acknowledgement they can proceed and negotiate a data channel. In the latter step, devices might be refused to access resources essentially because of protocols constraints. Proven that it is possible to model the two steps separately, in this thesis it's derived a novel formulation of the second one, based on a network of queue under product form solution, that overcomes limitations of state of the art approaches, such as the well known Erlang-B formula.

Mobile network issues, such as overloading and real-time constraint, might also be studied from a different point o view. In recent years, coalesce efforts of a crowd towards a common objective has been of great appeal. That is, crowd-sourcing attracted interest in many different fields; from fundraising campaigns to crowdsourcing Internet marketplace. Borrowing the idea of cooperation, mobile devices within the same cell can group up into coalitions (a subset of devices close to each other) and communicate with the network by means of a single coalition component. That is, Device to Device (D2D) communication enables direct communications among components of a coalitions and the device depicted to communicate with the network will act as a data gatherer and relay node. Eventually it is also shown how such mechanism is beneficial to mobile networks, both in case of HTC and MTC.

Introduction

The promise of next generation mobile network is a unprecedented flexibility. Indeed, the 5G cellular system is designed to offer application specific requirements, such as high reliability and ultra-low latency for factory automation. KPIs define the target performance the fifth Generation (5G) systems will have to meet in each specific application domain, or vertical industry. Overall the the Key performance Indexes (KPI) are shared among verticals, and most of the application specific scenarios require *i*) ultra-low latency ($\leq 5ms$), *ii*) support to high terminal density ($\geq 1M terminals/km^2$) and *iii*) high data volume per geographical area (up to $0.75 Tb/s$ in a stadium) [2].

To cope with such extremely demanding requirements is challenging, and meeting KPIs will only be possible approaching the challenge from multiple points, gathering improvements from several innovative solutions–from network slicing to drone-mounted cells. However, most of these approaches lay on the same Radio Access Network (RAN) , which it is expected to remain untouched in the 5G.

In a cellular system, the RAN is the component that implements radio interfaces and offers the suitable protocols to manage and use it. Given its relevance, RAN has been widely studied and many improvements have been proposed. Nevertheless, there are only a few inclusive studies investigating RAN and its components performance, such as [1]. Therefore, given a metric –e.g. Quality of Service (QoS) or access time– it is not easy to understand how a specific RAN component affects it. Moreover, without knowing how components relates each other it is impossible to evaluate how the system will benefit of a component's improvement.

This thesis approaches the gap in the RAN performance evaluation in massive communication scenarios. It provides a comprehensive analysis of the access procedure in current and future cellular systems. In particular, with reference to the connection establishment, both HTC and MTC are studied, keeping in mind the specific requirements of each scenario. Furthermore, we also investigate an approach to cope with access and capacity issues in cellular network, and proposes a possible cooperative implementation—i.e., user aided. Eventually, it is proposed an innovative model for the data access phase which fills the dearth of a performance model encompassing the RAN's peculiarities, overcoming limitations of traditional models.

The work presented in this doctoral thesis is divided as follow, and each chapter overviews the reference state of the art:

Chapter 1 provides an brief description of the mobile network development and its future directions. Moreover, it provides an overview of fourth Generation (4G)—and most likely 5G—architecture and protocols.

Chapter 2 discusses cellular network performance in case of high user density and human generated traffic. Such overcrowded situation it is foreseen to become the normal use case scenario in next generation networks according to both the 5GPPP KPIs and the ever growing trends of connected mobile devices [3]. The analysis conduced here investigates how Access Class Barring and Random Access procedure impact on the typical network usage pattern in HTC. From this analysis it is possible to infer useful insight for network designing and management. Specifically, the analytical model locates the operational points marking significant changes in network performance. Such insights are useful in the designing process to carefully tune the specific cell to a target population and QoS requirements. Moreover, this results will play a fundamental role in case of proactive network management, allowing a timely reconfigurations in a SDN. In addition to that, the analysis is extended to a cooperative scenario in which UEs can coalesce their requests and jointly access network resources, assuming the existence of a Device to Device (D2D) [4] mechanism to enable UEs direct communications. A part of the material of this chapter is reported in: *Castagno, Paolo, Vincenzo Mancuso, Matteo Sereno, and Marco Ajmone Marsan. "Why your smartphone doesn't work in very crowded environments." In A World of Wireless, Mobile and Multimedia Networks (WoWMoM), pp. 1-9. IEEE, 2017.*

Chapter 3 studies the use of D2D communications to enhance cellular network capacity, involving UEs in the delivery of a multimedia content. Specifically, the focus of this chapter relies on persuading UEs to employ their own resources to help out the network. Moreover, it proposes a *bargaining-like* mechanism to set up a D2D overlay network which allows service providers to get an optimal load redistribution—between cellular and D2D connections—resulting in economical savings. A part of the material of this chapter is presented: *Castagno Paolo, Rossano Gaeta, Marco Grangetto, and Matteo Sereno. "Device-to-device content distribution in cellular networks: a user-centric collaborative strategy."*

In Global Communications Conference (GLOBECOM), 2015 IEEE, pp. 1-6. IEEE, 2015.

Chapter 4 analyses MTC in a SF scenario. Indeed, it is an extremely challenging scenario and, as pointed out in [5], it is a valid choice to separately verify both the impact of high device density and critical real-time traffic. In this chapter, an additional use case is proposed: monitoring and actuating critical machineries in a assembly line generate time-critical traffic from a high number of co-located devices, arising a massive and time critical scenario. The model presented in this chapter, encompasses a detailed description of the access phase (ACB, RA) in cellular networks and of the data connection phase, keeping in consideration the limitations and the delay at each step. Given the scarcity of works encompassing all the phases required to access network resources, this model is a particularly interesting tool to boost the investigation of latencies and bottleneck preventing KPIs fulfilment. The material of this chapter is reported in:*Castagno, Paolo, Vincenzo Mancuso, Matteo Sereno, and Marco Ajmone Marsan. "A Simple Model of MTC in Smart Factories." In IEEE Conference on Computer Communications (INFOCOM), IEEE, 2018.*

Chapter 5 contrasts a novel formulation of the data access phase with traditional formulae. Indeed, models employed to represent network services are mostly the same since the definition of the Erlang-B formula, and are not capable of capturing many key features of cellular networks. Specifically, RRC allows UEs to perform short periods of inactivity between subsequent transmissions in which network resources are kept reserved, generating a flow of returning network requests. Such returning clients have a non-negligible impact on system performances, but are widely overseen. The formulation presented in this chapter offers a generic model easily integrable in other performance studies of different system components. Indeed, the solution can be computed efficiently and, more important, it does not depend on system time distribution. The material of this chapter is reported in: *Castagno, Paolo, Vincenzo Mancuso, Matteo Sereno, and Marco Ajmone Marsan. "Closed Form Expressions for the Performance Metrics of Data Services in Cellular Networks." In IEEE Conference on Computer Communications (INFOCOM), IEEE, 2018.*

Chapter 6 provides some insights towards future investigations and concludes the thesis.

Chapter 1

Mobile networks

Cellular networks became a pervasive and ubiquitous part of everyone daily life. Nowadays, it is almost impossible to imagine spending a whole day, or even a few hours, without chatting or checking the news from our smartphones. Mobile communications are not only changing human interactions, but are also becoming a key driver in industrial production processes.

From its early steps in the 1980s, cellular networks have greatly improved. The analog first generation (1G) cellular system supported only voice services. With the introduction of digital transmissions in the second generation of cellular systems, such as GSM (Global System for Mobile communications[1]), came an opportunity to increase network capacity and to give a more consistent quality of service [6].

The introduction of digital transmissions was a break through innovation in mobile systems, and it still is a fundamental pillar of today cellular networks as it will be for the forthcoming and future ones. It was a turning point in cellular network development in the sense that in 2G networks primary data services were introduced, such as text messages and circuit-switched data services. Nevertheless, at this stage 2G networks were providing a modest peak rate of 9.6 Kbps and only with the advent of 3G, the third generation of mobile systems, that cellular network mission enlarged to content delivery. With the introduction of Higher Order Modulation encoding scheme, such as 64-QAM, and carrier aggregation, 3G networks have been optimised for mobile broadband.

Following the direction undertaken in the development of 3G cellular systems, Evolved Universal Terrestrial Radio Access Network (E-UTRAN), often referred as Long Time Evolution (LTE), completed the trend of expanding

[1] Originally the acronym GSM was short form of Groupe Spcial Mobile

service provision beyond voice calls towards a multiservice air-interface. Although it was already a key aim of the 3G systems, LTE marked a big difference with its predecessors: it was designed from the beginning with the purpose of evolving the radio access technology under the assumption that all the services would be packed switched. Furthermore, the development of LTE came together with an evolution of the non-radio aspects of the whole system, under the name System Architecture Evolution (SAE), which included the Evolved Packet Core (EPC) network. Together, LTE and SAE comprised the Evolved Packet System (EPS) in which for the first time both the core network and the radio access were fully packed-switched [7].

Despite the improvements introduced with LTE, cellular networks were witnessing an exponential growth in data traffic demand. To cope with capacity and coverage challenges, in LTE advanced (LTE-A) the concept of Heterogeneous Network (HetNet) has been introduced. The HetNet concept brings a hierarchical structure among base stations with the legacy macro base stations network providing coverage for the users, and with a new type of low power nodes (picocells and femtocells) to provide users high data rates.

On this basis, the 5GPPP (5^{th} Generation Public Private Partnership) is laying down the fundamentals of the next generation cellular systems. Meanwhile, Machine Type Communications (MTC), transmissions generated by sensors and machinery without human intervention, and Internet of Things (IoT) impose a rule changing in network usage. Indeed, MTC traffic usually is characterised by sporadic transmissions of small size packets, an opposite trend respect to the HTC on which cellular networks are optimised on. As pointed out in [8], 5G systems are expected to offer native support to the traffic generated by devices without human intervention. From this point of view, some steps have already been undertaken, and beginning from LTE Release 13 3GPPP has introduced some optimisation for MTC. Narrowband IoT (NB-IoT) is a radio access system that has been designed on the basis of LTE, and therefore supports most of its functionalities but with many simplifications and some optimisation to support low-cost, low-power and low data-rate IoT services. NB-IoT supports only frequency division duplex, and at a higher level it includes idle mode mobility, extended discontinuous reception, power saving mode, paging, positioning based on the existing location services architecture, and access control. On the other hand, it has been optimised for small data transmissions, by enabling data transfer through the control plane, via Signaling Radio Bearer. In addition to that, NB-IoT supports sporadic data transmission introducing the possibility to suspend and resume Radio Resource Control connection, preventing the need to re-establish a new connection at each reporting instance [9].

The number of cellular connection generated by IoT devices, in 2015, has been estimated as a figure of 0.4 billion, and this number is expected to grow to 1.5 billion in 2021, equivalent to a yearly growth rate of 27% [10]. That is, the enormous amount of data collected by sensors (e.g., sensing devices that measure specific parameters to be sent to remote servers or other machines) combined with data produced daily from personal devices, smart-phones and wearable devices, has enabled a wide range of commercial opportunities. Moreover, with the boost of IoT devices a wide range of new challenges arise and the 5GPPP foresees the role of next generation mobile systems as a pillar for shaping a better society and enhancing the industry. Indeed, 5G networks are intended to support a high number of vertical industries, leading to a new conception of the industrial processes. In particular, 5GPPP has identified five primary vertical sectors of application and analysing requirements of each scenario it has been possible to define specific targets the 5G technology has to meet, the so-called Key Performance Indexes (KPI). The five vertical industry are as follow:

Health-care : the term m-health, mobile health-care, indicates all the procedures in health-care systems employing mobile networks to deliver services to patients. Advancements in sensing and communication technology have opened up new possibilities for remote health monitoring; for example, there already exists wearable IoT devices employed to sense patients physiological signals [11]. That is, 5G technology and IoT devices are going to provide a paradigm shift in health-care systems by providing a real-time, decentralized and highly personalized care. Moreover, the increased network capacity of 5G systems will also be fundamental in improving e-health, health-care practice supported by electronic processes, such as remote surgery. In a robotics-assisted tele-surgery scenario, the requirements for cooperative action coordination are immense and include the guaranteed and reliable availability of information from back-end databases and real-time data streams from a large variety of sources [12]. Derived from the aforementioned use cases, and from many others in [12], the main KPIs in health-care are can be summarized in: support to massive MTC, latency guarantees (e.g., 99.999% of packets needs to fulfil an upper latency of 30 ms in order to be usable for remote surgery), and increased security.

Energy : the energetic industries' history is over one century long, and it has developed through separate branches, namely energy generation and power supply. Production and supply have always been separated and easily managed: energy was produced in large central thermal and hydro generation stations following an easily predictable user demand. Nowadays, with the introduction renewable generation stations whose production is unpredictable and discontinuous, and combined with changing end-user energy demand energy industries are now required to be flexible both in producing and in distributing energy. Therefore, introducing such new dynamics into the rigid supplying infrastructure requires new Smart Grids [13], meaning that it will have to incorporate both smartness and communication capabilities, able to deal with unpredictable demands and offers; indeed users now are not only passive consumers but are also becoming producers of renewable energy. Based on systems criticality Smart Grid will require from low data-rate ($\leq$ 1kbps) and delay tolerant ($<$ 1s) communications between Smart Meters and secondary substations to high data-rate (in the range of Mbps and Gbps) and low latency ($\leq$ 5ms) combined with high reliability (packet loss $< 10^{-5}$) required in transmissions controlling the grid backbone [14].

Automotive : connected vehicle services have existed from a long time (e.g., car crash notifications and so on), but with the advent of autonomous driving cars it has gathered a lot of attention from the public opinion, the academy and the industry as well. Future vehicles will be aware of the surrounding environment both through sensing it and communicating with it. The paradigm of highly connected vehicles goes under the name of Vehicle to everything (V2X) communications and comprise of exchanging information with other vehicles (Vehicle to Vehicle, or V2V), communicating with the roadside infrastructure (Vehicle to Infrastructure, V2I), and Vehicle to Pedestrian (V2P) communications, among others. All such information automatically exchanged will then be exploited to provide new services for vehicle users, and in particular, the automotive industry foresees two main applications: automated driving, road safety and traffic efficiency services. As most of the attention and efforts have been caught by autonomous driving, and also because it takes to extremes the requirements, for autonomous driving of level 5[2] KPIs span from low latency ($\leq$ 30 ms) and high reliability (99.999 % $\leq$) to precise positioning ($\leq$ 0.3m) [15].

Media & Entertainment : over the last years M&E have been revolutionised, and a new type of needs are arising. Not only the content is available on-demand basis but, and most surprising, produced by the consumers themselves who become at the same time producers and consumers of multimedia contents. The biggest change driving all others in M&E is associated with the fact that the individuals themselves do not only passively consume, but interact, share, chat, talk, tweet, while walking, running, driving, commuting by subway and almost all the rest of the time. Therefore, network resources consumption due to M&E services is increasing more and more because of higher usage but also due to the larger amount of data exchanged per content (e.g. higher video and voice quality). In order to optimise the use of the network capacities, it is necessary that the on-demand resource parameters (i.e., latency, bandwidth, security, connectivity) are allocated and configured as required by the service. Hence, to support M&E industry, 5G net-

[2] Driving automation of level 5 means self-driving cars

works will have to provide high-quality QoS in terms of volume of data exchanged (UHD and 4K video streams), and ubiquitous coverage [16].

Smart Factory : this term is often referred to the evolution of traditional industrial plants with high integration of manufacturing robots and sensors to ease and enhance the production process. One of the pillars of Industry 4.0 [17] is the extensive use of Internet not only as an inexpensive channel to connect machines, devices, sensors, and people, but as a way to create products' functionalities. Features related to the capability of using the Internet as a source of information, such as advanced diagnostic and predictive maintenance based on large scale sensor data collection, e.g., from multiple locations and plants, will significantly reduce maintenance cost, increase asset availability, and create new usage-based business models [18]. Therefore, Smart Factories (SF) arise two challenging scenarios, namely massive MTC and time critical MTC. In the first case, SF will have to provide wireless connection to a multitude of sensors, employed to measure and ensure the quality of the production process– generating a ultra-dense scenario with a peak device density up to a hundred devices per square meter. On the other, time critical traffic will be generated by autonomous manufacturing robots and heavy machinery operating under the remote guidance of technicians, requiring ultra-low latencies and ultra-high reliability [5]

5G verticals industries cover the five most promising applications of next generation mobile systems, and even thought the number is limited, they raise many unsolved challenges. First of all, it is clear that 5G network will have to cope with requirements changing in time and on the base of which applications UEs are running. That is, next generation mobile systems are required to be flexible and easily programmable. The telecommunication community, to deal with such flexibility, developed the idea to apply the software defined networks (SDN) paradigm to cellular systems [19]. The SDN paradigm relies on the separation among user and control plane and on the introduction of novel network control functionalities, based on an abstract representation of the network [20]. Indeed, introducing a programmable abstraction layer, that controls networks' lower layers, opens an wide spread of new opportunities: separating the data plane and the control plane, makes network switches in the data plane simple packet forwarding devices, leaving a logically centralized software to control the behaviour of the entire network [21]. Network Function Visualisation (NFV), building on top of SDN, implements network functions through software virtualization techniques, that is NFV allows instantiating on-demand virtual appliances without the installation of new hardware [22]. An example of the synergy of such technologies is what is called Network Slicing [23]. It essentially extends the concept of channel, allocating spectrum resources to specific traffic classes with some QoS requirements.

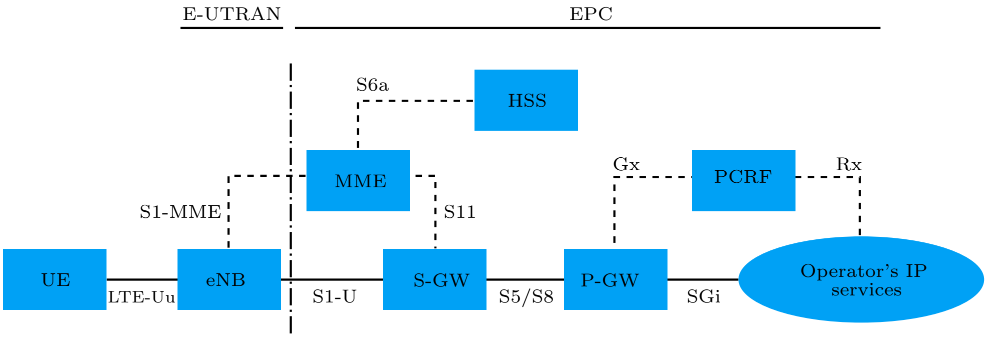

Figure 1.1: Evolved Packet System (EPS) architecture

Network Slicing provides an essential step forward, towards dynamic networks with a programmable design. Nonetheless, meeting the incredibly challenging KPIs posed by the 5GPPP is still questionable and for this reason it is of paramount importance to clearly understand how 4G and 5G system work. Indeed, a deep knowledge of network operational latencies and performance is primary to pinpoint system bottlenecks, leading future technological developments towards the KPIs fulfilment.

1.1 LTE system at a glance

System architecture

As mentioned before, the EPS encompasses both the LTE radio access and the non-radio aspects, which includes the EPC network. To provide a seamless IP connectivity between the User Equipment (UE) and the packed data network (PDN), such as the Internet, EPS uses the concept of EPS bearer. A bearer is an IP packet flow with a defined quality of service (QoS) routed between the gateway and the UE. Obviously, multiple bearers might be allocated to a single UE to provide either multiple IP flows with individual QoS requirement or connectivity to different PDNs. This is achieved by means of several EPS network elements, and Figure 1.1 depicts an overview of system architecture including both network elements and interfaces.

At an high level, the EPS comprises of two components, the core network (CN) and the access network. While the CN is structured in several logical entities, the access network is made up of essentially one node, the evolved Node-B (eNB), which connects to the UEs. The CN is responsible for the overall control of the UE and establishment of the bearers. Its main components are the PDN gateway (P-GW), the Serving gateway (S-GW) and the Mobility Management Entity (MME). In addition to that, EPC also includes further logical functions and entities such as the Policy Control and Changing Rules Function (PCRF) and the Home Subscriber Server (HSS).

The main purposes of EPS entities can be summarized as follows:

HSS stores users' SAE subscription data, such as access restrictions for roaming and EPS-subscribed QoS profile. It also holds information about the PDNs to which the user is allowed to connect, either in the form of an access point name or IP address. In addition, the HSS holds dynamic information such as the identity of the MME to which the user is currently attached or registered.

PCRF is responsible for QoS handling and charging functionalities in the Policy Control Enforcement Function (PCEF). The PCRF provides QoS authorization to define how a certain data flow have to be handled in the PCEF and ensures the accordance to users subscription

P-GW provides the interface to the PDN networks; it is responsible for UEs' IP address allocation, as well as QoS enforcement and flow-based charging according to rules from the PCRF. Moreover, the P-GW dispatches user's downlink IP packets into the different QoS-based bearers.

S-GW is the user-plane node connecting the EPC to the LTE RAN and it acts as a mobility anchor when terminals move across several eNBs. Further, S-GW manages the collection of information and statistics necessary for charging.

MME accounts for the control-plane node of the EPC: the functionalities operating between the EPC and the terminal are often referred to as the Non-Access Stratum (NAS), opposed to the Access Stratum (AS) which handles functionality operating between the terminal and the radio-access network. MME main functionalities might be categorised either as related to bearer or to the connection. Functionalities associated to bearer management include the establishment, maintenance and release of bearers. On the other hand, connection establishment and security management between network and the UE fall into the connection related functionalities.

The LTE radio-access network employs a flat architecture with a single type of node, the eNB, which provides all radio-related functions in one or several cells. Indeed, it is important to note that an eNB is not a physical node but a logical entity. One common implementation of an eNB is a multi-sector site, where a single base station handles transmissions in all the cells. As it can be seen in Figure 1.1, the eNB communicates with the EPC by means of the S1 interface. More specifically, it is connected to the S-GW by means of the S1 user-plane part, S1-U, and to the MME by means of the S1 control-plane part, S1-MME. Moreover, an eNB can be connected to multiple MMEs/S-GWs for load sharing and redundancy purposes. Eventually, the X2

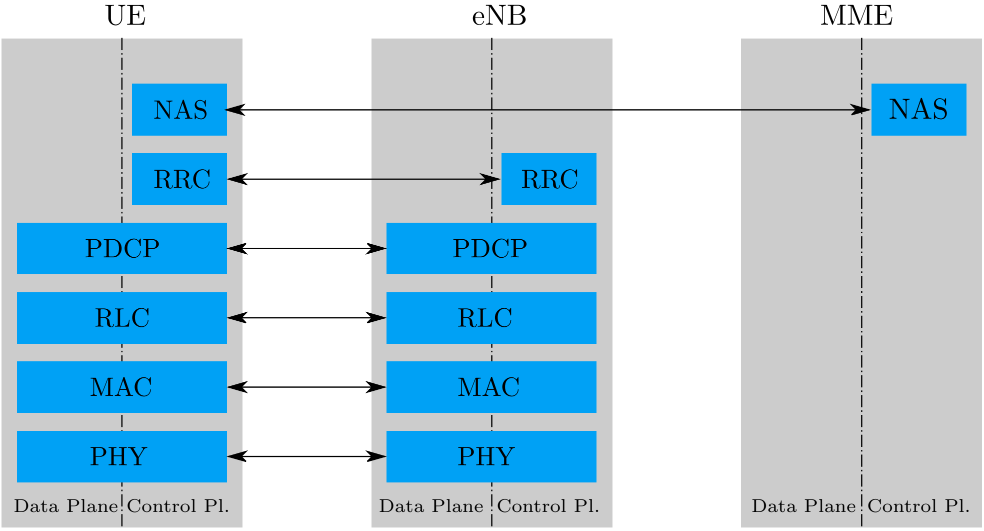

Figure 1.2: RAN protocol architecture

interface, connects eNBs to each other and is mainly used to support active-mode mobility.

System protocols

As stated before, the E-UTRAN is responsible for all radio-related functionalities that can be divided in:

Radio resource management: covers all functions related to the radio bearers, such as radio bearer control, radio admission control, radio mobility control, scheduling and dynamic allocation of resources to UEs in both uplink and downlink.

Header Compression: allows an efficient use of the radio interface by compressing the IP packet headers preventing significant overhead, especially for small packets.

Security: obviously, all data sent over the radio interface is encrypted.

Connectivity to the EPC: handles the signaling toward MME and the bearer path toward the S-GW.

Reflecting the separation of control and data flows in the network architecture, the RAN protocol architecture is divided in user plane and control plane. Figure 1.2 illustrates the RAN protocol architecture for both data and control plane[3]. It can be noted that many entities in Figure 1.2 are common

[3] In the figure, MME is only depicted for completeness: as discussed before, the MME is not part of the RAN.

to the two, therefore their description in many respects applies to both user and control plane. The main protocol entities of the RAN are:

Radio-Link Control (RLC) is responsible for segmentation/concatenation, retransmission handling, duplicate detection, and in-sequence delivery to higher layers. The RLC also provides services to the PDCP, in the form of radio bearers and, in a UE, one RLC entity is configured per each radio bearer.

Packet Data Convergence Protocol (PDCP) performs IP header compression and ciphering. For the control plane, PDCP is also responsible for integrity protection of the transmitted data, as well as in-sequence delivery and duplicate removal for handover. At the receiver side, the PDCP protocol performs the corresponding deciphering and decompression operations. Again, for a UE there is configured one PDCP entity per each radio bearer.

Medium-Access Control (MAC) handles multiplexing of logical channels, hybrid-ARQ retransmissions, and uplink and downlink scheduling. The scheduling functionality is located in the eNB for both uplink and downlink while the hybrid-ARQ protocol part is present in both the transmitting and receiving ends of the MAC protocol. The MAC provides to the upper layer, the RLC, services in the form of logical channels.

Physical Layer handles coding/decoding, modulation/demodulation, multi-antenna mapping, and other typical physical-layer functions. The physical layer offers services to the MAC layer in the form of transport channels, which are mapped onto the logical ones.

While the user-plane handles transmission of data generated at the UEs, control-plane protocols are in charge for connection set-up, security and mobility. Control messages sent to from the network to UEs might be originated either in the CN, and specifically from the MME, or from the Radio Resource Control (RRC), located in the eNB. The RRC is responsible for managing all the RAN-related procedures, which include:

- Broadcast of system information necessary for the UE synchronisation with a cell.

- Transmission of paging messages originating from the MME: this messages are used to notify the UE about incoming connection requests when it is not connected to a particular cell – that is, an UE in `RRC_IDLE` state.

- Connection management, including setting up bearers and mobility within LTE. This includes establishing an RRC context – that is, configuring the parameters necessary for communication between the terminal and the radio-access network.

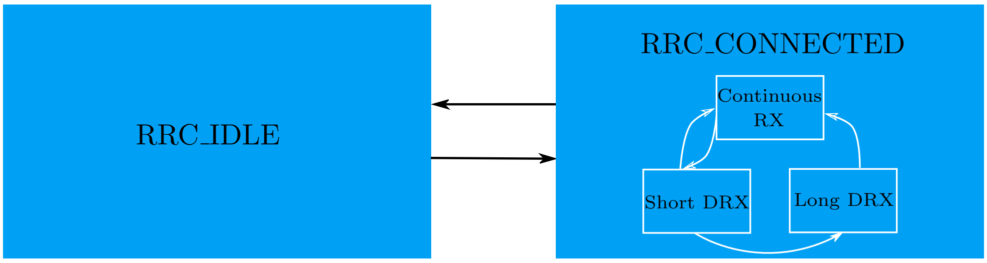

Figure 1.3: RRC state automata

- Mobility functions such as cell (re)selection, measurement configuration and reporting.
- Handling of UE capabilities: for backward compatibility and UE diversity, when a connection is established the terminal will announce its capabilities, as all UE are not capable of supporting all the functionality described in the LTE specifications.

1.2 Application requirements and RAN issues

Given the vertical scenarios diversity it is useful to notice that many of them, although with different nominal values, have common KPIs. That is, abstracting from 5GPPP KPIs specification 5G network will have to hit the following KPI:

- 1000 times higher mobile data volume per geographical area.
- 10 to 100 times more connected devices.
- 10 times to 100 times higher typical user data rate.
- 10 times lower energy consumption.
- End-to-End latency of < 1ms.

Keeping the LTE outline in mind, it can be noticed that many of the previous targets are tightly related to how resources are accessed and how those are managed. That is, although network capacity has always increased in each subsequent cellular system generation it is obviously bounded and the way UEs make use of it deeply impacts on latency, reliability and number of UEs served per unit of time.

As stated before, RRC protocol handles all the aspects related to the connection and, more in the details, it does it through a two state automata representing UEs' RRC state. As reported in Figure 1.3, an UE might be in two different RRC states depending on its recent activities, namely `RRC_CONNECTED` and `RRC_IDLE`. That is, UEs that are not transmitting and neither receiving data from a long time ($\sim$ tens of seconds) as well as UEs just switched on do not share a valid RRC context at the eNB – it compounds all the parameters necessary for communication between the two entities. Without an RRC context UEs in `RRC_IDLE` state do not belong to any specific cell and, in order to save energy, UEs sleep most of the time.

To keep up with possible incoming transmissions UEs periodically wake up in order to receive paging messages from the network. Since upload synchronisation in `RRC_IDLE` state is not maintained, data transmission is not possible. The only uplink communication available at this state is the one required to perform Random Access Procedure, which allows to re-establish uplink synchronization and eventually switch to `RRC_CONNECTED` state.

When moving to `RRC_CONNECTED` the RRC context needs to be established in both the radio-access network and the terminal, and as long as the uplink is synchronized, uplink transmission of user data and control signalling is possible. While in `RRC_CONNECTED` state UEs, to save battery, might be configured to perform cyclic sleep periods, called DRX cycles, when they are not actively using the network. The Discontinuous Reception (DRX) mechanism, for some fixed time-outs, establishes specific sleeping periods interleaved by timeslots in which downlink transmission, if any, will take place. At the reception of an incoming transmission from any of the DRX states UE moves to the continuous reception state, in which it continuously checks for incoming data. When the network activity ends and the inactivity period lasts longer than the time-out *drx-InactivityTimer* [24] UE moves to the first sleeping cycle. With the same mechanism based on inactivity periods, but with a different time-out (namely *drxShortCycleTimer*), UE moves to the *longDRX-Cycle* state. The *longDRX-Cycle* differs form the previous state not only for the time-out, which is the *longDRX-Cycle*, but also because in case of its expiration the UE eventually falls back in the `RRC_IDLE` state and network resources are released.

The impact of the RRC protocol has been mainly studied from the the energy saving perspective, evaluating the performance of the DRX cycle [25], [26] and how to optimize this mechanism according to a specific traffic shape [27], [28]. More recently, [29] explored idea of increase DRX flexibility introducing an higher number of sleep cycles and a probabilistic mechanism to move among them. Due to limitations in resources, the DRX cycle has gained great attention in IoT and MTC scenarios, [30]. Interestingly, most of the work on RRC and DRX cycle evaluate delays introduced by the mechanism, but in isolation from the rest of the system.

However, limiting our analysis to the RRC and DRX only, neglects the interactions that incurs among such mechanism and how resources are allocated, and more important how and when UEs release such resources. That is, based on the traffic characteristics the maximum time of inactivity allowed to an UE in `RRC_CONNECTED` state deeply impact on latency of next access to network resources. Indeed, if the UE sends traffic within this time window network resources are readily available. On the other hand, if it falls in the `RRC_IDLE` state uplink synchronisation with the network is lost and RRC context expired.

While in `RRC_IDLE` state the UE does not belong to any specific eNB and that requires the UE to perform the Random Access (RA) procedure. Through the RA procedure an UE requests a connection set-up with a specific eNB. In addition, RA pursue several purposes which, among the others, include:

- initial access, when establishing a radio link among UE and eNB – ascribable to the transition from `RRC_IDLE` towards `RRC_CONNECTED` state,
- re-establish a radio link after a failure,
- re-establish uplink synchronisation:
    - with a new eNB in case of an handover,
    - with the same eNB, in case of time alignment loss while in `RRC_CONNECTED` state.

Uplink synchronisation is one of the main objective of RA, but when establishing an initial radio link RA procedure assigns the UE a unique identifier, the Cell Radio Network Temporary Identifier (C-RNTI), which is the way eNB will refer to a specific UE after the RA procedure.

Since time synchronisation is critical in LTE transmissions, due to its orthogonal nature, RA procedure might be either contention-based or contention-free: the first one is commonly adopted for the initial access while the latter is reserved for time critical tasks, such as cell handover. The contention-based RA procedure compounds of four steps, and each step represent the transmission of a message on a specific physical channel. It is important to notice that, among all RA messages sent by the UE, only the first one happens on a dedicate channel[4] – while the other make use of normal uplink/downlink channels.

The first step in the contention-based RA procedure is initiated by the UE which sends a RA preamble signature (PS) to the eNB. Since UE and eNB do not yet share time synchronisation, PS transmission may happen only on a specific set of resources – called Resource Blocks (RBs) – in the time-frequency domain, the Physical Random Access Channel (PRACH), and the eNB broadcasts informations about its positioning to all terminals in a cell.

This first transmission announces the eNB the presence of a RA attempt, that is an UE willing to acquire some resources. Moreover, the PS transmission allows the eNB to estimate the delay between itself and the UE that will be used in the second step to adjust the uplink timing. An integral part of this step is the PS selection: in a LTE cell there are 64 PSs available divided in two set, one for contention-free and the other for contention-based RA procedure, and the UE randomly chooses one signature from the contention-based set.

[4] For a detailed discussion of LTE physical and logical channels, please refer [7]

As long as no any other UEs simultaneously pick the same PS signature the attempt will be detected from the eNB with an high probability. That is, even if the PS transmission happens without a contention in the resource block the eNB might not detect the PS because of an unsuitable transmission power setting at the UE. In the first RA attempt, the UE sets the transmission power by estimating path-loss on the cell downlink reference signal and then, in case of misdetection, RA procedure allows power ramping: TX power is gradually adjusted by a configurable step size for each unsuccessful RA attempt.

In case of PS detection, in the second step, the eNB will transmit a message on the Downlink Shared Channel (DL-SCH) containing the index of the detected PS, the timing correction computed at the PS reception and a temporary identifier TC-RNTI, Temporary C-RNTI, used to distinguish UE(s) associated with the specific PS transmission; if multiple copies of the same PS are detected, eNB response is valid for all the sending UEs. At this step, all the UEs involved in the RA procedure will monitor the control channels for the RA response for a predefined time window. All the UEs receiving a RA response within such time window proceed, while unanswered RA attempts will be considered failed. Obviously, in case of successful detection and in presence of contention colliding, UEs will react to the same RA response and the contention resolution will have to be resolved in subsequent steps.

Upon RA response reception UEs will adjust their time alignment and proceed with the third step. In this step, UE's transmission conveys both the TC-RNTI assigned with the previous message and either the C-RNTI, if the UE has already connected to the cell, or a unique identifier. Such redundant identifier is useful to eventually resolve collisions, if any has happened in the following step. In addition to that, this message is also used to begin the instantiation of the RRC context, that is the initial step for the transition toward the `RRC_CONNECTED` state.

The last step in the RA procedure consists of a downlink message for contention resolution, which again is sent over the DL-SCH. Note that, from the second step, multiple UEs performing simultaneous RA attempts with the same PS in the first step listen to the same response message in the second one and therefore they have the same temporary identifier. Hence, in the fourth step only UEs matching both the identifiers will declare the RA procedure successful and, if the terminal has not yet been assigned a C-RNTI, its TC-RNTI is promoted to the C-RNTI. On the other hand, collided UEs, as well as the one that have not received the downlink message within the appropriate time window, need to restart the procedure from the first step.

Contention-free RA procedure has been designed for all such situations that require a timely reaction, such as cell handover, positioning and uplink synchronization re-establishment upon downlink data arrival. In contrast to the contention based one, it requires the eNB to signal the UE what PS pick to avoid contention in the first RA step. Therefore the last two steps, which are useful to resolve the possibly conflicting UEs, are not required any more. The contention-free RA is, indeed, a two step procedure that saves precious time to operations sensitive to variable latency characterising the contention-based RA approach.

For its design, the first step in RA procedure is comparable to a multi-slotted Aloha channel [31] whose limitations and performance limits are well known from a long time [32]. In order to mitigate congestion in the RAN, and in particular in the early steps of RA procedure, in current networks system a prioritising mechanism is employed. Access Class Barring (ACB) controls incoming RA requests by setting a probabilistic threshold,that is the eNB broadcasts the threshold and UEs aiming to perform the RA draw a random value which either allows them to proceed, if it is greater than the threshold, or force them to perform a backoff period. Many works on ACB concerning Human Type Communications (HTC), such as [33], are focussed on the effect of bursty massive Machine Type Communications (MTC) on typical, i.e. human, users. On the other hand, with the great attention gathered by IoT, many works focus on the ACB mechanism, its performance and its possible evolutions. Authors in [34] perform a combined analysis of RA procedure and ACB mechanism while in [35] also Extended Access Barring (EAB) is taken into account. EAB extends the mechanism of ACB by introducing barring classes, each one with its corresponding configuration of barring probability and backoff time. The fundamental idea of EAB has been widely adopted to propose several priority based barring schemes such as [36], [37] and [38]. As for the HTC case many performance studies of the access phase, ACB plus RA procedure, have been done. The limitations of such works is that they consider this subsystem as isolated from the rest of network procedures. That is, they are limited in investigating how this subsystem impacts on the rest of the network and vice versa.

Chapter 2

Human type communications

2.1 Introduction

Our common experience is that wireless access networks perform poorly in very crowded environments. When we enjoy a football match or a rock concert in an extremely crowded stadium, and we would like to share our emotions with friends, we discover that placing a phone call or sending a short video, even posting a picture, is not possible, due to network congestion. When large numbers of networking experts gather at top international conferences in their field to discuss the latest research results, reading emails during the occasionally uninteresting talk is a problem, because the wireless access network is not able to sustain the very large number of email clients. These phenomena were quantitatively observed in [39], by collecting measurements over a tier-1 cellular network in the US during crowded events, and showing substantial performance degradations with respect to normal conditions.

The problem can only get worse. The Cisco Visual Networking Index forecast 2015-2020 [40] estimates that by 2020 the number of devices connected to IP networks will be more than three times as high as the world's population, generating an overall traffic of 2.3 ZB (equal to $2.3 \cdot 10^{21}$ B). Two thirds of this traffic will come from wireless devices, and 30% of the total will be generated by smartphones. The total mobile data traffic in 2020 will reach 30.5 EB (over $3 \cdot 10^{19}$ B) per month, with the highest volume in the Asia Pacific region, and the highest growth in the Middle East Africa region.

The 5G Infrastructure Public Private Partnership, in short 5G PPP, initiated by the European Commission, together with companies and research institutions of the field, shares those extreme visions [41]. Among the key challenges for 5G, a prominent position is given to the connection of over 7 trillion wireless devices serving over 7 billion people, and to the service of extremely crowded environments, such as a stadium, providing capacities of the order of 0.75 Tb/s over the stadium area, and an automated factory, comprising terminal densities up to 100 devices per m^2, and requiring sub-ms latency.

This chapter looks at the performance of wireless access networks in extremely crowded environments, focusing as an example on the case of a group of LTE cells covering a stadium. The main contributions are the following:

- We develop a simple analytical model that captures the key aspects of the behaviour of a cell and we use it to understand the main sources of poor performance.
- We validate the analytical model with detailed simulations, which prove the validity of the assumptions introduced for analytical tractability.
- We show how the model can be instrumental for a correct dimensioning of crowded cellular systems.
- We propose the adoption of device-to-device (D2D) communications [42] as a means to improve performance in extremely crowded environments, and we quantify the benefits that can be achieved with the D2D approach, showing that D2D clusters of size k are more beneficial to system performance than a costly increase of system capacity by a factor k (e.g., through the deployment of k more cells).

2.2 Scenario

The reference scenario that we use in our analysis is a large stadium, with capacity roughly comprised between 50 and 100 thousand spectators. Many such structures exist around the world, including, e.g.: the Maracana in Rio de Janeiro, the San Siro Stadium in Milan, the Santiago Bernabeu in Madrid, the Stade de France in Paris, the Wembley Stadium in London, the Camp Nou in Barcelona, the Rose Bowl in Pasadena, the Azteca Stadium in Mexico City, and the the Melbourne Cricket Ground, just to name a few. These structures regularly host important sport events, and occasionally also music concerts of famous rock and pop stars, and in the latter case the structure capacity grows by up to 50 thousand attendees.

Of course, such extraordinary numbers of people (terminals) imply a wide variety of services: spectators may want to send to their friends short videos or pictures of the event, may receive all sort of messages, as well as phone calls, and at the same time terminals may be involved in content downloads.

We primarily focus on services which imply the human intervention, such as the transmission of a picture with a messaging application. In this case, the human user is in the service loop, so that the basic sequence of the service operations is made of a request for the radio access network resources, possibly repeated several times, until resources are granted, then the use of the network resources, followed by a think time before the next service request.

We will see that in some cases the system bottleneck is in the request for the radio access network resources, mostly because cellular systems use an Aloha-like contention-based scheme for this operation. It may thus happen that, while the network resources are available, request collisions do not allow their allocation. Under these circumstances, a reduction of the number of requests is mandatory to restore acceptable network performance. This can be obtained by reducing the number of users who are allowed to issue requests, or by forcing users to *coalesce* during the request phase. This is where D2D comes into play. If end user terminals are allowed to form clusters or are instructed to form clusters by the network—through appropriate commands issued by the eNB—only one request is issued whenever multiple terminals of the same cluster require access to the network resources, as proposed in [43] for opportunistic scenarios.

2.3 Accessing Resources in LTE

In 3GPP standards like LTE, LTE-A and the upcoming 5G, end UEs have to proceed through the RA procedure to access data channels, if not already connected to the eNB. The access procedure begins with the UE sending a message on the Physical RACH (PRACH). Two types of random access procedures are defined: contention-based (implying an inherent risk of collision) and contention-free [24]. In each LTE cell a fixed number (64) of orthogonal preamble signatures (PSs) are available, and the operation of the two types of RACH procedure depends on a partitioning of these PSs between those for contention-based access and those reserved for allocation to specific UEs on a contention-free basis. The contention-free RA procedure is reserved to delay-sensitive cases, such as incoming traffic and handovers [7]. A contention-based random access PS is chosen at a UE to send a random access signal to the eNB. A conflict occurs if more than one UE use the same PS and time-frequency resources, resulting in undecodable messages at the eNB. The contention-based procedure consists of an exchange of four messages to set up a connection among UE and eNB.

Step 1: UE → eNB (Random Access Preamble). A first message conveys the randomly chosen RACH PS. The UE selects one of the available PSs and transmits it in a time-frequency slot. Several UEs may choose the same PS and the eNB may not be able to decode it. After the PS transmission, UE begins to monitor the downlink control channel (PDCCH) looking for an answer.

Step 2: UE ← eNB (Random Access Response – RAR). This message is sent by the eNB on the PDCCH, and addressed with an ID identifying the time-frequency slot in which the PS was decoded. Whether multiple UEs have collided or not, if no RAR matching message has been received within the RAR window, they must repeat the RACH procedure, after a backoff delay. The duration of such backoff is randomly chosen in the range $(0, B]$ where B is the maximum number of subframes in a backoff period, and varies in $(0 - 960]$ ms.

Step 3: UE → eNB (Scheduled Transmission). The UE that receives the RAR message responses a scheduled transmission request that includes the ID of the device and a radio resource control (RRC) connection request message on the uplink shared channel (ULSCH).

Step 4: UE ← eNB (Content Resolution). Contention resolution is released from the eNB on the PDSCH. This identifies that no conflict on the access procedure exists. The UE can transfer data to eNB.

Once a UE has successfully performed the RACH procedure, it owns an active duplex connection and is in the `RRC_CONNECTED` state. Keeping a connection running requires that the eNB reserves physical resources devoted to this connection, even if there is no traffic available for the intended UE. Therefore the eNB can handle only a limited number of connected devices. Moreover, the connected UE has to continuously verify if there is any incoming traffic, monitoring control channels, and therefore incurs high battery consumption.

As long as the communication is alive, the UE remains in the `RRC_CONNECTED` state, but after an inactivity period, it begins to perform sleep cycles, from which it can return to the `RRC_CONNECTED` state without performing the contention-based RACH procedure.

Since the above-described access mechanism is based on a multichannel slotted Aloha, each PS representing an Aloha channel, its performance degrade beyond the threshold of 1 request/slot per PS. Hence, in dense scenarios, congestion can happen and become a system bottleneck. To alleviate congestion, state of the art solutions adopt the Access Class Barring (ACB) mechanism, which segments devices in several classes [35]. Devices within each class are managed through two parameters, namely the access barring probability and the barring time. With ACB, devices that are ready to attempt a random access are probabilistically *barred*, and barred devices wait for a barring time before making another barring decision, i.e., a device can be barred multiple times in a row. ACB is effective in smoothing peaks of access requests, but it does not change the RACH load under steady-state conditions. Moreover, ACB introduces a stochastic delay.

2.4 Analytical Model

We model the operations of n end-user terminal devices located in the same cell, under the coverage of one eNB. The notation used in this chapter is

Table 2.1: Notation and Cell Parameters used in Section 2.6

Quantity	Symbol	Value
Number of devices (or clusters)	n	
eNB capacity	C	150–1500 [Mb/s]
Network max accepted requests	M	200
Number of Random Access preambles	N	54
Slot time	τ	0.01 [s]
Backoff time RACH	B_0	av. 0.15 [s]
Backoff time Network	B_1	av. 1 [s]
ACB access probability	p_a	0.05–0.95
ACB barring time	B_a	av. 4-512s
Transmitted data volume	F_S	av. 1.5 [MB]
Transmission time	S	
Think time	T_{TH}	av. 30 [s]
Device uplink speed limit	R	
Probability to skip RACH procedure	p_J	≤ 0.5
Access delay	A_T	
Thinking subsystem throughput	λ	
Network subsystem throughput	ξ	
Random Access subsystem input	γ	
Arrival rate at Network subsystem	σ	
Collision probability	p_C	
Rejection Probability	p_B	

summarized in Table 2.1.

Each device generates uplink transmission requests according to the 3GPP contention-based RACH procedure briefly described in Section 2.3 to obtain a transmission grant from the eNB. We account for the fact that the establishment of downlink flows might provide the devices with extra opportunities to obtain transmission grants, skipping contention through the contention-free RACH procedure.

In the following, we derive a model for access requests and service operation in the cell, and show how to compute network utilization, access delay, and in general how to assess the behavior of the system as a function of the number of devices in the cell, for a given eNB configuration (in terms of capacity, number of RACH channels, RACH slot duration, backoffs experienced upon failed RACH procedures, etc.).

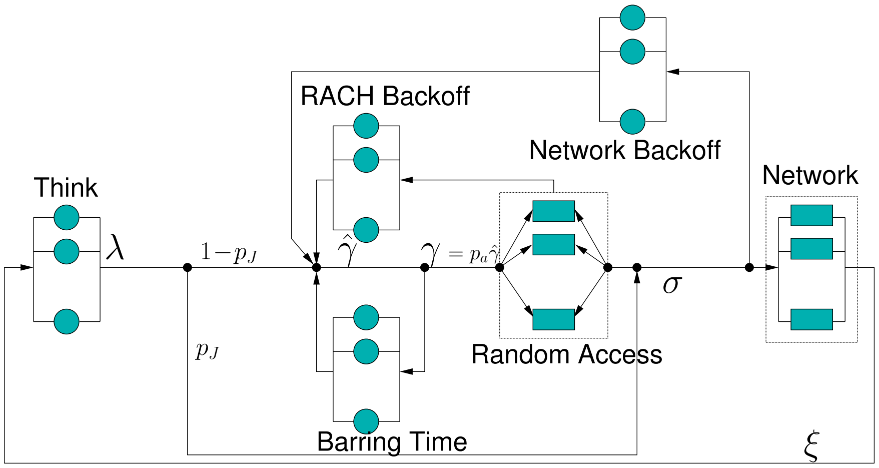

Figure 2.1: Closed queueing network model of a cell

Closed representation of the system

The eNB has uplink capacity C, in bits per second, and can share its capacity among at most M devices at a time (i.e., there can be up to M devices in state `RRC_CONNECTED`). The number of RACH channels (i.e., orthogonal preamble signatures - PS) available for Random Access is N and the interval between two consecutive Random Access Opportunities (RAOs) is τ. If during τ a single device selects a given RACH channel, then the RACH procedure is successful, otherwise the RACH channel is either unused or a collision happens with multiple devices attempting to use the same PS.

A RACH collision results in a random backoff B_0, after which a RACH retry follows. In case of successful RACH procedure, the device is granted transmission only if there are less than M devices under service at the eNB, otherwise the device goes through a random backoff B_1 followed by another RACH procedure. The model also considers ACB with uniform access probability p_a for all classes, and barring time with average duration $E[B_a]$.

For what concerns the traffic generated by end-user terminals, we consider human-operated wireless devices, and assume that each device produces a new transmission request, with random data volume F_S, only after its previous request has been served. More specifically, upon service completion, we assume that the devices enters a "think time" period with random duration T_{TH} before generating the next request. Unless otherwise specified, the average service time $E\left[S\right]$ only depends on C, M and the average value $E\left[F_S\right]$, i.e., we assume that the serving speed is fixed and equal to C/M, so that $E\left[S\right] = \frac{M \cdot E[F_S]}{C}$. However, we will also show how to account for the equal sharing of the eNB capacity among the actual number of devices under service in the system, and for service speeds limited by a device uplink speed R.

The resulting system model is depicted in Fig. 2.1. The model comprises

6 main components: *i)* Think, representing the end-user think time between the end of a service and the generation of a new access request; this is modeled with an infinite server queue with exponential service time with average $E\left[T_{TH}\right]$; *ii)* Random Access, representing the RACH contention-based procedure; this is modeled as a set of N parallel slotted Aloha channels, receiving each $\frac{1}{N}$ of the total load offered to the RACH; the slot duration for any of the N slotted Aloha channels is τ; *iii)* Barring Time, which models ACB operation as an infinite server with average service time $E[B_a]$ affecting a portion $1-p_a$ of the flow directed to the Random Access; *iv)* Network, representing the eNB resources, modeled as an M/G/M/0 queue with average service time $E\left[S\right]$. The Network queue is fed by the output of the Random Access subsystem and by the requests that skip the Random Access because of transmission opportunities generated by downlink traffic requests; these are modeled by means of the "jump probability" p_J, which is the probability to perform the contention-free RA procedure, and access directly the eNB resources. *v)* Network Backoff, and *vi)* RACH Backoff, representing the two backoffs, which are modeled by means of infinite server queues with exponential service times, with averages $E\left[B_0\right]$ and $E\left[B_1\right]$, respectively.

Fig. 2.1 also shows that the system is closed, i.e., the population is finite, with the number of customers fixed to n. We denote by λ the output of the Think subsystem, and by ξ the output of the Network subsystem. Because of the closed structure of the system, $\lambda = \xi$. We indicate with γ the total arrival rate at the N RACH channels in the Random Access subsystem, and we assume that RACH requests follow N parallel and i.i.d. Poisson processes with intensity $\frac{\gamma}{N}$. Although devices decide to send RACH requests asynchronously, such requests are cumulated over τ seconds and physically sent at the same time over the same frequency band. Thus, the successful output of each of the N RACH channels is that of a slotted Aloha system with $\frac{\gamma\tau}{N}$ arrivals per slot, which is given by $\frac{\gamma\tau}{N}e^{-\frac{\gamma\tau}{N}}$ successes per slot, as known from the the standard analysis of multichannel slotted Aloha [44]. The maximum throughput per slot of such multichannel slotted Aloha system is $\frac{N}{e}$, which is achieved for $\gamma\tau = N$.

With the above, the arrival rate at the network service is $\sigma = \gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda$, the arrival rate at the RACH backoff B_0 is $\gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right)$, and the one at the Network backoff B_1 is $p_B\sigma$, where p_B is the blocking probability, given by the Erlang-B formula with M servers and load $\rho = E\left[S\right]\sigma$. The load accepted and served by the network service is $\xi = \left(1-p_B\right)\sigma$. For analytical tractability, we introduce the simplifying assumption that all arrival processes are homogeneous and independent Poisson processes.

In the described system, quantities λ, σ, and ξ (and therefore also ρ and p_B) are functions of γ. It is possible to write a recursive equation in γ by considering that γ is $\hat{\gamma}$ minus what enters the Barring Time block. $\hat{\gamma}$ results from the sum of four arrival rates: $\lambda\left(1-p_J\right)$ from the Think subsystem, the output of backoffs B_0 and B_1, plus the recycle caused by ACB:

$$\hat{\gamma}=\lambda\left(1-p_J\right)+\gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right)+p_B\left(\gamma e^{-\frac{\gamma\tau}{N}}+p_J\lambda\right)+(1-p_a)\hat{\gamma},$$

which, combined with $\gamma = p_a\hat{\gamma}$, yields a recursive expression for γ, which does not depend on ACB operation at all:

$$\gamma = \lambda\left(1-p_J\right) + \gamma\left(1-e^{-\frac{\gamma\tau}{N}}\right) + p_B\left(\gamma e^{-\frac{\gamma\tau}{N}}+p_J\lambda\right). \tag{2.1}$$

The recursive expression (2.1) has two unknowns: γ and λ (note that p_B can be written as function of ξ, and $\xi = \lambda$). Unfortunately, this expression is not enough to identify the operating point of the system, because it contains no dependency on the population size n. However, to introduce n in the loop, and remove λ, we can apply Little's law to different blocks in the modeled system, as presented in the following.

Solving system equations requires iteration, whose proof of convergence is provided in Section 2.4.

Dependence on the population size n

From the model described in the previous subsection, we can easily derive the expressions for the network utilization, the number of devices under service and in any of the system blocks depicted in Fig. 2.1, the time of a complete cycle between two transmissions, and the delay to access the service. All these quantities can be expressed as function of γ, and γ can be expressed as function of the population size n.

Utilization and distribution of devices. The network utilization ξ is equal to $\sigma\left(1-p_B\right) = \left(\gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda\right)\left(1-p_B\right)$. Therefore, since $\xi = \lambda$, it is immediate to obtain the following expressions for ξ, λ, σ and ρ:

$$\xi = \lambda = \frac{\gamma e^{-\frac{\gamma\tau}{N}}\left(1-p_B\right)}{1-p_J\left(1-p_B\right)}; \tag{2.2}$$

$$\sigma = \frac{\xi}{1-p_B} = \frac{\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J\left(1-p_B\right)}; \tag{2.3}$$

$$\rho = E\left[S\right]\sigma = \frac{E\left[S\right]\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J\left(1-p_B\right)}. \tag{2.4}$$

Note that, since ρ in (2.4) only depends on γ and p_B, we have that p_B actually depends only on γ. Thus, all the quantities representing arrival rates in the system model are functions of γ only, for fixed values of the other system parameters.

The number of devices under service, that cannot exceed M, is computed by applying Little's law at the Network, i.e., $n_S = \xi E\left[S\right] \leq M$, which also implies that utilization cannot exceed $M/E[S]$. Similarly, the average number of devices in Think is proportional to the average number of devices under service, i.e., $n_{TH} = \xi E\left[T_{TH}\right] = n_S \frac{E[T_{TH}]}{E[S]}$.

The rest of the devices $n - n_S - n_{TH}$ are attempting access, either waiting for the next RACH opportunity (including after a barring event) or in one of the backoff queues, so applying again Little's law we obtain:

$$n - n_S - n_{TH} = \gamma\left(\frac{\tau}{2} + \frac{1-p_a}{p_a}E[B_a]\right) + \\ + \gamma\left(1 - e^{-\frac{\gamma\tau}{N}}\right)E\left[B_0\right] + \frac{p_B \gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J\left(1-p_B\right)}E\left[B_1\right],$$

where the average delay incurred in a RACH attempt is computed as half of the slot duration because of the Poisson arrival assumption. The total number of devices in the network can therefore be expressed as a function of γ:

$$n = \gamma\left(\frac{\tau}{2} + \frac{1-p_a}{p_a}E[B_a]\right) + \gamma\left(1 - e^{-\frac{\gamma\tau}{N}}\right)E\left[B_0\right] + E\left[B_1\right] \\ \cdot \underbrace{\frac{p_B\ \gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J\left(1-p_B\right)}}_{\frac{p_B}{1-p_B}\xi} + \left(E\left[S\right] + E\left[T_{TH}\right]\right)\underbrace{\frac{\gamma e^{-\frac{\gamma\tau}{N}}\left(1-p_B\right)}{1-p_J\left(1-p_B\right)}}_{\xi} \qquad (2.5)$$

This is a monotonic relation between n and γ, which can be inverted (although not in closed form) to express γ as a function of n. However, we have seen that all quantities of interest in the system are functions of γ, so that we can conclude that they are eventually functions of n only, i.e., of the device population's size.

Cycle duration. The average time for a complete cycle in the system (e.g., the cycle between two consecutive service completions) is denoted by $E\left[T_{cycle}\right]$ and can be easily computed from the model of Fig. 2.1, by considering that: *i)* the probability to collide on a slotted Aloha representing the RACH channel with Poisson arrivals of intensity $\frac{\gamma\tau}{N}$ arrivals per slot is $p_C = 1 - e^{-\frac{\gamma\tau}{N}}$, and *ii)* collisions are assumed to be independent. Hence, we can write that:

$$E\left[T_{cycle}\right] = \frac{p_B}{1-p_B}E\left[B_1\right] + E\left[S\right] + E\left[T_{TH}\right] + \left(\frac{1}{1-p_B} - p_J\right) \\ \cdot \left[e^{\frac{\gamma\tau}{N}}\left(\frac{1-p_a}{p_a}E[B_a] + \frac{\tau}{2} + E\left[B_0\right]\right) - E[B_0]\right]. \qquad (2.6)$$

The term in brackets in (2.6) is the average time spent in the loop formed by the RACH and the RACH backoff blocks, which has to be counted $\frac{1}{1-p_B}$ times on average (i.e., the average number of Bernoulli trials before a success, including the success that occurs when a device finds the Network available), except for the case in which a request skips the RACH, which occurs with probability p_J. The quantity $\frac{1-p_a}{p_a}E[B_a] + \frac{\tau}{2} + E\left[B_0\right]$ is the time to complete one of such RACH loops—which includes, on average, $\frac{1-p_a}{p_a}$ passages through the ACB backoff—and there are, on average, $\frac{p_C}{1-p_C} = e^{\frac{\gamma\tau}{N}}-1$ collisions before a successful RACH attempt (in which case the RACH backoff does not occur). The network backoff is traversed only after a failed network access (i.e., $\frac{p_B}{1-p_B}$ consecutive times, on average), whilst the Network and Think subsystems are traversed only once per cycle. $E\left[T_{cycle}\right]$ depends on γ since we have shown that p_B also depends on γ. So, using (2.5) we conclude that $E\left[T_{cycle}\right]$ can be written as function of n.

Access delay. The access delay, indicated as $E\left[A_T\right]$, is the time spent in a cycle, excluding the think time and the service, and is therefore easily obtained from (2.6):

$$E\left[A_T\right] = E\left[T_{cycle}\right] - E[S] - E\left[T_{TH}\right]. \tag{2.7}$$

As for $E\left[T_{cycle}\right]$, this is an expression that depends on γ, and therefore on n. An alternative expression for $E\left[A_T\right]$ is obtained by applying Little's law to the part of the system that excludes network service and think time:

$$E\left[A_T\right] = \frac{n - n_S - n_{TH}}{\lambda}. \tag{2.8}$$

Since $\lambda = \xi$, (2.8) reveals that the access delay is (practically) linear with the population size if ξ is (roughly) constant in a range of n, so that also n_S and n_{TH} are constant. As we will show later, such range exists if the Network saturates before the Random Access. That range is very relevant, because any point in it leads to maximal utilization.

QoE indexes

We use two indexes to express the quality of experience (QoE) for the end-user. The first index η_S compares the service time with the time spent waiting before service starts, and it decreases with the access delay:

$$\eta_S := \frac{E[S]}{E[S] + E\left[A_T\right]}. \tag{2.9}$$

The second index is η_A, which is inversely proportional to the service time and fades exponentially with the access delay. Service time and access delay used in η_A are normalized to their values obtained with the smallest population n that causes the presence of M devices under service (denoted by n'):

$$\eta_A := \frac{E[S]|_{n=n'}}{E[S]} e^{-\frac{E\left[A_T\right]}{E\left[A_T\right]|_{n=n'}}}. \tag{2.10}$$

Differently form η_S, index η_A is very sensitive to relative increases of delay rather than to absolute increases.

Both the indexes defined in (2.9) and (2.10) represent the end-users satisfaction resulting from the comparison between the time spent using the network and the time required to get access to it. Unlike η_S which is a linear combination of the measures of interest, η_A relates the performance perceived by the end-users with the performances of the system at its desired operational point. Therefore, (2.10) express the percieved QoE in terms of how the system improve or worsen with respect to the operational point n'

Analysis with D2D support

When D2D is used to alleviate RACH contention problems, terminal clusters come into play, each of them behaving as a single device. Thus, we can use the same formulas as above, with n, n_S, n_{TH} denoting the number of clusters in the system, under service and in think time, respectively. Similarly, all arrivals and services refer to clusters. The main effect of clusters is the reduced load to the Random Access. The impact is non-linear because γ does not scale linearly with n.

Cluster formation. Clusters form either spontaneously, when a device announces its willingness to wait for other users to join in a random access attempt, or under the control of the eNB, when RACH collision probability becomes problematic.

Cluster service time. If k is the average cluster size, i.e., the average number of devices in a cluster, the service time becomes k times higher than for the case without clusters.

Cluster think time. In the case of clustered RACH access, the think time increases as well. Indeed, for a cluster, the think time corresponds to the think time of the device that initiates the cluster, plus the time needed for the other members to join. However, assuming a very high density of devices, forming a cluster of a few units is very quick. For instance, clusters of k devices have an average think time of $E\left[T_{TH}\right] + \sum_{i=1}^{k-1} \frac{E[T_{TH}]}{m-i} \simeq E\left[T_{TH}\right]\left(1 + \frac{k-1}{m}\right)$, where $m \gg k$ is the number of devices that can join the cluster. In practice, the think time increase is negligible in crowded environments, and so we ignore it in the model.

Impact of resource sharing under non-saturated conditions

If we consider that the eNB resources can be shared by active connections, it is obvious that underloaded systems offer higher rates to the active devices. Therefore, the analysis proposed so far is valid in the region in which the Network subsystem is fully loaded, while it contains an approximation elsewhere. To fix this approximation, let us consider a Network subsystem that shares equally its resources among the connected devices, up to a rate R that can be interpreted as the maximum rate achievable by a device or as the maximum rate specified in the user's service level agreement.

$$E\left[S\right] = E\left[F_S\right] \max\left\{\frac{1}{R}, \frac{1}{C/n_S}\right\}. \tag{2.11}$$

Note that $E\left[S\right]$ is equal to $\frac{E[F_S]}{R}$ when the number of devices under service is not enough to saturate the Network subsystem. The adaptation of $E[S]$ to the number of devices under service introduces a further element of dependence on n, and a non-linearity. However, the impact on system performance is quite limited and can be neglected, as shown next in Section 2.5.

Convergence

With constant $E[S]$. For a given user population n, we find iteratively γ, p_B and λ, and then compute all other quantities. More specifically, we use two nested iterations. First, we note that we can evaluate the value of n that corresponds to an input γ by using (2.5) and compare it with the target. Since we have a monotonic relation between n and γ, it is enough to start with a random value of γ and then keep increasing (or decreasing) such input value until the n computed with (2.5) matches the target value of n. However, the evaluation of (2.5) requires to know the value of p_B. This can be computed with a nested iteration, for a fixed γ. Specifically, we use the following two equations to find p_B (and λ at the same time, as a byproduct):

$$\begin{cases} p_B & = \frac{\left[E[S]\left(\gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda\right)\right]^M / M!}{\sum_{j=0}^{M}\left[E[S]\left(\gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda\right)\right]^j / j!}; \\ \lambda & = \frac{\gamma e^{-\frac{\gamma\tau}{N}}(1-p_B)}{1-p_J(1-p_B)}. \end{cases} \tag{2.12}$$

On the one hand, increases of λ correspond to increases of p_B in the first expression in (2.12) because, for fixed γ, λ is proportional to the load of the Network subsystem $\rho = E[S]\left(\gamma e^{-\frac{\gamma\tau}{N}} + p_J\lambda\right)$. On the other hand, increases of p_B correspond to monotonic decreases of λ computed with the second expression in (2.12). Therefore, it is enough to start with input $\lambda = 0$, which produces output $p_B \geq 0$ and $\lambda \geq 0$ and keep increasing the input value of λ, which in turn increases p_B and produces an output λ monotonically decreasing, until the value of input and output for λ coincide.

With $E[S]$ depending on n_S. In this case, to the iteration on γ with nested iteration on p_B, we need to add a third nested iteration on $E[S]$. For fixed γ and λ, $E[S]$ has to satisfy the following recursive equation obtained from (2.11) with a number of devices under service computed with Little as $n_S = \lambda E[S]$:

$$E[S] = E\left[F_S\right] \max\left\{\frac{1}{R}, \frac{\lambda E[S]}{C}\right\}. \tag{2.13}$$

Here, considering that $\lambda E\left[F_S\right]$ is the throughput of the Network subsystem in bits per second and cannot exceed C, it is clear that $\lambda/C \leq 1$ as soon as at least one bit has to be transferred in a message. Therefore, the L.H.S of (2.13) increases with slope 1 while the R.H.S. is constant for small values of $E[S]$ and then increases with a slope smaller than 1. The result is that (2.13) admits only one solution that can be found by searching from its minimum value $\frac{E[F_S]}{R}$ and with small increases, until $\frac{E[F_S]M}{C}$, which is the maximum for $E[S]$. The value found for $E[S]$ is then used in (2.12) to update λ. Since $E[S]$ is non-decreasing in λ and p_B increases with both $E[S]$ and λ, the convergence of (2.12) is guaranteed also in this case by starting from $\lambda = 0$ and searching by increases.

2.5 System Behavior

Here we study the bottlenecks of the system, point out some notable points in the performance curves, and analyze how performance is affected by the number of devices present in the cell and by the introduction of D2D-based clusters.

Bottlenecks

The model depicted in Fig. 2.1 has two potential bottlenecks: the Random Access and the Network subsystems. The former filters network access attempts, and asymptotically prevents any network request as γ grows with the population n. The Network subsystem has finite capacity, and therefore cannot serve more than M simultaneous requests.

Fig. 2.2 shows a typical case in which the maximum throughput of the Random Access is below the capacity of the Network subsystem, and thus is the only bottleneck, for all population sizes. In this case, the Network subsystem throughput ξ and the input σ of the Network subsystem are equal, since the blocking probability p_B is negligible. From (2.7), the access delay becomes a linear affine function of $e^{\frac{\gamma\tau}{N}}$, and therefore grows with e^n. However, as shown in Fig. 2.2, a system in which the Random Access saturates before the Network subsystem does not suffer high delay. The range of device populations that roughly maximizes network utilization is quite narrow, and corresponds to a rather small interval around the peak efficiency of a multichannel slotted Aloha system, i.e., to values of n close to the one that yields $\gamma\tau = N$ (about 400 in the figure). This is the context that was previously analysed in the literature for the case of machine to machine (M2M) communications [45], with a system model similar to ours, and studied by means of a Markov chain. Here we focus on the more complex two-bottleneck case, in which the Network subsystem saturates before the Random Access, which is typical for the stadium scenario.

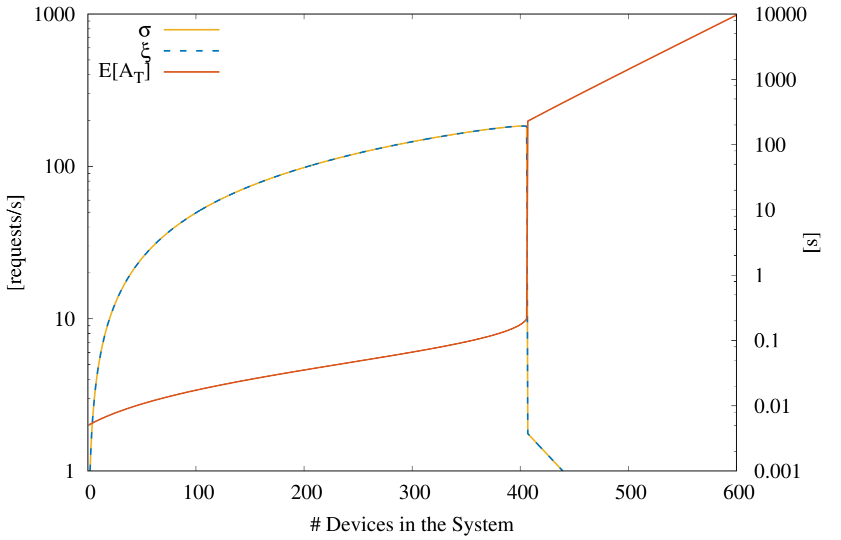

Figure 2.2: Random Access-limited model behavior. Left scale for σ and ξ, right scale for access delay.

Fig. 2.3 shows an example of the model behaviour when both the Random Access and the Network subsystem can become the system bottleneck. Indeed, the Network subsystem is a bottleneck for lower values of population size, until the Random Access reaches success probabilities so low to starve the Network subsystem. In the figure we can identify three operational regions. In the first region (low number of devices and low load: roughly below 550 devices for the specific example), p_B and p_C are close to zero, $\sigma \simeq \xi$, and the delay is practically negligible. In the second region (shaded in the figure, roughly from 550 to 7500 users), the throughput of the Network subsystem is constant, while σ follows the familiar bell-shaped curve of slotted Aloha, and the delay grows linearly with the user population, as visible from (2.8) (note the logarithmic vertical scale on the right). In the third region, p_B is negligible again, so that $\sigma \simeq \xi$ like in the first region, but the delay now grows exponentially with γ, and therefore with n. Out of such three regions, only the second one is desirable for system operation, since the Network subsystem resources are not wasted, and delay scales linearly with the number of devices in the cell.

Notable operational points

Random Access saturates first. In this case, $p_B \simeq 0$, so that $\xi \simeq \frac{\gamma e^{-\frac{\gamma\tau}{N}}}{1-p_J}$, the average number of devices in service is $E\left[n_S\right] = \xi E[S] < M$, and the average service time has to be constant (as remarked in Section 2.4, $E\left[S\right]$ must be equal to $\frac{E[F_S]}{R}$, otherwise the Network subsystem saturates). The network throughput is maximal when the output of the Random Access is maximal.

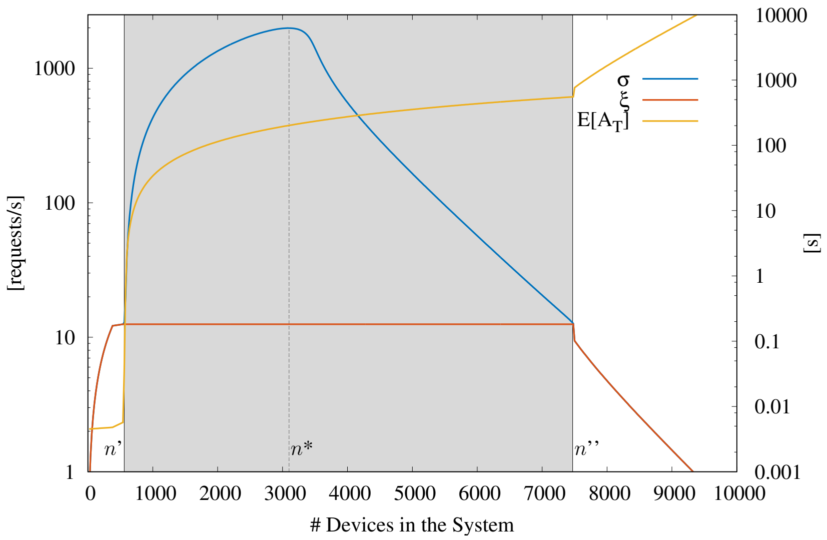

Figure 2.3: Model behavior with Network subsystem saturation. Left scale for σ and ξ, right scale for access delay.

This occurs for a number n^* of users that results in $\gamma = \frac{N}{\tau}$. From (2.5) we obtain the approximation (linear in N):

$$n^* \simeq \frac{N}{\tau}\left[\frac{E[S]+E\left[T_{TH}\right]}{e(1-p_J)}+(1-e^{-1})E\left[B_0\right]+\frac{1-p_a}{p_a}E[B_a]\right]+\frac{N}{2}.$$

With Network saturation. In this case, to characterize the behavior of the system in the three operational regions shown in Fig. 2.3, in addition to n^* we characterize n' and n'', i.e., the values of n that correspond to the first and the second knee of the curve representing ξ vs. n.

Note that $n' \leq n^* \leq n''$, and the throughput of the Network subsystem is constant and equal to $\frac{C}{E[F_S]}$ for all values in the interval $[n', n'']$. Therefore, (2.1) reduces to:

$$\gamma e^{-\frac{\gamma\tau}{N}} = \frac{C}{E\left[F_S\right]}\frac{1-p_J\left(1-p_B\right)}{1-p_B}, \quad \forall\gamma \mid n \in [n', n''].$$

At the extremes of the considered interval $[n', n'']$, the Network subsystem has exactly enough resources to satisfy the demand, so that we can consider $p_B \simeq 0$:

$$\gamma e^{-\frac{\gamma\tau}{N}} \simeq \frac{C}{E\left[F_S\right]}\left(1-p_J\right), \quad \gamma \mid n \in \{n', n''\}. \tag{2.14}$$

Considering that the l.h.s. of (2.14) is a non-negative continuous function of γ that starts from 0, grows until it reaches the value $\frac{N}{e\tau}$ at $\gamma = \frac{N}{\tau}$ and then decreases asymptotically to 0, expression (2.14) admits two (possibly coinciding) real solutions only if $\frac{C}{E[F_S]}(1-p_J) \leq \frac{N}{e\tau}$. So, a range of values of n such that the throughput of the Network subsystem is constant and maximal exists if and only if

$$N \geq \frac{e\tau C}{E\left[F_S\right]}\left(1-p_J\right). \tag{2.15}$$

The distance between the zeros of γ in (2.14) decreases logarithmically with C increasing (and with p_J decreasing). Since γ is monotone with respect to n, this means that the interval $[n', n'']$ becomes smaller with larger capacities C (and with smaller probabilities p_J), and $n' = n'' = n^*$ when (2.15) holds as equality. If (2.15) does not hold, the Network subsystem cannot saturate, and we fall back to the Random Access-limited scenario of Fig. 2.2.

The above condition also tells that the number of RACH channels needed to allow network saturation scales linearly with the capacity of the network and with $(1-p_J)$.

The notable points described above and the asymptotic behavior of γ vs. n can be approximated by means of the following closed form expressions that can be readily derived, as shown next.

Approximated values for the notable points. The value of n' can be approximated in closed form if the network saturation throughput is much less than the maximum Random Access throughput. In this case, there is neither network blocking nor collisions (i.e., $p_B \simeq 0$ and $p_C \simeq 0$) at n' and $\xi = \lambda = \sigma = \frac{C}{E[F_S]}$ while $\sigma \simeq \gamma + p_J\lambda$. The result is that (2.5) reduces to:

$$n' \simeq \frac{C}{E\left[F_S\right]}\left[E[S] + E\left[T_{TH}\right] + (1-p_J)\left(\frac{\tau}{2} + \frac{1-p_a}{p_a}E[B_a]\right)\right]; \tag{2.16}$$

Therefore, n' scales linearly with network capacity, slot duration, and probability of skipping the Random Access.

For what concerns the value of n^*, we can again use (2.5) with $\gamma\tau = N$, $\xi = \frac{C}{E[F_S]}$, $\rho = E[S]\left(\frac{N}{e\tau} + p_J\frac{C}{E[F_S]}\right)$. Considering that $\rho \gg M$ if the network saturates well before the Random Access, then $p_B \simeq 1 - \frac{M}{\rho}$ and $n_S \simeq M$. In conclusion, the following approximation holds:

$$\begin{aligned} n^* \simeq\, & M + E\left[T_{TH}\right]\frac{C}{E\left[F_S\right]} + \frac{N}{2} + \frac{N}{\tau}\left(1-e^{-1}\right)E\left[B_0\right] \\ & + \frac{N}{\tau}\frac{1-p_a}{p_a}E[B_a] + \left[\frac{N}{e\tau} - \frac{C}{E\left[F_S\right]}\left(1-p_J\right)\right]E\left[B_1\right]. \end{aligned} \tag{2.17}$$

Therefore, the value of n^* scales linearly not only with C, but also with M, N, p_J and τ^{-1}.

The value of n'' has to be computed numerically. From (2.5) computed for $p_B \simeq 0$ and with the product $\gamma e^{-\frac{\gamma\tau}{N}}$ given by (2.14), it results that:

$$n'' \simeq n' + \left[\gamma'' - \frac{C}{E\left[F_S\right]}\left(1-p_J\right)\right] \cdot \left(\frac{\tau}{2} + \frac{1-p_a}{p_a}E[B_a] + E\left[B_0\right]\right), \tag{2.18}$$

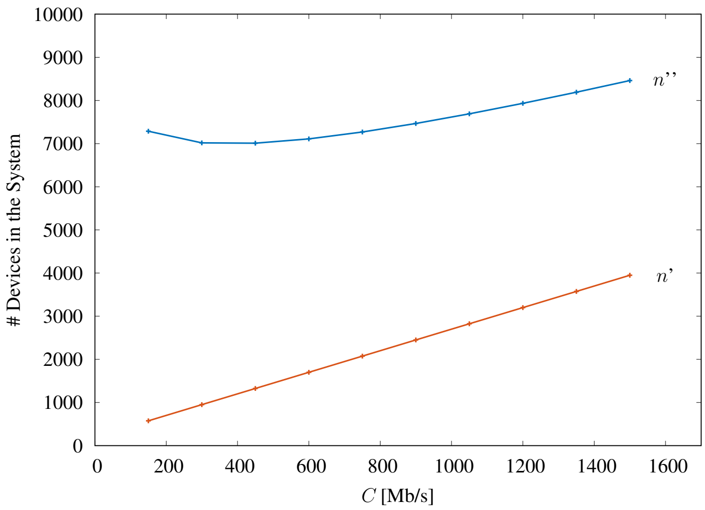

Figure 2.4: Behavior of n' and n'' versus the cell capacity (computed without ACB).

where γ'' is the largest root of (2.14). One can notice that while n' increases linearly with C and $1-p_J$, the distance $n''-n'$ decreases logarithmically with the same parameters, due to (2.14). Therefore, n'' decreases with C and $1-p_J$ for small values of such parameters, where the log decrease is superlinear, and then increases when the logarithm becomes sublinear. Moreover, n'' grows with N, because N increases the L.H.S. of (2.14), and therefore it has the effect of spacing apart the zeros of that equation, while n' is not affected by N, as noticed before with (2.16).

The behavior of n' and n'' vs. C is shown in Fig. 2.4 for the same example discussed in Fig. 2.3 (in there the cell capacity was fixed to $C = 150$ Mb/s, while now we consider more values). The behavior of n' is clearly linear, whereas n'' initially decreases and afterwards grows. Overall, n'' exhibits a quadratic behavior. In all cases, the distance $n'' - n'$ diminishes with C.

Asymptotic behaviour of γ vs. n. For $n > n''$, the Network throughput starts to vanish exponentially, and the result is that the device population progressively moves to the Random Access block. Asymptotically, there are no devices either in service or in think time, so that all devices loop between the Random Access subsystem and its backoff subsystem:

$$\lim_{n\to\infty} \frac{n}{\gamma} = \frac{\tau}{2} + E[B_0]. \tag{2.19}$$

Fig. 2.5 shows the behavior of γ vs. n for the example of Fig. 2.3. In the figure, γ grows very slowly for $n < n'$ (access requests do not collide and there is practically no blocking at the Network). Between n' and n'', the value of γ grows faster and faster, especially in the zone in which the RACH throughput has a negative slope (this is due to high collision probability and high Network

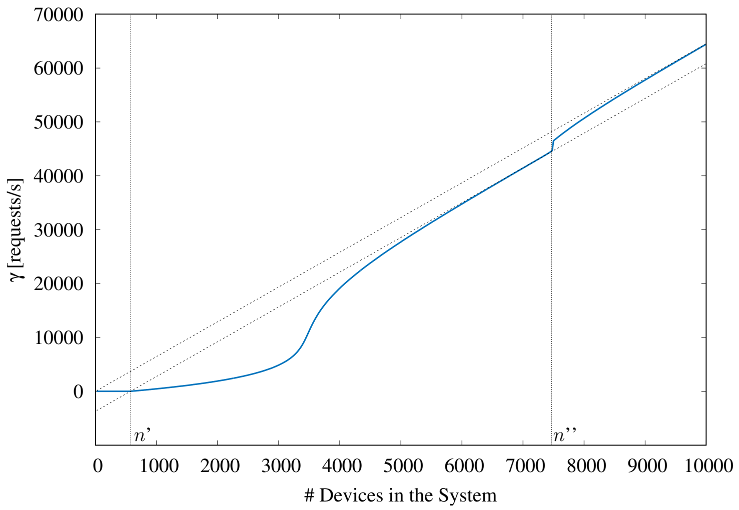

Figure 2.5: Monotonic relation between γ and n (computed without ACB).

blocking). However, as soon as p_B decreases again, due to excessive collisions, the curve of γ vs. n changes towards a linear relation (right before n'', where $p_B \simeq 0$). After n'' all devices move to the Random Access and backoff B_0 subsystems (no device will be under service, asymptotically), and eventually the relation between γ and n approximates (2.19).

With clusters. Clustering k devices results in transferring $kE\left[F_S\right]$ bits per network access, hence the cluster service time $E[S]$ becomes k times longer. So, n' decreases with increasing cluster size. However, the number of devices within clusters becomes kn'. Denoting by $E[S|1]$ the service time without clusters, we have:

$$kn' \simeq \frac{C}{E\left[F_S\right]}\left[kE[S|1]+E\left[T_{TH}\right]+(1-p_J)\left(\frac{\tau}{2}+\frac{1-p_a}{p_a}E[B_a]\right)\right], \tag{2.20}$$

which includes $(k-1)\frac{CE[S|1]}{E[F_S]}$ more devices w.r.t. the case without clusters. Similarly, we can observe that kn^* grows by M plus a number of devices proportional to N for each increase of 1 in the cluster size k.

The interval $n''-n'$ increases with the cluster size, because a factor k appears in the denominator of the r.h.s. of (2.14) when clusters are used. Therefore, the increase of the size of the network saturation region, in terms of devices, becomes $k(n''-n')$, which is more than a k-fold increase.

We can conclude that the beneficial impact of clustering is larger than the one obtained by increasing cell capacity, which is linear, and it comes at a much lower deployment cost.

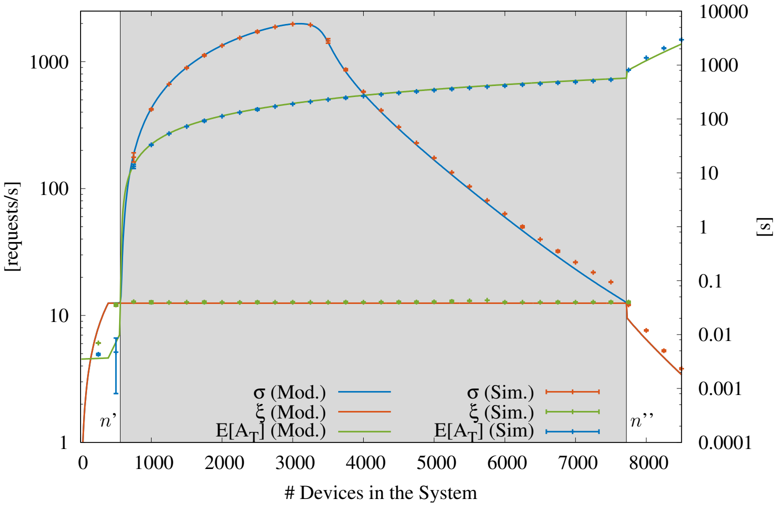

Figure 2.6: Model validation for a cell with 150 Mb/s capacity and $p_a = 1.0$. Left scale for σ and ξ, right scale for access delay.

Delay

The access delay $E\left[A_T\right]$ is negligible when the Random Access saturates first, and for $n < n'$ when the Network subsystem also saturates, unless ACB introduces high delay by using low values for p_a and/or high values for $E[B_a]$. When the Network subsystem is saturated, we know from (2.8) that $E\left[A_T\right]$ is proportional to n with coefficient $\frac{1}{\lambda} = \frac{E[F_S]}{C}$. For $n > n''$, the delay explodes exponentially. Therefore, the desirable range of population sizes goes from n' to $n' + \Delta n$, where Δn is such that the delay $\frac{\Delta n E[F_S]}{C}$ is bearable by the applications running at the devices in the network.

So, in practice, the study of n' and its approximation are key to tune system parameters properly during network design.

Validation through packet-level simulation

In order to validate the simplifying assumptions that we had to introduce for the analytical tractability of the model, we developed a packet-level simulator that reproduces the behaviour of the closed model in Fig. 2.1. However, in the simulator we used uniformly distributed (rather than exponentially) file sizes; in addition, the output of the Random Access subsystem is not a Poisson process, rather an impulsive process in which all successful RACH attempts are brought at the Network subsystem ingress at the same time. Of course, in the simulator, the assumption that all arrival processes are Poisson, homogeneous and independent does not hold.

Fig. 2.6 reports an example of the simulated results for σ and ξ, together with the analytical results. Specifically, we report numerical results for a cell

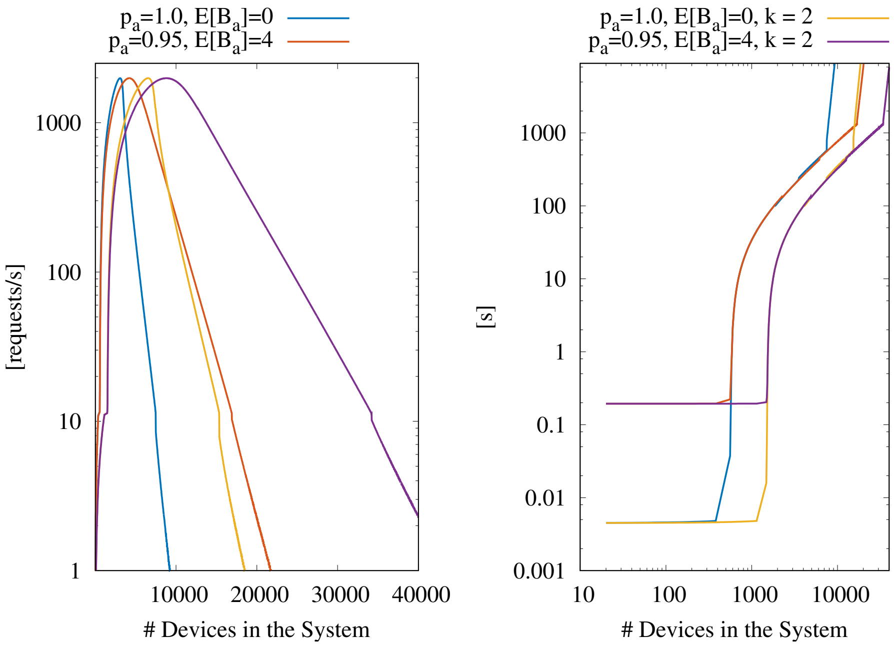

Figure 2.7: Throughput (left) and access delay (right): impact of ACB with $[p_a = 0.95, E[B_a]{=}4]$ (solid lines) and clustering (where k is specified) in the stadium scenario.

with $C = 150$Mb/s, $R = 10$Mb/s, $N = 54$, $M = 200$, and $\tau = 0.01$s, which are typical values for LTE eNBs. Moreover, we used $E\left[T_{TH}\right] = 30$s, $E\left[B_0\right] = 0.15$s, $E\left[B_1\right] = 1$s, $E\left[F_S\right] = 1.5$MB, to account for typical upload of pictures and small videos during crowded events by using applications like WhatsApp, with automatic file upload retry. We run each experiment a sufficient number of times to obtain small 95% confidence intervals. The figure clearly shows that the model is extremely accurate. We tested a wide range of values for all relevant parameters, and found very similar model accuracy in all cases.

2.6 Stadium: Numerical Results

We consider a stadium covered by a set of LTE cells. The system parameters are as reported in Table 2.1 (right-most column). Fig. 2.7 illustrates the impact of ACB and clustering in the specified scenario. We only report the results obtained with the ACB configuration that causes less delay, and one example of clustering ($k{=}2$). The figure shows that either ACB or clustering makes it possible to significantly increase the number of users in the system. In particular, clustering as few as groups of 2 users is very effective in increasing n'. ACB suffers large delays, so as to make it quite undesirable even for limited user population sizes. However, Fig. 2.7 also shows that ACB and clustering

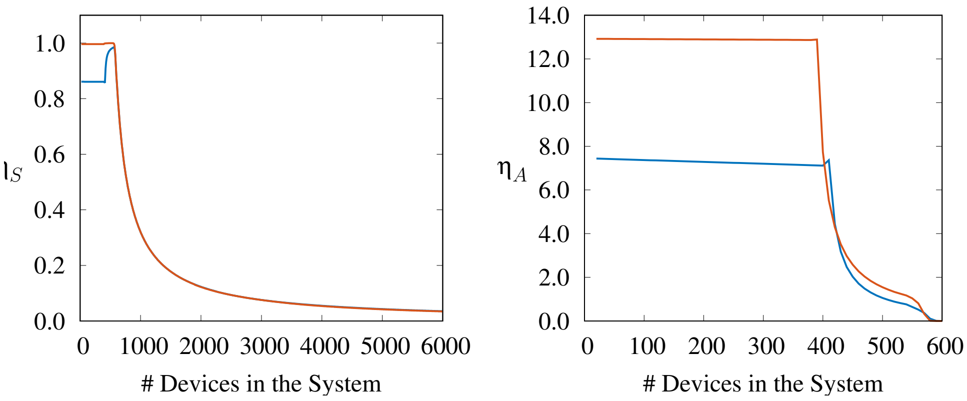

Figure 2.8: End-user QoE indicators η_S and η_A. Quality degrades with ACB $[p_a = 0.95, E[B_a] = 4]$ (solid lines) w.r.t. scenario without ACB (dashed lines), because of the additional delay it causes.

in combination achieve low delay and guarantee access to very large user populations.

Fig. 2.8 reports the values for the two QoE indexes η_S and η_A we defined in Section 2.4. Both indexes try to capture the user satisfaction, combining the service time and the access delay. In the first case we just compute the ratio between the service time and the sum access delay plus service time. In the second case we define a more elaborate parameter, which is inversely proportional to the service time, normalized to the service time when M users are under service, and exponentially fades with the access delay normalized to the access delay value at population value equal to n', thus being very sensitive to relative increases of delay rather than to absolute increases.

The curves of the QoE parameters show qualitatively similar trends. As regards η_S, with a low population of UEs, the network access time $E[T_A]$ is very low and mostly depends on RACH transit and ACB operation. Each device in service is guaranteed a rate equal to R, keeping η_S close to 1, unless ACB is used and $E[A_T]$ cannot be neglected. When the eNB can no longer provide the maximum rate R to each one of the n_S devices in service, $E[S]$ starts to increase, while $E[A_T]$ is practically constant (without ACB) or slowly increasing (with ACB), so that its weigh in η_S diminishes as the population increases. However, when the number of devices reaches the value n', the access delay $E[A_T]$ starts increasing fast (and linearly) causing a hyperbolic decrease of η_S towards zero. The QoE parameter starts dropping around 550 devices in the cell. In general, the figure shows that using ACB *is* detrimental in terms of quality experience in steady state conditions, especially with small populations, when the ACB delay is the most prominent component of the access delay.

For what concerns η_A, the figure shows that, without ACB, it starts from

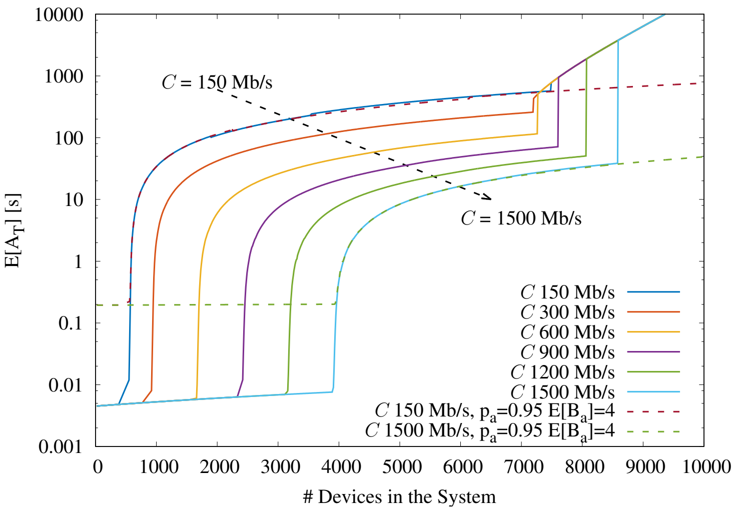

Figure 2.9: Access Delay for variable cell capacity with ACB $[p_a = 0.95, E[B_a]\!=\!4]$ (solid lines) and without (dashed lines).

the value $10/0.75 = 13.33$. This is the ratio between the data rate cap for each individual device, and the data rate given by the eNB to each user once the maximum number of users (200) is reached (150 Mb/s divided by 200 users means 0.75 Mb/s per user). With ACB, the additional delay due to barring decreases the initial value of η_A. In all cases, the curve stays close to the initial value as long as the access delay remains negligible, then it rapidly drops. Also in this case, the QoE parameter starts dropping around 500 devices. Note that this means that a coverage of the $50,000$ users in the stadium with good QoE would require about 100 cells, if each user carries just one device, 200 cells if each user carries two devices, and so on.

Of course, one possibility to improve performance is to use cells with higher capacity. In Fig. 2.9 we plot curves of $E[A_T]$ for cell capacities in the range 150-$1,500$ Mb/s. The critical element for QoE is given by the points where the access delay starts increasing significantly. This means about 550 devices with capacity 150 Mb/s and about $4,000$ devices with capacity $1,500$ Mb/s. The latter translates into 12 cells for $50,000$ devices, 25 in the case each spectator carries 2 devices.

In addition, Fig. 2.9 clearly shows that the operating area where both end-users and network operators wish "to be" is just before the curve's first knee. In such neighbourhood, ACB does not play any significant role, and $E[A_T]$ is a fraction of $E[S]$, before starting to rapidly move to bigger values. It is important to recall that this "change of phase" in the access delay is pinpointed by n'. The second knee of the curves corresponds to n'', and both knees change with the cell capacity. It is very important to notice that in the whole interval $[n', n'']$ the system bottleneck is the Network due to the

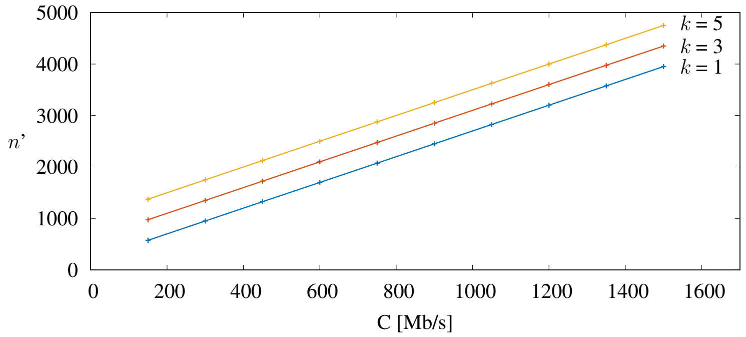

Figure 2.10: Values of n' versus the cell capacity, for variable cluster sizes. ACB curves are practically superposed to curves without ACB.

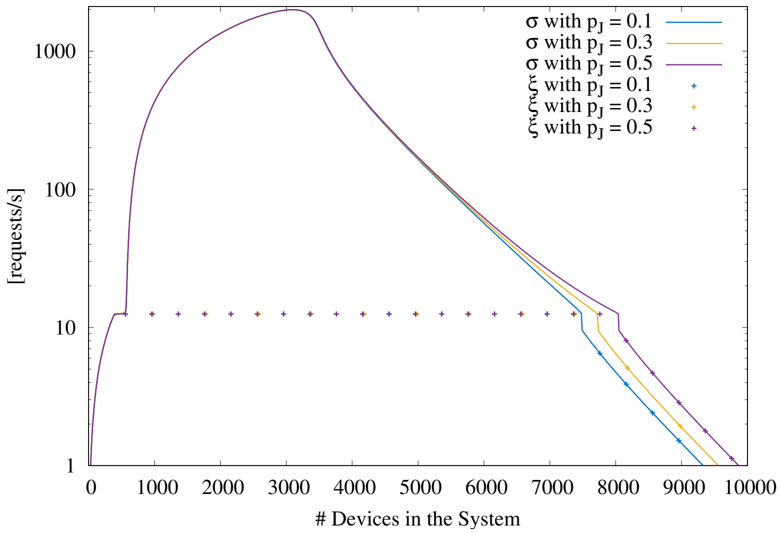

Figure 2.11: Impact of skipping the contention-based RACH procedure.

limitation of M `RRC_CONNECTED` devices. When the number of devices in the cell becomes larger than n'', we see a switch in the bottlenecks, and only from this point on the RACH subsystem becomes unstable and the access time explodes going asymptotically to infinite.

Increasing the cell capacity or the number of cells is quite costly, and may not be the most desirable solution to achieve good QoE in crowded environments. A much simpler option can be to allow users to coalesce in their network access attempts through the formation of clusters. Fig. 2.10 shows the values of n' as a function of the cell capacity, for variable cluster sizes. We immediately appreciate the advantages of clusters: the adoption of coalitions brings a gain comparable to the one obtained increasing C with a negligible cost (if any) to the network provider. Indeed, a gain equal to or larger than that obtained by doubling the cell capacity can be achieved by adopting a cluster size $k = 3$.

Finally, to evaluate the importance of reducing the load of the Random Access in presence of downlink traffic, we repeat our tests with different values of p_J. Skipping the contention-based RACH procedure introduces a small improvement in terms of the height of the point at n^*, allowing the RACH to sustain a slightly higher arrival frequency. However, as can be seen in Fig. 2.11 for the case with no ACB, the main impact of p_J on the performance of the system is reflected in the value of n'', which is moved towards larger values of n. It must be noted that the increase in height at n^* is so small not to be visible on the graphs, and that the increase in the value of n'' is not relevant from the point of view of applications, because at those numbers of users per cell, performance (e.g., in terms of access delay) is intolerably bad.

2.7 Practical Validity of the Model and its Limitations

The analysis presented in this chapter accounts for the main parameters of a 3GPP-compliant wireless access network. However, there are some parameters in the configuration of the network and in the application generating data to transmit that are not accounted for in the model and that are object of research and technical proposals. Those parameters can impose practical limitations to the behavior of the cellular access system. In particular, there are three main parameters that might be relevant and are not considered in the simple model presented and discussed so far: (*i*) the maximum number or RACH retries that a request can go through before being dropped; (*ii*) the timeout used in `RRC_CONNECTED` state, which guarantees that resources allocated to a device remain available beyond the last packet is transmitted, and which is useful to avoid incurring in a new RACH request procedure for customers returning within a few seconds; and (*iii*) the *patience* of a user, i.e., the maximum delay allowed by the application running at the user's device after which a traffic transmission request is dropped and a new request is issued after some application-specific backoff.

In the following we analyze and discuss the impact of each of such parameters by means of simulations, which are compared to the results of our model. The configurations parameters used in what follows are summarized in Table 2.2. As we will see, our model captures the behavior of the system in a wide spectrum of configurations and, most important, it always gives accurate results for the regions of operation (i.e., the ranges of the number of users) that have practical importance.

Impact of the maximum number of retries on the RACH

Let's denote the maximum number of RACH retries as $k_{\max}$. This is the maximum number of consecutive failed RACH attempts, after which an ac-

Table 2.2: Configuration parameters used to evaluate the practicality of our simple model

Quantity	Value
p_a	0
p_j	0
RRC_{TO}	$\{1000, 2000, 3000, 5000, 10000\}$ $[ms]$
Global timeout	$\{15000, 30000, 60000\}$ $[ms]$
Max attempts	$\{10, 20, 30\}$

cess request is dropped. If such drop occurs, the device that was not granted network access will retry after some time, according to a backoff mechanism that has a length comparable with the think time of our model (tens of seconds). We therefore simulate a network in which, after a request fails RACH access for $k_{\max}$ consecutive times, the devices goes back to the Think station of Fig. 2.1 without sending any data. Therefore, in this modified system, differently from the originally described one, it is possible to have *failures*. Note also that, for $k_{\max} \to \infty$, the system tends to the one we have modeled.

Fig. 2.12 shows that the impact of $k_{\max}$ on the throughput of the RACH (σ) and on the throughput of the Network station (ξ) is important only for population sizes $n > n^*$. Specifically, the smaller $k_{\max}$, the larger the distance $n'' - n'$ because n'' increases while n' does not practically change. Similarly, n^* remains practically unchanged. This can be explained by considering that the probability to fail $k_{\max}$ consecutive times on the RACH is an increasing function of the load γ, and for populations below n^* devices, the probability to fail even a few as four or five times in a row is low because the probability of a single RACH failure is of the order of $1 - 1/e$ or smaller. The figure also shows that the network throughput does not change before n''.

Let's now consider the impact of $k_{\max}$ on the access delay. As depicted in Fig. 2.13, $E[A_T]$ is impacted only starting from the value n^* identified with our model. For larger values of the population of devices, the access delay decreases, still it remains quite high. This delay reduction is however obtained at the expenses of the success probability experienced by a network access request. Indeed, as shown in Fig. 2.14, the system suffers high failure probability starting from $n = n^*$, which is the point at which the simple model shows a pick in the RACH throughput.

By jointly considering access delay and failure probability, it is clear that the network cannot be efficiently and satisfactorily operated with populations much larger than n'. As a consequence, our model offers accurate predictions well beyond the range of interest, since it is accurate up to $n^* > n'$.

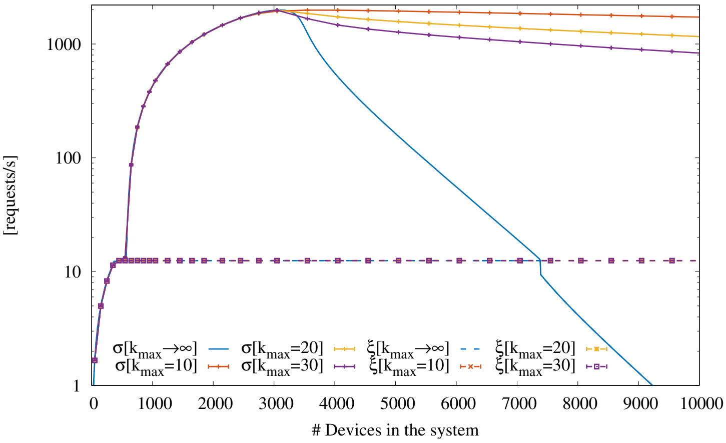

Figure 2.12: Effect of the max number of RACH retries $k_{\max}$ on σ and ξ

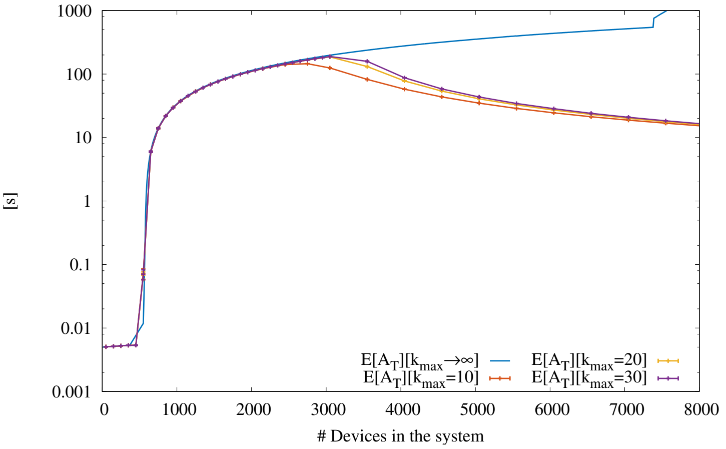

Figure 2.13: Effect of the max number of RACH retries $k_{\max}$ on the access time

Impact of timeout for the RRC_CONNECTED state

We now consider that, in real 3GPP access networks, devices that enter the RRC_CONNECTED state will leave that state based on an activity timeout, i.e., only after a time RRC_{TO} has elapsed, during which no transmission occurred. Otherwise, if after a file transmission is complete a device wants to initiate a new file transmission and the timeout has not expired, that device will not need to go through the RACH again. We simulate such system by (i) counting all devices in RRC_CONNECTED state, including the ones that have transmitted their file, in the set of devices that are under service, and whose number cannot exceed M; (ii) if a device exits the Think station and it is still in RRC_CONNECTED state, it will jump to the Network station of our system of Fig. 2.1. Note that our simple model corresponds to the case in which

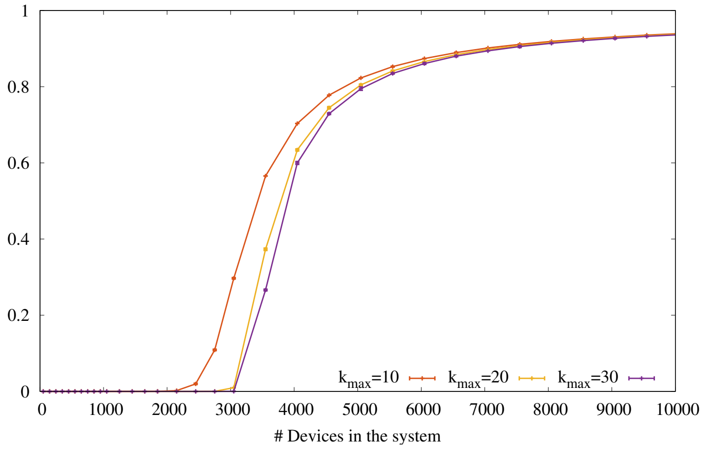

Figure 2.14: Failure probability introduced by the max number of RACH retries.

$RRC_{TO} = 0$. Note also that, differently from the case of $k_{\max}$ discussed before, the use of a timeout for the `RRC_CONNECTED` state does not lead to failures, although it incurs a reduced average number of devices using the transmission channel in at the same time, as commented in what follows.

In this case, the network throughput ξ is practically not affected, as shown in Fig. 2.15, except n'' shifts forward due to returning devices sustaining ξ even when σ fades off. The Network in our system is PS, which means that when less than M devices are under service, they still use all resources, i.e., they are served faster than with exactly M devices. Therefore, although the use of a timeout $RRC_{TO} > 0$ imposes that some devices in the Think station do not free their Network allocation before RRC_{TO} time units after their file is transmitted, Network resources are not wasted. Fig. 2.15 also shows that the throughput of the RACH with $RRC_{TO} > 0$ is barely affected by the timeout. However, the longer the timeout, the higher and the sooner σ grows before n', and the later it falls after n''. This is due to the fact that if devices return to the Network station before the timeout expires, the load of the RACH will be alleviated, with less collisions experienced. However, in the interval between n' and n'', σ is practically simply shifted backward because when Network is saturated the RACH sees a system with $M(1-r)$ serving slots instead of M, and our model can be applied to compute the resulting σ.

For what concerns access delay, Fig. 2.16 shows that only minor differences with our model can be appreciated. Specifically, a part for a backward shift of n' and a forward shift of n'' that we have commented above, the linear slope of $E[A_T]$ in between is only slightly changed because the value of λ to be used in this case in (2.8) is $\xi(1-r)$. This effect is due to the increased probability to fail Network access (where rM allocation slots are now reserved for returning customers) and the consequent return to the RACH of a fraction

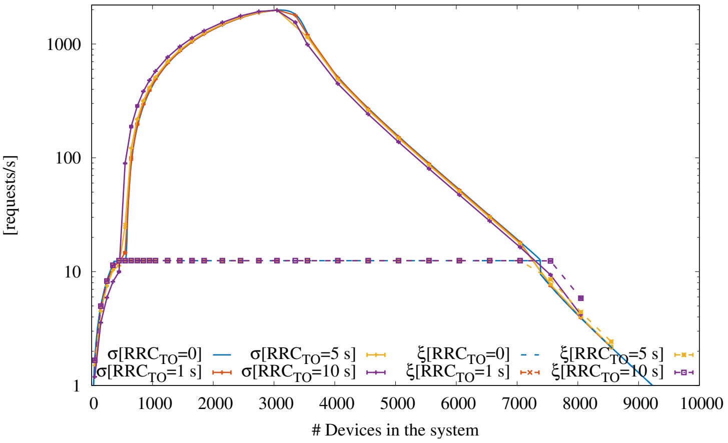

Figure 2.15: Effect of several configuration of RRC_{TO} on σ and ξ

of requests that grows with the value of RRC_{TO}. This is partially compensated by the short access delay experienced by devices for which the timeout does not expire. In the extreme case in which $RRC_{TO} \to \infty$, the access delay is minimized for returning customers (it reduces to the latency of the Network station), although it becomes infinite for the rest of users, so that, as soon as $n > M$, the average access delay diverges.

Fig. 2.17 shows the average number of devices under service as a function of RRC_{TO}. The ratio between M and the value plotted in the figure is the coefficient r discussed above, which grows with the timeout and causes increased access delays.

In practical circumstances in which RRC_{TO} is of the same order of magnitude as the Think time, or shorter, the resulting values for the probability r are limited to a few percents, and our model provides a good approximation for all values of the population n.

Impact of the application timeout

Finally, we consider the impact of the application timeout. We simulate the system of Fig. 2.1, although we interrupt and drop a service request if its access delay exceeds the application timeout AR_{TO}. Our model corresponds to the case $AR_{TO} \to \infty$. When a request is dropped, a failure occurs and the device goes back to the Think station.

Fig. 2.18 compares the throughput of our model with the one of the simulator using various values of AR_{TO}. The figure unveils that no differences can be appreciated in σ and ξ for population size up to n' and slightly above that point. Beyond that point, the curves of σ separate. Both n^* and n'' increase with AR_{TO} decreasing. To explain this behavior, consider that the average access delay in our model increases with the population size. So, as soon as

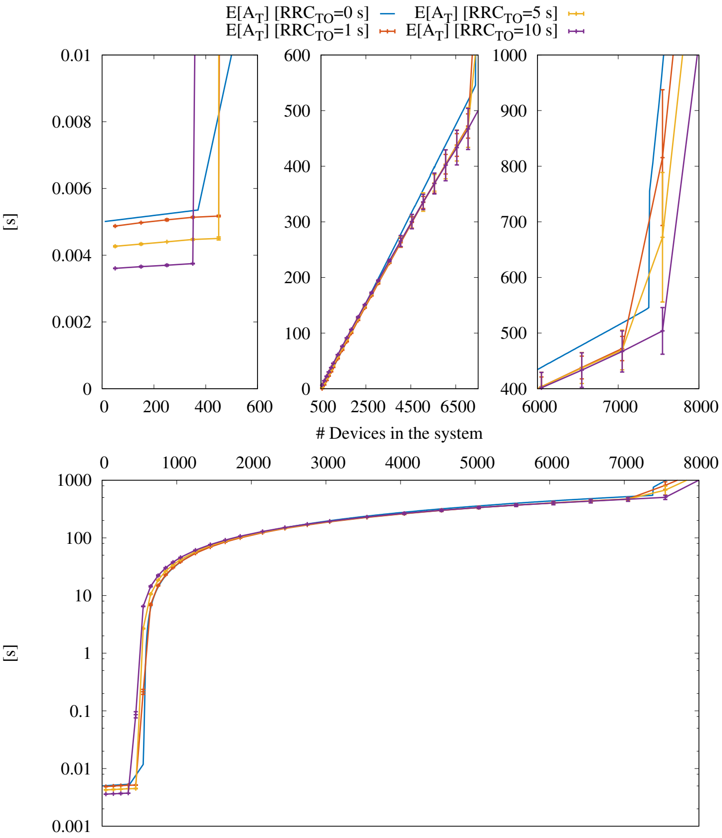

Figure 2.16: Effect of several configuration of RRC_{TO} on the access time. The top row plots represent in a linear scale and from left to right $E[A_T]$ for population in the ranges $[0, n'^{+}]$, $[n'^{-}, n''^{+}]$ and $[n''^{-}, 8000]$ respectively. In the bottom plot, in logarithmic scale.

the average access delay increases, the probability to trigger an application timeout increases. With AR_{TO} larger than the RACH latency (in the order of τ, which is extremely small for an application timeout), a non-negligible timeout probability is possible only when the RACH experiences significant collisions, i.e., starting with some value of n in between n' and n^*, which is what we observe in the figure. From that point on, requests that suffer timeouts take a Think time backoff, which is longer than the RACH or the Network backoff, thus resulting in reduced RACH load. This is why n^* and n'' move towards higher values.

For what concerns the impact on access delay itself, Fig. 2.19 shows significant differences from the point at which timeouts start occurring. Interestingly, the access delay with AR_{TO} grows very slowly after the curves in the figure split, and this is due to the hard bound imposed by the timeout.

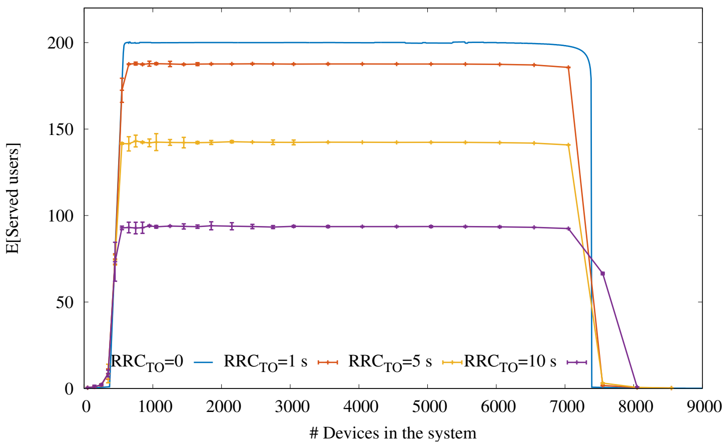

Figure 2.17: Number of users in service due to the use of a finite and not null timeout for the RRC_CONNECTED state.

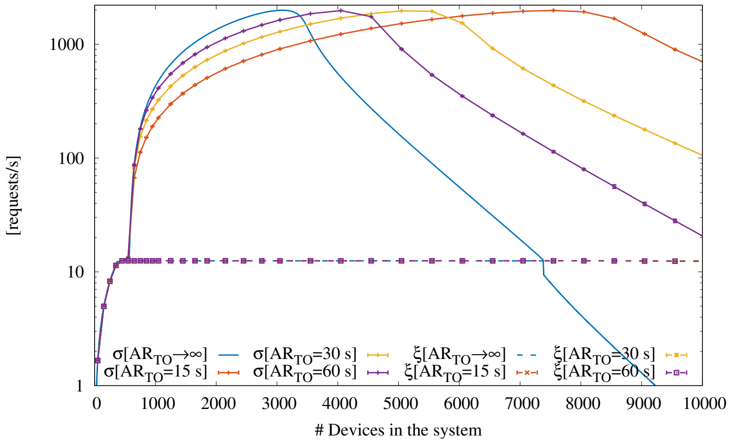

Figure 2.18: Effect of several configuration of the application timeout on σ and ξ

Therefore, using an application timeout could allow to use the system well beyond n' and even beyond n^*. However, as shown in Fig. 2.20, the failure probability becomes relevant well before n^*. So we conclude that using the system with populations much higher than n' is not a good idea even in this case.

In conclusion, using our model is very accurate for the range of population sizes in which the failure probability is negligible or bearable (below a few percents).

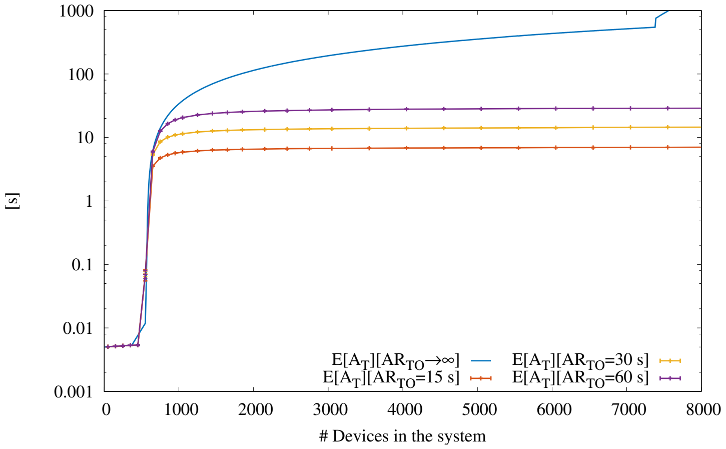

Figure 2.19: Effect of several configuration of application timeout on the access time

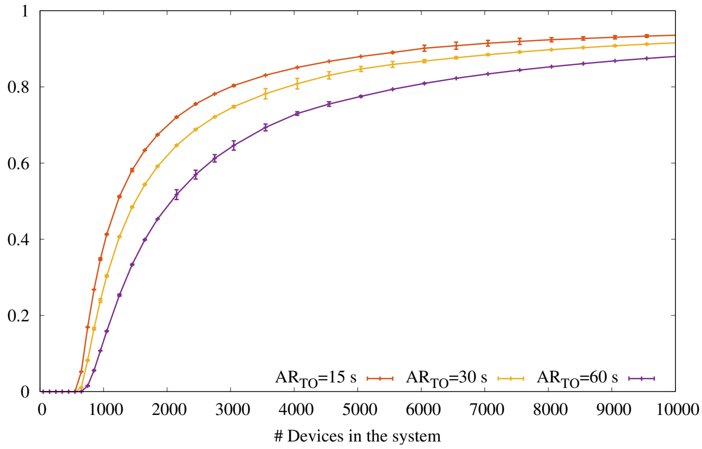

Figure 2.20: Failure probability introduced by the application timeout.

2.8 Related Work

In [46], 3GPP has identified the random access mechanism as a possible problem when the number of connected devices rises to tens of thousands. For this reason, MAC overload control has been investigated, and a broad literature exists on this topic. See [47] for a comprehensive overview. Simple models to estimate the probability of preamble collision in the PRACH channel are presented in a few 3GPP standard documents (e.g., [46]), and in the literature (e.g. [45], [47], [48], [1]). The conclusions of most of these studies point out that for Machine-Type Communications (MTC) applications, the PRACH procedure can drastically limit network performance. Possible approaches to modify the PRACH access procedure have been proposed in [49, 50].

Most of the previous studies on dense cellular environments have focused on MTC scenarios, and [47] shows that the differences between the human-based and the MTC scenarios are substantial. Nevertheless, the PRACH access mechanism, and its interactions with the other phases of the network usage cycle play an important role also in case of human-based scenarios. This was shown in [39], through a measurement-based study of cellular network performance during crowded events, showing that network access failures become orders of magnitude higher than those observed on routine days, and the interaction between access and transmission phases generates behaviors difficult to predict. The simple analytical model presented in this eNB provides a tool to understand the root causes of the behaviors measured in [39], and to quantify the impact of the crowd size on network performance, also indicating possible approaches to correctly dimension the network and to mitigate the negative impacts of crowds.

2.9 Conclusions

This chapter presents a model to capture the key aspects of the behaviour of a cellular networks in crowded environments. The main merit of the model lies in the insight that it brings on cellular system operations in very crowded environments, and in the possibility to use it to drive the correct dimensioning of the cellular system in very crowded environments. As an example, the model allows the assessment of the benefits achievable through the adoption of D2D communications to reduce the congestion on the RACH more effectively than with ACB, thus significantly improving performance and QoE. For example, our model shows that, instead of serving 50,000 terminals with 100 cells of capacity 150 Mb/s each, it is possible to use 25 cells, each of capacity 300 Mb/s, provided that clusters of 5 devices are formed to access the RACH. Nonetheless, in this chapter a coordination mechanism to set up D2D coalitions is not discussed and an example of distributed coalescing mechanism is presented in chapter 3.

Chapter 3

Device to Device & users' incentivisation

3.1 Introduction

Chapter 2 shows D2D is an effective approach to alleviate congestion in RAN by coalescing UEs' access requests. For example, according to the taxonomy in [51] *out-band controlled D2D communications* allow UEs to transmit over the unlicensed spectrum—e.g., WiFi—under the guidance of a central entity. However, it does not provide a practical approach to organise UEs into an appropriate structure that enables them to communicate directly.

Moreover, many previous works [51–53] also show that cooperation leads to an optimal outcome by offloading content transmissions over D2D links. However, in these works it is not taken into account that cooperation arises some issues on the user side. Indeed, relaying other UEs' transmissions is as an additional cost for the user because she/he sees her/his battery operated terminal draining more power to support a cluster of connected devices. So, what if any client is considered as a selfish and rational player? Is cooperation still the best choice?

It is clear that the cooperation convenience depend on which side of the D2D link the UE is: if it is the one who contributes with its own short range bandwidth to gather traffic and relays it towards the mobile network, then the cooperation is always unfavourable unless it is remunerated. Therefore, in this chapter we provide an approach to study UEs willingness to cooperate and a suitable mechanism to foster the D2D communications.

To this aim, in this chapter we consider a set of UEs under within the coverage area of an eNB and we assume that the users have subscribed a pay-per-view service to receive a common content, e.g., live streaming of a popular event. In this scenario, we define proper mechanisms to incentive collaboration by leverage on the subscription costs. Several other papers have shown that D2D strategies can be used to relieve the base station task. This approach clearly amounts to shift some of the efforts to UEs that are charged by an extra task (the relay service), hence increasing their energy consumption; such issues related to the UE point of view have attracted limited attention in the literature. The D2D-enhanced offloading scheme proposed in this chapter tries to fill this gap.

In particular, the proposed analysis is centred on the UEs, on the incentive that the eNB must offer them to participate to the content distribution, and on the relations between these issues and the user battery state. The main contributions of the chapter are: *i)* the proposal of an incentive mechanism encouraging terminals to organize into an *optimal* number of clusters from the point of view of both bandwidth capacity and power consumption; *ii)* the consideration in the overall balance of hidden terminal costs, e.g., the battery power drained for altruistic collaboration.

We validate the model representing our incentive mechanism against detailed simulation in both static and dynamic scenarios. We then exploit the model analysis to quantify the cost reduction of the BS thanks to energy savings and the increased spare bandwidth that can be allocated to other services.

3.2 Problem Formulation

System Description

As cellular networks became the prime medium for users to access multimedia and live contents online, network providers also become content providers. Indeed, recently it is quite common for users to subscribe special offers or contract with their network providers that comprise of discounted access to streaming services, either owned or in partnership. Furthermore, 5GPPP pose the Multimedia & Entertainment (M&E) is one of the five vertical sector—the most promising and challenging application scenarios for the 5G technology.

In [16] the M&E scenario has been detailed and KPIs defined accordingly. Although many KPIs are novel, e.g., related to the emerging trend of the users as content providers sources, the need for higher network capacity is common to requirements of previous cellular generations. Despite huge efforts are being spent to cope with the always increasing demand of bandwidth—e.g., millimetres waves—a solution easily employable and economically sustainable does not yet exists.

Therefore, Device to Device (D2D) communication has attracted great interest as a promising solution to enhance the standard cellular infrastructure, e.g., LTE and 5G networks, under several aspects such as improved spectral efficiency, larger capacity and lower energy consumption [54–57]. In particular, it has been recently shown that D2D can significantly boost the performance of group oriented services [52, 53], e.g., bandwidth demanding multicast or broadcast of multimedia.

In this chapter we consider N UEs in the range of an eNB that subscribed a pay-per-view service to receive a common content that is being transmitted by the network provider, such as a live streaming of a popular event, a football match ot the last episode of a trilling TV series. UEs are equipped with multiple radio interfaces and, among the others, with a *long range* link interface—employed for wireless transmissions over licensed frequencies operated by service provider, i.e., LTE or 5G—, and a *short range* link interface over the unlicensed spectrum, such as Wi-Fi or Bluetooth. As pointed out in Chapter 2 and in many other works [43, 57], such unlicensed and uncontrolled frequencies might be exploited to alleviate the congestion in LTE and 5G systems by increasing the frequency diversity. Out-band Device to Device (D2D), one of the suitable D2D configuration [51] which leverage on the multiple radio interfaces to enable direct communications among UE without the brokerage of the eNB, by means of the short range link interfaces i.e., Wi-Fi Direct [58].

In the absence of any strategy aiming at reducing energy consumption and/or bandwidth optimization, all N UEs would download the content over the cellular radio interface, each one requiring a transmission of the content they subscribed for. On the other hand, the UEs can cooperate and forward among each other, using the short range links, the content. As a result, the UEs can reduce the amount of data circulating over long range links targeting both a decrease in terms energy consumption and an increase in the capacity of the eNB.

In the following, we consider a generic cellular network uses such as LTE or 5G for the long range transmissions, while D2D communications employ the Wi-Fi Direct protocol to locally spread the multimedia content. Therefore, the transmissions are done in unicast on long range links whereas on the short range links multicast is used. We assume that wireless multicast optimization techniques (e.g., [59]) are deployed to increase the number of receivers.

The system under consideration is the one depicted in Figure 3.1, a single an eNB providing service to a population of UEs within its coverage area, and transmissions are normally operated according to the specific cellular technology—LTE or 5G. However, either to save energy or due to traffic congestion the eNB might require UEs to coalesce their transmissions and self-organize in *clusters*. A D2D cluster is a set of co-located UEs, within the coverage area of the short range interface A_{d2d}, among which one will be in charge of collect and dispatch transmission to and from all the other *cluster participants*.

Therefore, *cluster heads* will be the only UEs with an active connection toward the eNB arising, as a side effect, a remarkable reduction of the signalling overhead required to manage and keep cellular connection alive. Clearly, ac-

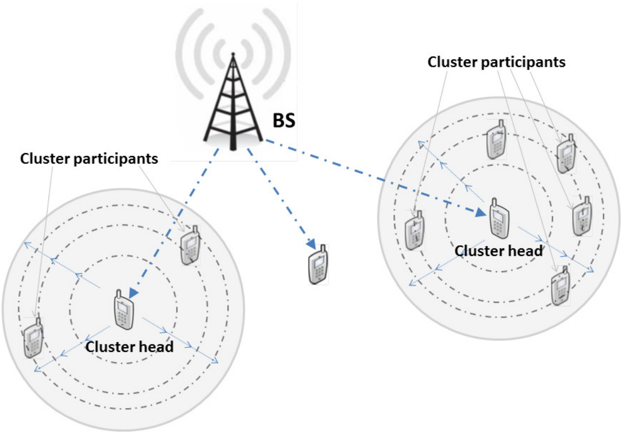

Figure 3.1: Illustrative example of D2D communication assisting cellular content distribution

cessing the network through some D2D relays, the cluster heads, is advantageous: it requires lower transmission power and the shorter distance to cover provides better SNR. On the other hand, relying cluster participants' transmissions is a demanding activity and it drains energy out of devices' battery. Therefore, coalescing is beneficial for both the eNB and cluster participants, but all the efforts fall back onto cluster heads, making this role unappealing.

Assuming UEs being rational—acting to pursue their own interests—, none of them will volunteer to take on the relying duty, requiring an incentive mechanism to boost the cooperation. In particular, such mechanism to be suitable for network providers, have to relay on as few information as possible to not incur in additional signalling overheads. At the same time, the incentive mechanisms have to encourage the UEs to organize themselves into an *appropriate* number of clusters, choosing the suitable incentive amount to maximise D2D benefit while reducing network provider operational costs.

In the following we assume that energy consumption is solely determined by transmissions. That is, we neglect the energy consumptions due to receptions both in the UE—this is not a cost introduced by the the cooperation—and in the eNB—whereas such traffic is negligible if compared to the stream of high-definition multimedia data. Therefore, we only consider transmission costs for the base station, denoting by $E_{tx,b2d}$ the power consumption rate.

Table 3.1: Notation used in Section 3.3 and 3.4

Quantity	Symbol
Number of devices	N
UEs' battery level	β
Threshold battery level	β_{min}
Minimum incentive required	c_0 [\$]
Cluster head role duration	τ [min]
Cluster head acceptance probability	p_{ch}
D2D uncovered probability	p_u
Average number of cluster head	n_{ch}
Average number of cluster participants	n_{cp}
Average number of D2D uncovered UEs	n_u
Cellular coverage area	$A_m 2d\ [Km^2]$
D2D coverage area	$A_{d2d}\ [Km^2]$

3.3 The D2D-enhanced Offloading Scheme

As reported before, we assume the short range transmissions being profitable with respect to the cellular ones from the point of view of the drained energy per transmitted data unit, as also [56] reports. Hence, in our analysis a rational UE would always prefer joining a nearby cluster and receive the content over the short range link, given that an UE has agreed to act as a cluster leader. Indeed, to set up a D2D mechanism requires to convince some UEs to devote a fraction of their own resources to relay neighbouring terminals transmissions'. As customers, UEs have subscribed a pay-per-view settlement and service providers can foster them helping out the system by providing discounts or direct payments. Therefore, the incentivisation process can be modelled as a bargaining in which, for each time window τ, the eNB offers a payment and the UEs can either accept or reject it.

Each terminal has its own minimum incentive that is willing to accept, and this minimum depends on the battery level. However, for a human user the worth of its mobile device battery is related to the expectations he/she has on the residual battery life. Indeed, with a completely charged device spending some energy has a risible importance, but when the battery has been almost drained, every single split of energy is worthy. Therefore, assuming UEs' battery level β being normalised on the interval $[0, 1]$, the monetary value of UEs battery can be modelled as a generic and invertible function $m(\beta)$ which *i)* when $\beta = 1$ it assumes the minimum incentive an UE is willing to accept to cooperate, c_0, and *ii)* it grows indefinitely when approaching a threshold value β_{min}.

Without loss of generality, it can be assumed that the distribution of battery levels in the population can be expressed as a continuous random variable X and that such distribution is known to the eNB. Moreover, we assume that, to establish an appropriate incentive, the only additional knowledge the eNB can relay on is the aforementioned function $m(\cdot)$. Therefore, the eNB needs to know the probability that an UE, whose battery level is unknown, accepts the incentive offered to become a cluster head. Therefore, we can define $Y \sim m(X)$, where Y is the r.v. representing the minimum incentive that an UE is willing to accept to become a cluster head. Unfortunately the domains of X and $m(\cdot)$, respectively $[0, 1]$ and $(\beta_{min}, 1]$ do not match, and the trivial solution cannot be applied. However, the conditional probability $Pr\{Y < y|X > \beta_{min}\}$ can be easily derived by substitution, obtaining

$$Pr\{Y < y|X > \beta_{min}\} = \frac{1 - F_X\left(m^{-1}(y)\right)}{1 - F_X(\beta_{min})}. \tag{3.1}$$

Hence, the probability an UE is willing to accept the offered incentive y, $p_a(y)$, is expressed as

$$p_{ch}(y) = Pr\{Y < y|X > \beta_{min}\}(1 - \beta_{min}), \tag{3.2}$$

where the latter term normalises the probability over the domain of X. Hence, for a given offer y UEs will react either accepting or rejecting it, meaning that the probability of having exactly l cluster head in a population of N UEs is binomially distributed

$$p_{ch}(y, l) = \binom{N}{l} p_{ch}(y)^l \cdot (1 - p_{ch}(y))^{N-l}, \tag{3.3}$$

while the average number of cluster heads is $n_{ch}(y) \approx Np_{ch}(y)$.

Let's imagine the eNB has enabled the offloading scheme, and let's assume that thanks to incentive y UEs are relying traffic. To understand how good the decision of choosing y as incentive amount the eNB can compute the number of uncovered UEs, that is the ones do not reside in the short link coverage area of any cluster head. Therefore, given the area covered by a cluster head (A_{d2d}), the whole area covered by the eNB (A_{m2d}), and assuming there are l cluster heads, we can approximate the probability that a generic UE does not fall in the area covered by the l cluster heads as

$$p_u(l) = \left(1 - \frac{A_{d2d}}{A_{m2d}}\right)^l. \tag{3.4}$$

Equation (3.4) has been derived under the hypothesis that the coordinates of the N UEs in the area covered by the BS are described by uniform independent random variables. Moreover, accuracy of the approximation described by (3.4) increases as the ratio $\frac{A_{d2d}}{A_{m2d}} \to 0$: in this case, the probability that a fraction of the A_{d2d} area falls outside the A_{m2d} area also approaches 0.

$$\text{minimize } \eta(y) = (n_r\,(y) + n_c\,(y))\,\xi + n_r\,(y)\,y \tag{3.7a}$$

$$\text{subject to}$$

$$n_r\,(y) + n_c\,(y) \leq N \tag{3.7b}$$

Figure 3.2: The optimization model for the service provider's costs minimisation.

Combining (3.3) and (3.4), the average number of UEs that are not covered by any cluster head is

$$\overline{n}_u(y) = \sum_{l=0}^{N} (N-l) \cdot p_{ch}(l,y) \cdot p_u(l). \tag{3.5}$$

Note that in our content distribution scheme the UEs that are not covered by any cluster head have to be served via long range links, introducing an unwanted cost for the eNB. Indeed, the total cost of the D2D offloading scheme can be written as follow

$$\eta(y) = (n_{ch}\,(y) + n_u\,(y))\,\xi + n_{hc}\,(y)\,y, \tag{3.6}$$

where the first term accounts for the transmission costs and the latter for the costs introduced by the incentives. Moreover, $\xi = \delta E_{tx,b2d} \cdot \lambda \cdot \tau$ accounts for the streaming cost per a time period of duration τ hours where λ is the transmission bit-rate and δ is the electricity cost.

Therefore, a service provider employing the proposed offloading method aims to find that optimal offer y^* which *i)* maximises the number of D2D content transmissions while *ii)* employing the minimum number of cluster heads—that is, minimise the costs due to long range transmissions and avoid redundant incentives to cover the same area. Doing so, requires to solve the minimization problem in Figure 3.2

3.4 Numerical Results

Uniform battery level distribution. In section 3.3 we derived a generic formulation of the offloading mechanism, making our approach easily configurable given a battery distribution X and a preference function $m\,(\cdot)$. At the best of our knowledge, neither one or the other have already been investigated and, for shake of simplicity, in the following we assume battery levels uniformly distributed, $X \sim U\,(0,1)$.

Here, we define β_{min} as the minimum power level that the battery must have to allow the UE of being cluster head for a relay period of τ hours. Basically, using the results on WiFi consumption presented in [60] we estimate

Parameter	Value	
A_{m2d}	0.1256	[Km^2]
A_{d2d}	0.0007065	[Km^2]
N	100	
$E_{tx,b2d}$	14.65	[Watt/Mbps]
λ	1	[Mbps]
τ	$1 \sim 16$	[minutes]
δ	0.1	[$/KWh]
c_0	0.1	[$]

Table 3.2: Experimental settings

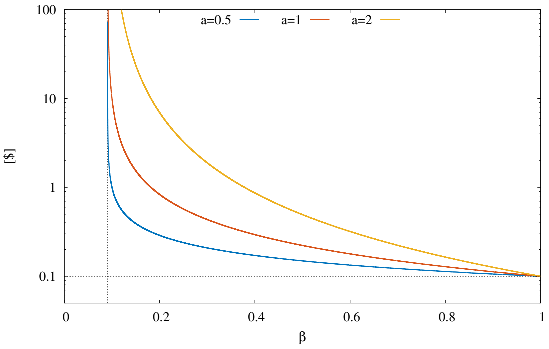

Figure 3.3: Plot of (3.8) for $a = 0.5, 1$, and 2.

reasonable values for β_{min}, assuming a typical battery capacity of 1400 mAh and using the Peukert's law to approximate the drained charge for the relaying period. Therefore, a possible implementation of the requirements for the function representing the minimum incentive accepted—that is, it assumes a small value c_0 when the UE's battery is fully charged and it increases as β tends to $\beta_{\min}$—is the following,

$$m(\beta) = \begin{cases} \dfrac{(1-\beta_{\min})^a}{(\beta-\beta_{\min})^a} c_0 & \text{for } \beta \in (\beta_{\min}, 1] \\ \text{undefined} & \text{otherwise,} \end{cases} \tag{3.8}$$

where a is a real number used to increase/decrease the concavity of the function. Figure 3.3 depicts the function defined in (3.8), with $x_{\min} = 0.0.091$, $co = 0.1$, and for three different values of a. Eventually, using $X \sim U(0,1)$ and (3.8) to derive an explicit formulation of the acceptance probability $p_{ch}(y)$

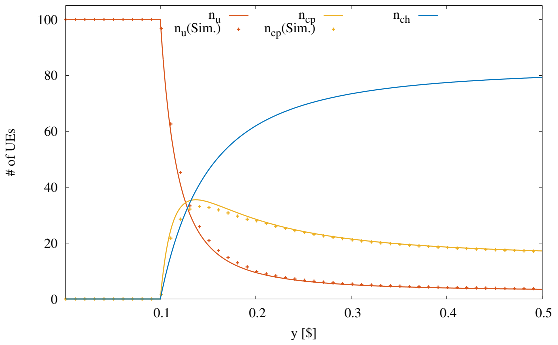

Figure 3.4: UEs roles as function of the offered incentive

(3.1), we have

$$p_a(y) = \begin{cases} 0 & \text{if } y < c_0 \\ \dfrac{1 - \sqrt[a]{\frac{(1-\beta_{\min})^a}{y}c_0} - \beta_{\min}}{1-\beta_{\min}} & \text{if } y \geq c_0. \end{cases} \tag{3.9}$$

Figure 3.4—note that here $n_{cp} = N - (n_{ch} + n_u)$ denotes the average number of cluster participants—describes the mode of operation of our offloading mechanism:

$\mathbf{y < c_0}$, the incentive is not enough to convince any of the UEs to cooperate and all the communications happen on long range links, i.e., $n_{ch} = 0$ and $n_u = N$;

$\mathbf{c_0 \leq y \leq y^*}$, when the incentive y grows above c_0 some UEs, the ones with higher battery level, begin to accept the payments and relay transmissions. At the same time, also the number of cluster participants grows up to $y^* \simeq 0.122$, the optimal incentive for this scenario;

$\mathbf{y^* < y}$, when the incentive the eNB pays for grows higher then y^*, the additional UEs accepting the deal are the ones that with a lower incentive would be served through D2D links, and indeed n_{cp} decreases for values of $y > 0.11$.

Therefore, our method allows to save cellular bandwidth by offloading part of the connections to unlicensed frequencies, and at the same time it reduces eNB consumptions. Indeed, Figure 3.5 shows the operational costs of an eNB implementing the offloading scheme. Obviously, since (3.6) is a linear combination of the values in Figure 3.4, the pattern is easily ascribable to the trend of the different UE roles. However, it is interesting to notice that, the optimal minimum operational cost of our offloading scheme leads to a

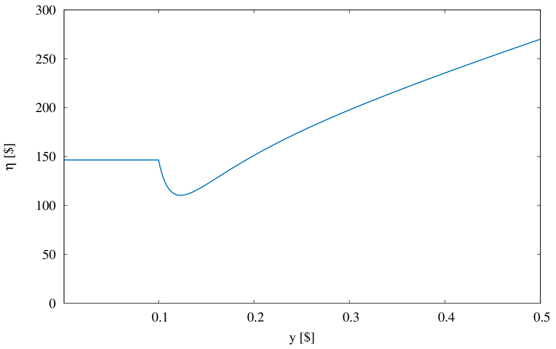

Figure 3.5: BS costs as function of b

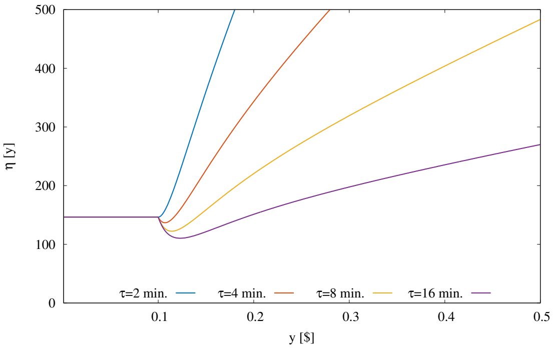

Figure 3.6: Effects of different durations of the relay periods

configuration where we have that $n_{ch}(b^*) \simeq 27$, $n_u(b^*) \simeq 39$, and $n_{cp}(b^*) \simeq 34$: it reduces the cellular transmissions, but to avoid wasting incentives almost two third of the UEs are not covered by D2D links.

Eventually, in Figure 3.6 several relaying periods of different duration τ are compared. To ease the comparison costs the eNB bears for each short period are summed up to compare them with the longer one. In Figure 3.6 periods last respectively from 1 to 16 minutes, and result shows that the longer periods present lower costs. Indeed, shorter relaying time windows require more frequent topology adjustments, each one requiring at least the payment of the minimum incentive of c_0 to all the relying UEs. On the other hand, we will see that refreshing the coalition topology makes the offloading scheme more resistant to users mobility.

Non-uniform battery distribution. To investigate the impact of the battery level, here we use a beta distribution with several parameter configurations, and in Figure 3.7 the respective p.d.f are depicted. In particular, we model three different scenarios:

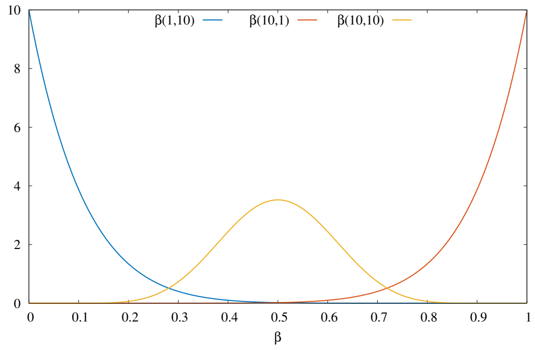

Figure 3.7: Effects of different battery distributions

$\beta\,(1,10)$ generates a distribution biased towards low battery levels;

$\beta\,(10,1)$ generates a distribution opposed to the previous one, biased toward high battery level;

$\beta\,(10,10)$ provides a population with two picks of battery level, close to zero and close to one.

Figure 3.8 shows, for each one of the previous battery distribution, the D2D offloading scheme's costs. A distribution heavily biased towards the full charge, allows low incentives. In that case, the minimum incentive an UE accepts rapidly fades to its minimum after the threshold c_0. Decreasing the UEs' average energetic level such cliff move forewords, as the price to convince an user to cooperate increases. Eventually, it can be noticed that our mechanism has precise interval where to find the optimal incentive. Indeed, for values smaller than c_0 nothing happens and the overall cost is equivalent to the one without the offloading scheme. On the other side, the range in which the optimum lies ends at the point in which the cost grows higher than the initial value. After this point, costs begin to grow linearly because of the fact that the number of cluster participants becomes negligible with respect to the population.

System validation. A simplified C++ discrete event simulator has been used to validate the approximation we introduced by (3.4). In this case we simulate static UEs randomly placed in a circle representing the eNB coverage area and for each offer y we count the number of UEs that can accept to be cluster head. Then, based on the UEs spatial configuration of the cluster heads, we compute the actual number of cluster participants. Therefore, we compared the results with the one computed analytically with the same configuration. Figure 3.4 shows the simulation points superimposed to the analytical model, showing a good approximation of the real UEs positioning.

Moreover, to investigate the effect of UEs mobility in the range of $[0-2]$ m/s—a speed compatible with people walking while enjoying a multimedia

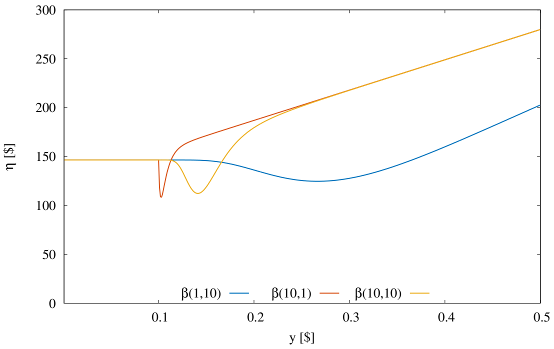

Figure 3.8: Effects of different battery distributions

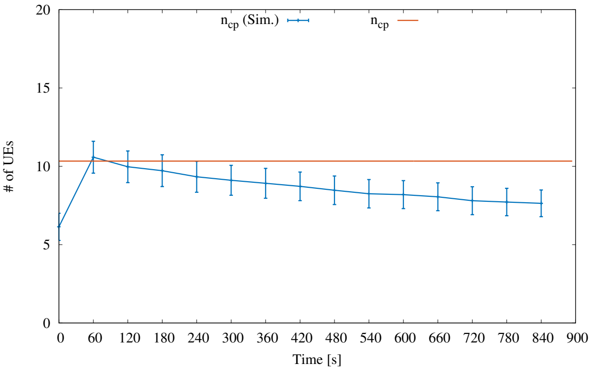

Figure 3.9: Simulation result derived by using ns-3 (the plot shows the average number of cluster participants and the 95% confidence interval).

content and avoiding possible obstacles—, we also simulated our mechanism implementing it in the well known ns-3 network simulator. Here, we also consider physical transmissions to take into account of the actual energy consumption in the UEs and transmission relate issues. Therefore, in Figure 3.9 we show how the average number of cluster participants changes in time, when a low UEs mobility is taken into account. It can be noted that at the beginning our system is very close the the simulation, although a little pessimistic, and as the time elapses—with the UEs moving around randomly—the average number UE reachable through the initial cluster fades. Here τ is equal to ten minutes and the good approximation of the initial period suggest that redesigning the D2D overlay network more frequently we would be able to improve system performances, and results closer to the ones predicted by the analytical model.

3.5 Related work

Several taxonomies of possible D2D architectures have been proposed (e.g., [51]). In particular, according to the used radio spectrum we can have D2D communications that occur on cellular spectrum (i.e., in-band) or unlicensed spectrum (i.e., out-band). Concerning the out-band communications, the coordination between radio interfaces is either controlled by the eNB (i.e, controlled) or by the UEs (i.e., autonomous). Most of the research proposals that focus on in-band D2D communications study the problem of interference mitigation between D2D and cellular communications (e.g., [54, 55, 61, 62]). Concerning out-band D2D communications, the research focuses on power consumption (e.g., [56, 57]) and inter-technology architectural design.

In this chapter we focus on D2D communications where the cellular network operator controls the communication process to provide better user experience and make profit [61, 63]. In particular, we consider how to efficiently use D2D communications for enhancing the quality of wireless multicast services in cellular networks. There are several other studies that address similar issues, see for instance [52, 53, 64–66]. In these papers several paradigms for multicast content delivery in LTE systems based on the joint use of cellular and short-range D2D communications have been proposed. All these proposals point out the potentialities of the different approaches in terms of energy consumption and of bandwidth resources. Our research focusses on the same type of applications but with a different viewpoint: the cooperation of the mobile users. That is, the exploitation of the D2D communications depends on the cooperation of the mobile users that must help the eNB in the content distribution process. A cooperative mobile user consumes an extra amount of energy (e.g., the energy to forward the contents to other mobile users by using D2D communications) and this reduces the battery level. A fundamental issue in this case concerns the definition of appropriate incentive strategies that a cooperative user should receive for his/her additional work (i.e., the relaying). In particular, the papers [67, 68] provide an (abstract) analysis of this issue. Although, we restrict to a particular scenario, the investigation we present in this chapter can be seen as implementation of the abstract strategies proposed in [67, 68] to boost the mobile users cooperation.

3.6 Conclusions

In this chapter we investigate the effectiveness a D2D approach in a simple scenario, where multiple nearby UEs enjoy a common content. We propose a D2D-based strategy for collaborative content delivery in mobile cellular networks, which is suitable for services with high bandwidth demands, such as for instance streaming services. The proposal is based on the cooperation of the mobile users that collaborate with the base station by distributing the received contents to other neighbouring terminals. Therefore, by introducing our D2D based schema an eNB reduces its load by offloading a fraction of it on the mobile users that decide to cooperate in the content distribution process.

However, it is also well known that for their collaboration mobile users incur in additional power consumption. From this it follows that a crucial issue for a successful exploitation of D2D communications concerns the incentive strategies that the network provider must use to involve the mobile terminals in the content distribution process. Moreover, the collaborative attitude of the mobile terminals, their propensity in incurring in a extra power consumption, depends on their battery levels. In other words, a mobile terminal with low battery level is more reluctant to collaborate, and on the other hand when it has a fully charged battery the collaborative attitude is higher.

The strategy presented in this chapter allows the base station to derive an incentive mechanism based on a statistical knowledge of the battery levels of the mobile terminals in its coverage area. The strategy originates an optimization method that the base station can use for deriving an estimate of the offer to provide to the mobile terminals. Eventually, we validated the model representing our incentive mechanism against detailed simulation in both static and dynamic scenarios. We then exploited the model analysis to quantify the cost reduction of the BS thanks to energy savings and the increased spare bandwidth that can be allocated to other services.

Therefore, results presented in this chapter can be used as a foundation for future development. For example, just by introducing some consideration about additional battery consumption, it is possible to apply our approach to UE specific content delivery. Moreover, our model could be enhanced taking into account content popularity, for example to better coalesce terminals. Eventually, as showed before, UEs mobility has a non-negligible impact and therefore it is necessary to integrate it as a feature of the model.

Chapter 4

Machine Type Communications

4.1 Introduction

Factory automation, under the buzzwords Factories of the Future (FoF), or Smart Factories (SF), is a key pillar of the Industry 4.0 concept, and one of the key vertical sectors for 5G technologies [69], together with automotive, healthcare, energy, media and entertainment [41]. 5G classifies the most stringent performance requirements of this application domain in the use case family termed *Tactile Internet / Automation* since they require *time-critical process optimisation to support zero-defect manufacturing.*

Key Performance Indicators (KPIs) defined by the 5G PPP for SF are exceedingly stringent: end to end (E2E) latency between 100 μs and 10 ms, device densities between 10.000 per square km and 100 per square meter, service reliability higher than 99%. Such utmost device densities suggested the identification of SF as the paradigmatic environment for massive Machine-Type Communication (MTC) and a very challenging example of the Internet of Things (IoT).

While the present 5G activities are addressing scenarios that are *either* massive (i.e., with extreme user densities) *or* critical (i.e., with stringent latency requirements), it is quite likely that future evolutions of 5G research will also consider massive *and* critical scenarios, which will emerge in several domains, most notably automotive, health, and, in particular, SF. Therefore, investigating how the 5G technology can cope with an extremely demanding environment such as SF is very important, especially to determine the type and the density of base stations (BSs) that can meet the required KPI targets, together with the associated cost.

To accomplish such task, little exists in the literature that can help to understand the impact of those procedures needed to access resources in a cellular network under extreme operational conditions. The most relevant work in this field is the analytic study described in [1]. In there, the authors developed a probabilistic model for MTC using the LTE technology, and compared the model results to simulation predictions, to show a good match between the two approaches. The model in [1] incorporates many features of the LTE procedures, but does not account for blocking at the BS, does not allow for differentiation of traffic classes, does not generate the latency distribution, and does not provide a closed form solution for the main performance indicators.

In this chapter we describe a stochastic model of the behaviour of environments that, like SF, can be massive, or critical, or massive and critical, incorporating features that will be part of the 5G operations, and evaluating the performance of scenarios typical of a SF environment. The model allows us to evaluate operational conditions, and to derive the distribution of latencies experienced by network access requests. Our network performance analysis proves to be very accurate when results are compared to the predictions of a detailed simulator or to the very detailed analytical model in [1].

Our results show that, for example, with standard system parameters (details are given in the section on numerical results), in order to achieve a success probability not less than 0.9, and a latency not higher than 70 ms, one BS should serve no more than ~ 1400 devices. With a device density equal to 10.000 per square km, this means that the BS can cover an area of radius equal to approximately 200 m. Instead, if we consider the most extreme density envisioned for SF, equal to 100 devices per square meter, the BS can cover only 14 square meters, hence a circle of radius just over 2 m, which would be practically unfeasible even in future SF scenarios! This shows how important it is to carefully evaluate the performance of cellular access in massive MTC environments, and how impactful device density is, which should be definitely taken into account in the design of future wireless access techniques for super-dense device layouts.

4.2 System

We focus our analysis on a single cell, with n Machine-Type Devices (MTDs) that generate new uplink transmissions with a given aggregate rate. In the following, we distinguish two different types of requests: time-critical and non-time-critical, which we identify with two flows, namely the primary and secondary flows. The requests belonging to the primary flow (with intensity λ) can wait at most T_O seconds before being served; otherwise, they are dropped. On the other hand, the requests of the secondary flow (with intensity ℓ) have no timeout, and represent traffic with lower priority, referring to non-real-time

Table 4.1: Notation and Cell Parameters used in the analytical model

Description	*Notation*	*Range*
RACH interval	τ	1 ms
Minimum time needed to reply to a RACH request	$T_{\min}$	0.2 ms
Maximum time allowed to replay to a RACH request	$T_{\max}$	$0.4 \sim 1$ ms
Maximum time needed to establish an RRC connection after a RACH exchange	$W_{\max}$	1 ms
Maximum number of RACH attempts	$k_{\max}$	$10 \sim 40$
RACH collision probability	p_C	$0 \sim 1$
Probability of failure in the RRC connect	$p_{\bar{R}_i}$	$e^{-k_{\max}} \sim \frac{1}{e}$
Maximum number of requests that a base station can serve in a RACH interval	Θ	$12 \sim 24$
Network blocking probability	p_B	$0 \sim 1$
Average RACH backoff at stage i	B_i	10 ms
ACB deferral probability (for flow ℓ)	p_A	$0.05 \sim 0.95$
Random ACB backoff, after the j-th ACB barring event	A_j	$4 \sim 512$ s
Primary/secondary flow rate	λ/ℓ	$9 \sim 0.11$
Timeout (primary flow)	T_O	≤ 10 s
Number of Random Access Preambles	N	54

applications. Table 4.1 shows the notation we use.

In order to access the network, each MTD has to first complete the random access procedure, which initiates as soon as a RACH (Random Access CHannel) opportunity is granted by the BS. The MTD has to go through the RACH each time it has a new message to transmit because downlink traffic is assumed to be sporadic [24].

A request is successful only when resources are actually allocated to the MTD; that is, we take into account also signalling messages that are exchanged $a$fter the random access procedure successful completion. Indeed, the 3GPP-defined procedure to access resources includes the RACH phase and the RRC (Radio Resource Control) connect phase, resulting in the exchange of four messages. When either of the two phases fails, the MTD retries after a random backoff interval. Multiple timeouts are used in the overall procedure, in the event of a collision, of an early access failure (when no RRC connect message is exchanged before a time $T_{\max}$ from the beginning of the RACH opportunity used by the MTD) or a late access failure (when the RRC connect phase starts, but no final resource allocation is notified to the MTD within a window $W_{\max}$ from the beginning of the RRC connect phase). More details on the timing of a request will be provided in the next section when describing our model.

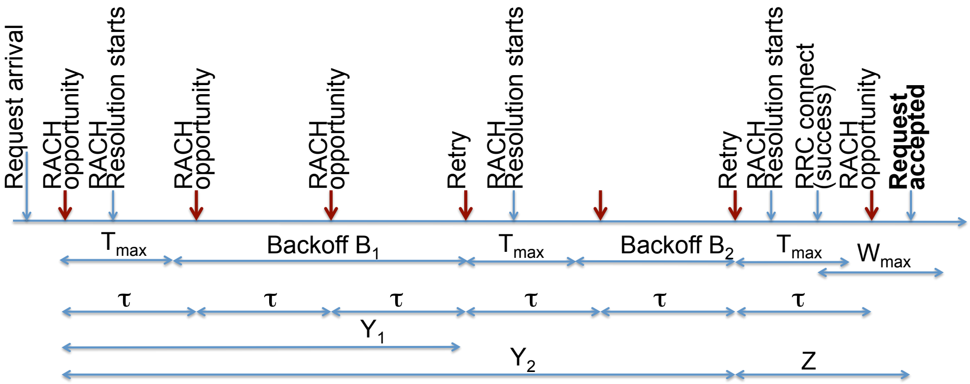

Figure 4.1: Timing example for a primary flow request served after 3 attempts.

Fig. 4.1 shows an example of access request that succeeds after 2 retries over the Random Access.

In the system we just described, a message transfer can take place after the successful completion of two subsequent steps: the RACH and the RRC connection procedures. The RACH can be divided into $k_{\max}$ sequential stages, one for each allowed RACH attempt (after $k_{\max}$ attempts, a request is dropped). Access requests move from one stage to the next in case of collision (with probability p_C) and in case the request gets lost (i.e., it is not correctly received and acknowledged by the BS). The latter event occurs with probability $p_{\bar{R}_i}$, which is different at each attempt, due to the standard power ramping mechanism: nodes progressively increase the power used to transmit RACH requests after each failed attempt [46].

The dynamic of RACH requests in the system is presented in Fig. 5.1. In each stage, a request can leave because of a success (the MTD transmits its data). The request can however also leave the system because of a failure, which can consist in either a network blocking due to a shortage of queueing resources at the network processor after a successful RRC connection procedure or because of a timeout. Moreover, a request can move from stage i to $i+1$ because of a collision, or any event that precludes the success of the RRC connection procedure: either the request is not decoded by the BS, or the BS does not have resources to send an acknowledgement and decides to drop the request (we indicate with Θ the maximum number of requests the base station can acknowledge in each RACH interval τ). In addition, a request can retry the RACH procedure at most $k_{\max}$ times. Otherwise, it leaves the system with a failure. Notice that passing from a stage to the next incurs a random delay due to backoff. Moreover, the secondary flow incurs RACH access deferring with fixed "barring" probability p_A, and multiple back-to-back deferrals are possible, so that secondary flow requests incur additional delay, due to standard Access Class Barring (ACB) operation [70].

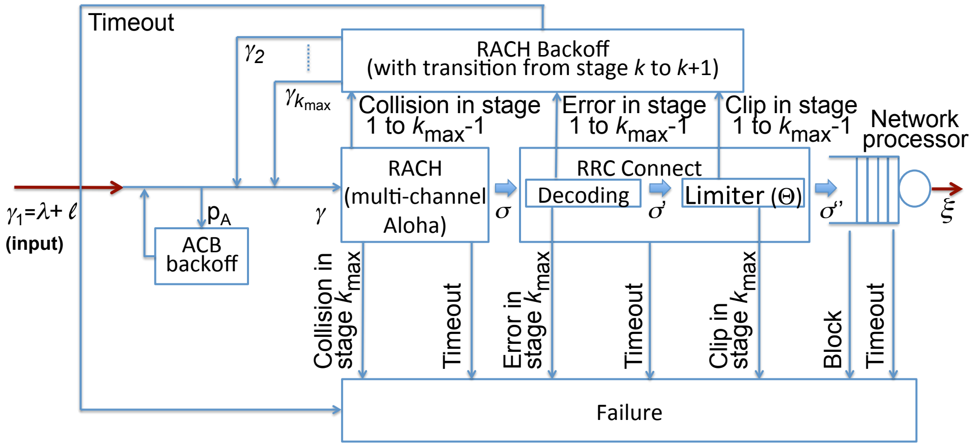

Figure 4.2: Block diagram representing the system with primary and secondary flows accessing resources via RACH channels and RRC connect procedure. Flow ℓ is subject to ACB and all accepted requests are served in FIFO order.

4.3 Analytical model

For the sake of compactness and readability, we provide the reader with the definitions of the variables used in the analysis in Table 2.1. Moreover, the random variables representing intervals of time used in the model are pictorially presented in Fig. 4.1, while flows entering and leaving the RACH system are indicated in Fig. 5.1 jointly with the system throughput ξ. For the sake of tractability, the input to the considered system is assumed to be a Poisson process with intensity γ_1. The accuracy of such assumption will be later validated in the numerical evaluation section (see Fig. 4.4).

Structure of the request sojourn time

A RACH request enters the system in stage 1 and leaves in any stage $i \in \{1, \ldots, k_{\max}\}$ upon a success, a network blocking, an excessive number of retries, or a timeout. If a request leaves the system from stage i because of either a success or a network blocking, it has been in the system for a time Y_{i-1} (more precisely, either $Y_{i-1}^{(\lambda)}$ or $Y_{i-1}^{(\ell)}$, depending on the flow considered) which consists of $(i-1)$ times the interval $T_{\max}$ and $i-1$ backoffs, plus a random interval Z needed to model the delay between RACH request and network grant (see Fig. 4.1 for $i=2$)—the latter being independent from Y_i—and a random number of barring backoffs for requests of the secondary flow. Similarly, in the case of an excessive number of retries, the time spent in the system is $Y_{k_{\max}-1} + Z$. In the case of timeout, of course, the time spent is T_O. When passing from stage i to $i+1$, the time spent until the stage transition is simply Y_i. As we will see later, the above quantities are sufficient to describe the entire sojourn in the system and to evaluate the performance of the system in terms of, among other quantities, network blocking probability, timeout probability, throughput and sojourn time. With the notation described in Table 2.1, the distribution of $Y_i^{(\ell)}$ is

$$F_{Y_i^{(\ell)}}(x)=\Pr\left\{ i\,T_{\max} + \sum_{k=1}^{i} B_k + \sum_{j=0}^{L} A_j \leq x \right\}, \tag{4.1}$$

where L is the random number of back-to-back deferrals experienced because of ACB. The distribution for the primary flow, namely $F_{Y_i^{(\lambda)}}(x)$ is omitted because it can be derived from $F_{Y_i^{(\ell)}}(x)$ by plugging $L=0$. The backoff random variables B_k and A_j are independent among them and from Z, although not necessarily identically distributed. In contrast, variables $Y_i^{(\lambda)}$ depend on $Y_k^{(\lambda)}, \forall k < i$. Similarly, variables $Y_i^{(\ell)}$ depend on $Y_k^{(\ell)}, \forall k < i$.

The distribution of the time spent by a request until the resolution of the i-th RACH attempt, for the secondary flow, is expressed as the distribution of the random variable $Y_{i-1}^{(\ell)} + Z$:

$$F_{Y_{i-1}^{(\ell)}+Z}(x)=\Pr\left\{(i-1)T_{\max}+\sum_{k=1}^{i-1} B_k+\sum_{j=0}^{L} A_j+Z\leq x\right\}; \tag{4.2}$$

and for the primary flow it is enough to use (4.2) with $L = 0$ to derive $F_{Y_{i-1}^{(\lambda)}+Z}(x)$. Since Z is independent from $Y_i^{(\lambda)}$ and $Y_i^{(\ell)}$, and denoting by f_Z the p.d.f. of Z, the following useful results also hold: $F_{Y_{i-1}^{(\lambda)}+Z} = F_{Y_{i-1}^{(\lambda)}} * f_Z$ and $F_{Y_{i-1}^{(\ell)}+Z} = F_{Y_{i-1}^{(\ell)}} * f_Z$.

Stage probabilities

At stage i, a request leaves the system because of either a success, a network blocking, or a timeout (primary flow). In all other cases, the request moves from stage i to stage $i + 1$, with the exception of stage $k_{\max}$ for which an attempt to pass to stage $k_{\max+1}$ results in a failure due to an excessive number of retries. Here we derive stage probabilities for the primary flow only. However, the same equations hold for the secondary flow by replacing λ with ℓ and using $T_O \to \infty$.

Stage transitions. Denoting by p_C the RACH collision probability, which is the same for all RACH attempts, and by $p_{\bar{R}_i}$ the probability of an error in the RRC connect procedure after the i-th RACH attempt, stage transition probabilities $P_N^{(\lambda)}(i)$ are computed as the probability to reach stage $i+1$ going through all previous i stages. The described quantities only depend on the aggregate load in the RACH and on the resources available at the BS.

For a request in the primary flow, the transition to the next stage occurs when there is either a collision or an RRC connect failure, therefore with probability $1-(1-p_C)(1-p_{\bar{R}_i})$, but only if the timeout has not expired before the end of the RACH backoff in that stage, i.e., with probability $F_{Y_i^{(\lambda)}}(T_O)$. This results in the following iterative computation, $\forall i \geq 1$:

$$P_N^{(\lambda)}(i) = P_N^{(\lambda)}(i-1)\left[1-(1-p_C)\left(1-p_{\bar{R}_i}\right)\right] F_{Y_i^{(\lambda)}}(T_O); \tag{4.3}$$

where $P_N^{(\lambda)}(0)=1$ by definition. Note that, since $k_{\max}$ is the maximum retry number, $P_N^{(\lambda)}(k_{\max})$ is a failure probability.

Success. The probability of a request succeeding in stage i, $\forall i \geq 1$, is the probability of reaching stage i and then have no collision in the RACH, no error in the RRC connect phase, and no network blocking. At the same time, no timeout has to occur while waiting for the resolution of the i-th RACH attempt. Hence, denoting the conditional network blocking probability by p_B, given that a request succeeds on the RACH, the following recursive relation holds:

$$P_S^{(\lambda)}(i) = P_N^{(\lambda)}(i-1)(1-p_C)\left(1-p_{\bar{R}_i}\right)(1-p_B)F_{Y_{i-1}^{(\lambda)}+Z}(T_O). \tag{4.4}$$

We denote by $P_S^{(\lambda)}$ the total success probability for the primary flow. Such quantity is computed by summing the success probabilities (4.4) over the stages.

In the case of success, the request receives service, and the time spent in the system before service, for a request on the primary flow, results to be a random variable $Y_{i-1}^{(\lambda)} + Z$.

Blocking. When a request successfully passes both the RACH and RRC connect phases, it can be either admitted to the service or blocked because of lack of resources at the network processor of the BS. The probability that a request is blocked by the network in any stage i, can be computed as

$$P_B^{(\lambda)}(i) = P_N^{(\lambda)}(i-1)\,(1-p_C)\left(1-p_{\bar{R}_i}\right) p_B\, F_{Y_{i-1}^{(\lambda)}+Z}(T_O). \tag{4.5}$$

We denote as $P_B^{(\lambda)}$ the total blocking probability of flow λ.

In the case of blocking, the time spent in the system is exactly like in the case of success (now excluding the service time), i.e., for a request on the primary flow, it is $Y_{i-1}^{(\lambda)} + Z$.

Timeout. Requests of flow λ can experience timeout in stage i if they reach stage i and: 1) either the random access or the RRC connect fail, and the backoff delay leads to exceeding the timeout; or 2) the RRC connect attempt is not resolved within the timeout. The time spent in the system is of course T_O, but it is also a value obtained from the r.v. $Y_i^{(\lambda)}$ or $Y_{i-1}^{(\lambda)} + Z$. The resulting timeout probability can be expressed via the cumulative functions of those r.v.'s:

$$P_{TO}(i) = P_N^{(\lambda)}(i-1)\left\{(1-p_C)\left(1-p_{\bar{R}_i}\right)\left[1-F_{Y_{i-1}^{(\lambda)}+Z}(T_O)\right]\right.$$
$$\left.+\left[1-(1-p_C)\left(1-p_{\bar{R}_i}\right)\right]\left[1-F_{Y_i^{(\lambda)}}(T_O)\right]\right\}. \tag{4.6}$$

We further denote as P_{TO} the total timeout probability.

Closed form for probability expressions. Although we have presented iterative expressions, one can notice that all of the above expressions can be easily re-written in closed form. Indeed, it is enough to notice that stage transitions probabilities can be put in the following closed form, $\forall i \geq 1$:

$$P_N^{(\lambda)}(i) = \prod_{k=1}^{i}\left\{\left[1-(1-p_C)\left(1-p_{\bar{R}_k}\right)\right]F_{Y_k^{(\lambda)}}(T_O)\right\}. \tag{4.7}$$

The above expressions can be used in all other expressions found in this section to derive probabilities in closed form.

Remark on the generality of stage probability expressions. All expressions derived in this section are valid independently from the distribution of backoff events and ACB configuration, and can be easily generalised for the case with no limit on the number of RACH attempts (i.e., for $k_{\max} \to \infty$). As it is easy to check, the sum of success, blocking, and timeout probabilities, plus the stage transition probability in stage $k_{\max}$, i.e., the sum over all events in which a request leaves the system, is identically 1 for all possible values of parameters and distributions used, which has to hold because an MTD request eventually has to leave the system.

Analysis of random access operation

To compute the expressions for p_C, p_B and $p_{\bar{R}_i}$ to plug in the stage probability expressions derived above, we model the RACH operation as a multi-channel slotted Aloha system with random backoff after a collision and with a finite number $k_{\max}$ of attempts. We consider the typical 3GPP procedure in which access requests are transmitted with increasing power after each failure and the BS can receive corrupted RACH messages even in the case of no collision, with probability e^{-i}, with the power used in stage i, as modeled in [46]. Moreover, the BS can serve a limited number of requests per RACH opportunity interval, namely Θ access requests each τ seconds, where τ is the spacing between two subsequent Random Access Opportunities (RAOs) and users can choose between N orthogonal RACH preambles to request access.

RACH collision probability. Given that, regardless the actual stage, all the requests performing random access share the same resources, the collision rate is the same at all stages, and depends on the total RACH load γ, including both primary and secondary flows. Hence, the collision probability in the resulting multi-channel slotted Aloha with N channels and slot duration τ, is simply expressed as $p_C = 1 - e^{-\frac{\gamma\tau}{N}}$.

With one primary flow of intensity λ arrivals per second, plus a secondary flow of intensity ℓ, the load of the RACH is given by the sum of arrivals at each stage of the RACH:

$$\gamma = \gamma^{(\lambda)} + \gamma^{(\ell)} = \sum_{i=1}^{k_{\max}} \gamma_i^{(\lambda)} + \sum_{i=1}^{k_{\max}} \gamma_i^{(\ell)}. \tag{4.8}$$

where γ_i is the RACH load due to attempts of connections that have already failed the random access $i-1$ times.

In turn, the load entering stage i due to the primary flow is simply given by the total intensity of the flow times the probability to reach stage i, which is given by (4.7), i.e.:

$$\gamma_i^{(\lambda)} = \lambda \prod_{k=1}^{i-1} \left\{ \left[1 - (1-p_C)(1-p_{R_k})\right] F_{Y_k^{(\lambda)}}(T_O) \right\}. \tag{4.9}$$

The expression of $\gamma_i^{(\ell)}$ is similar, but for the fact that $\lim_{T_O \to \infty} F_{Y_k^{(\ell)}}(T_O) = 1$, and therefore we omit it.

Failure of RRC connect. After a success in the random access phase, an access request may not receive an answer either because of channel errors or because the BS is saturated, which happens when the output σ of the multi-channel slotted Aloha is greater than a maximum rate Θ.

As concerns channel errors, since the power ramping mechanism is taken into account, at each subsequent stage, requests are detected with an increasing probability $1 - e^{-i}$, where i is the current stage index [46].

As concerns exceeding the base station capacity Θ, let's consider the output of the RACH at each stage, namely σ_i, which is simply given by the load at that stage, times the probability of having no collision, i.e.: $\sigma_i = \left(\gamma_i^{(\lambda)} + \gamma_i^{(\ell)}\right)(1 - p_C)$. However, part of the non-collided RACH requests are received incorrectly by the BS, depending on the stage in which they are, so that the actual number of requests to accommodate is $\sigma_i' = \sigma_i \left(1 - e^{-i}\right)$, which is $\sigma' = \sum_{i=1}^{k_{\max}} \sigma_i'$ in total.

With the above, the number of correctly received requests in a RAO is, on average, $\sigma'\tau$. Considering that the RACH behaves as a slotted Aloha system with N independent channels (one for each orthogonal RACH preamble) with binary output, the number of correctly decoded access requests at the BS can be modeled as a binomial process with success probability $\sigma'\tau/N$. Note that the throughput of a multi-channel slotted Aloha is upper-bounded by the number of channels, which guarantees that $\sigma'\tau/N \leq 1$. The resulting mass probability function can be written as follows:

$$\pi'_j = \binom{N}{j}\left(\frac{\sigma'\tau}{N}\right)^j\left(1-\frac{\sigma'\tau}{N}\right)^{N-j}, \ \forall j \in \{0,\dots,N\}. \tag{4.10}$$

At most Θ requests can be answered in a RAO, and we denote by $\sigma''\tau$ the average value of the corresponding random process. The average loss N_L due to clipping to Θ is

$$E\left[N_L\right] = \left(\sigma' - \sigma''\right)\tau = \sum_{j=\Theta+1}^{N} \left(j-\Theta\right)\pi'_j. \tag{4.11}$$

Since clipping is enforced independently of the RACH stage, the losses are uniformly spread over the stages: $\sigma''_i = \sigma'_i\left[1-\frac{E[N_L]}{\sigma'\tau}\right]$. Hence, combining the probability to incorrectly decode a request or that the BS cannot answer the request, we derive the RRC connect failure probability:

$$p_{\bar{R}_i} = 1 - \frac{\sigma''_i}{\sigma_i} = 1 - \left(1-e^{-i}\right)\left(1-\frac{E\left[N_L\right]}{\sigma'\tau}\right). \tag{4.12}$$

Notice that the computation of $\gamma_i^{(\lambda)}$, $\gamma_i^{(\ell)}$, p_C and $p_{\bar{R}_i}$ requires an iterative approach, which can be solved by finding the fixed point for $\gamma = f(\gamma)$, where $f(\gamma)$ results from using the expressions of p_C and $p_{\bar{R}_i}$ in $\gamma_i^{(\lambda)}$ and $\gamma_i^{(\ell)}$ and summing to compute the aggregate RACH load.

Blocking probability. The maximum number of MTDs allowed to access the network for packet transmission per unit of time is constrained by the transmission rate C of the devices (which equals the rate at which the BS operates) and the mean packet length P_L. Denoting with $E[S] = \frac{P_L}{C}$ the network service time, the flow of requests approaching the network exceeds the BS capacity as soon as the offered load $\rho = \sigma'' E[S]$ becomes greater than 1. The latter happens when the number of accepted requests in a RAO, $\sigma''\tau$, is larger than $\frac{\tau}{E[S]}$. The maximum number of MTDs' requests that can fit in a RAO unit is then $m = \lfloor \tau/E[S] \rfloor$. Requests in excess of m are blocked. Since the BS replies to access requests in an interval that can be considered as uniformly distributed and with no memory, to compute the blocking probability, we use σ'' as the arrival rate of a M/D/1/m queue. The resulting blocking probability is [71]:

$$p_B = (1-\rho)E_m/(1-\rho E_m), \tag{4.13}$$

where $E_m = 1-(1-\rho)\sum_{j=0}^{m}\frac{(-1)^j\rho^j(m-j)^j e^{\rho(m-j)}}{j!}$.

Network throughput. From the above simple approximate analysis, the resulting flow of requests successfully accessing the network is simply $\xi = \xi^{(\lambda)} + \xi^{(\ell)} = \lambda P_S^{(\lambda)} + \ell P_S^{(\ell)}$.

Sojourn time distribution

Primary flow. The distribution of the time spent in the system (not including the service time) for an access attempt in the primary flow is computed by noting that a request exits the system at a generic stage i if one of three disjoint events happens: 1) success, 2) blocking and 3) timeout. In addition to this, at stage $k_{\max}$, any failure in the random access causes a drop as well, even if the timeout has not expired. All the described events are mutually exclusive and cover the entire space of probability for the event of leaving the system. Hence, the CDF of the time $T^{(\lambda)}$ spent in the system by a request can be written by using the total probability formula as follows:

$$F_{T^{(\lambda)}}(x)=\sum_{i=1}^{k_{\max}} P_{TO}(i)\,\mathcal{U}\,(x-T_O)+\frac{F_{Y_{i-1}^{(\lambda)}+Z}\,(x)}{F_{Y_{i-1}^{(\lambda)}+Z}\,(T_O)}\sum_{i=1}^{k_{\max}}\Big(P_S^{(\lambda)}(i)$$

$$+\;P_B^{(\lambda)}\,(i)\Big)+P_N^{(\lambda)}(k_{\max})\,\frac{F_{Y_{k_{\max}-1}^{(\lambda)}}\,(x-T_{\max})}{F_{Y_{k_{\max}-1}^{(\lambda)}}\,(T_O-T_{\max})}, \tag{4.14}$$

where $\mathcal{U}$ is the unit step function centred in T_O. However, if we consider that failures for blocking or excess retries are equivalent to timeouts, we consider as T_O the latency in case of any failure and then simplify the above formula as follows:

$$F_{T^{(\lambda)}}\,(x)=\sum_{i=1}^{k_{\max}}\left(P_S^{(\lambda)}(i)\,\frac{F_{Y_{i-1}^{(\lambda)}+Z}{}^{(x)}}{F_{Y_{i-1}^{(\lambda)}+Z}{}^{(T_O)}}+\Big(1-P_S^{(\lambda)}(i)\Big)\mathcal{U}(x-T_O)\right). \tag{4.15}$$

For designing and dimensioning purposes, a more insightful indicator should only take into account the time spent within the system until a success. Hence, we derive the cumulative probability function of $T^{(\lambda)}$ given a success as

$$F_{T_{|S}^{(\lambda)}}\,(x)=\sum_{i=1}^{k_{\max}}\frac{P_S^{(\lambda)}\,(i)}{P_S^{(\lambda)}}\,\frac{F_{Y_{i-1}^{(\lambda)}+Z}\,(x)}{F_{Y_{i-1}^{(\lambda)}+Z}\,(T_O)}. \tag{4.16}$$

Secondary flow. In case of an access request belonging to the secondary flow, the expressions of the sojourn time $T^{(\ell)}$ are similar to the ones derived for the primary flow, except for the absence of timeout events (i.e., $T_O\to\infty$).

4.4 Model positioning

The model described so far is rather simple, and its solution requires low computational complexity. The heaviest part consists in computing the CDFs of $Y_i^{(\lambda)}$, $Y_i^{(\ell)}$ and Z, which can be done just once, offline. Moreover, deriving those distributions in closed form is trivial in case of simple distributions of backoffs. Indeed, the first ones are convolutions of exponential random

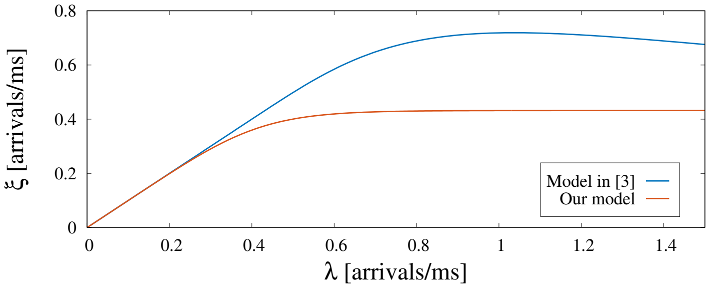

Figure 4.3: Comparison between our simple model and the M2M model by Madueño *et* al. [1], which follows the behaviour of LTE-A signalling operations in detail. The latter is not meant to describe the behaviour of the system in saturated conditions, and hence a fair comparison with our model is possible only in the leftmost part of the figure.

variables (i.e. Erlang distributions) while the latter, Z, is a convolutions of an exponential random variable and a uniform one. After computing the CDFs, one only needs to solve iteratively the equations described above. However, few iterations are enough for accurate results (observations not reported here for lack of space show that less than 5 iterations are needed) and each iteration scales linearly with the number of stages $k_{\max}$.

Our model is generic, since it can be used for arbitrary population sizes and time constraints, so that it can be useful to design massive as well as mission-critical SF scenarios.

Our model does not consider in deep detail the operations of signaling channels and access techniques of real networks, e.g., LTE/LTE-A. This implies that we need to validate our model against realistic simulations. However, before proceeding with a complete validation and performance evaluation, here we show that the results of previous very detailed models do not substantially depart from ours. In particular, we consider a model recently proposed by Madueño *et* al. [1], which can be used for the evaluation of M2M unsaturated scenarios, with sparse traffic and small payloads, RACH retries and dropped requests. The main differences between the model in [1] and ours consist in the fact that [1] models LTE-A signalling channels very accurately, that requests are never dropped because of lack of transmission resources, but only because of user impatience, and that users never return to the network before the RRC timeout.

Fig. 4.3 compares the predictions obtained with our model and with the model in [1]. In order to perform a fair comparison, we used the same configuration parameters for the two models. Specifically, we used the parameters suggested in [1] for M2M traffic, with a narrowband LTE-A cell (1.4 MHz, resulting in 12 OFDMA resource blocks per ms, $\tau = 10$ ms, $N = 54$) and a slow modulation and coding scheme (3.456 Mb/s) for all data and signalling

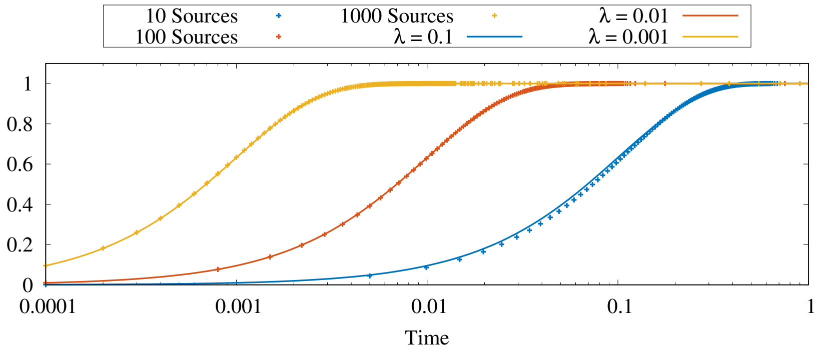

Figure 4.4: Comparison of the CDF of a Poisson arrival stream against the one resulting from the superposition of $10, 100, 1000$ arrival streams with interarrival times distributed according to a $\mathcal{U}(0.9, 1.1)$.

channels. We use 1 kbyte as fixed payload size and 40 ms as maximum waiting time for a request queued for service. Accordingly, in our model, we use $m = 4$ and $\Theta = 72$, which correspond to queue and serve RACH request in at most 40 ms. Fig. 4.3 shows the system throughput vs. the exogenous arrival rate generated by users. The two models behave quite similarly at low loads, i.e., in the range for which the model in [1] was designed. However, when approaching saturation, the two models substantially deviate from each other. Indeed, the model in [1] achieves unrealistically high throughputs, beyond the feasible bound imposed by channel speed (the flat region in the curve of our model) because that model does not consider that messages can be dropped because of lack of transmission resources over the PUSCH channel. Those resources are instead limited, as taken into account by our model. This comparison proves that our simple model can be as accurate as a more complex and detailed one, while at the same time resulting in a much more flexible and suitable tool for the evaluation of SF radio access.

4.5 Numerical Results

Arrival process Poisson approximation. In this chapter, we consider industrial (i.e., SF) scenarios where the network traffic consists of data from large numbers of MTDs. In the case of real-time control, data is normally generated from MTDs at quasi-deterministic intervals. On the contrary, data generation for monitoring and maintenance applications can be assumed more random.

As a consequence, modelling the request arrival processes as Poisson might appear an unacceptable simplification. However, it is well known that (in general) the Poisson process is the limit collective behaviour for increasing number of sources that independently generate arrivals. Specifically, the Palm-

Khintchine theorem expresses that a large number of not necessarily Poissonian renewal processes combined will have Poissonian properties. Therefore, we performed a set of simple simulation experiments, comparing the interarrival time CDF generated by a Poisson process against the one produced by different numbers of sources. Fig. 4.4 shows some of the results we have obtained. In particular, in this figure, we compare Poisson arrivals against the process resulting from superpositions of processes with interarrival times distributed according to a uniform distribution in the range $[0.9, 1.1]$ (Fig. 4.4-b). In the experiments, we vary the number of sources that generate arrivals, as well as the average number of total requests for each case. We can clearly see that the CDFs are very similar already for 10 independent sources, and become identical for 1000 sources. Since in SF scenarios the number of MTD is extremely high, we consider Poisson arrivals a reasonable approximation, even for relatively small numbers of MTDs.

SF experiment parameters. Since the focus of this work is on traffic generated by autonomous and automatic MTDs reporting to a central entity collecting data in the SF, single transmissions are of negligible dimensions and we assume $P_L = 1000$ bits as a realistic value. Based on application-specific constraints (due to real-time sensing and control), the traffic has a cyclic nature; therefore the duration of the cycle depends on the maximum allowable latency. In the following, we use a timeout $T_O = 100$ ms and a message generation interval equal to $\frac{4}{3}T_O$, so that any MTD generates a new message every 133.3 ms, on average. Moreover, we assume that MTDs can transmit at $C = 10$ Mb/s. With the above, the number of requests that can be served in a RAO is $m = 10$. Latencies strongly depend on the frequency of RACH opportunities. Here we use $\tau = 1$ ms, which corresponds to a RACH opportunity every 10 data slots in upcoming LTE Advanced Pro and 5G systems [72]. RACH and RRC connect timers are set to be of the order of magnitude of τ. Specifically, we use $T_{\min} = 0.2$ ms, $T_{\max} = 0.8$ ms and $W_{\max} = 1$ ms (respectively 2, 8 and 10 time slots). The number of RACH channels is $N = 54$, which is a typical value in 3GPP specifications. Unless otherwise specified, we use $\Theta = 18$ requests/ms which is realistic for 4G/5G base stations in which there can be up to 3 acknowledgements per time slot during $T_{\max} - T_{\min}$, and set the maximum number of retries to $k_{\max} = 10$. Note that, although in the simulator we consider many operational details of resource request and grant procedures, we do not enter into the details of the signalling channel protocol, which are specific of each cellular implementation. As concerns backoff timers, we use $E[B_i] = 10$ ms and $E[A_{ij}] = 4$ s for RACH and ACB retries, respectively, and $p_A = 0.5$, although the importance of $E[A_{ij}]$ and p_A is not shown since they only affect the latency of flow ℓ without impairing any throughput.

System behaviour and model validation. Fig. 4.5 presents the most significant quantities to characterise the system behaviour in presence of the primary flow only. With the parameters described above, the upper part of the figure illustrates the dome-shaped relations between the system input λ and

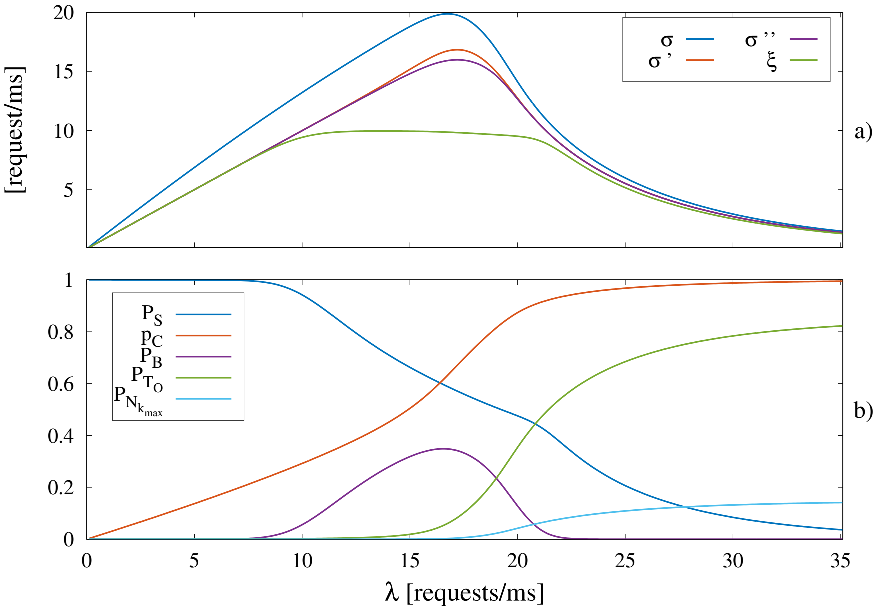

Figure 4.5: System behaviour in the reference scenario in presence of flow λ: a) network throughput and flows leaving the RACH and the RRC connect phases with a success; b) stage probabilities.

i) the amount or requests per unit time that pass the RACH without collision (σ, which is at most $\frac{N}{e}$, i.e., the max throughput of an N-channel Aloha), ii) the amount or requests per unit time that reach the base station with no decoding error (σ'), iii) that complete the RRC connect phase (σ'', which is limited by Θ), and iv) that eventually receive service (ξ, which is capped by m). Because of the structure of the system, the typical Aloha output flow σ is progressively scaled and flattened to become the system throughput ξ. We can identify 3 regions for ξ. An initial linear region in which the throughput grows almost linearly with the input; a flat region in which the throughput is practically constant or slightly recessing; and a breakdown region in which small increments of the input cause large throughput degradation.

Fig. 4.5-b gives some insight into the system reactions to progressively higher traffic loads. It is clear that in the linear region, the system works just fine: p_C is quite low, and both $P_{TO} = \sum_{i=1}^{k_{\max}} P_{TO}(i)$ and $P_B = \sum_{i=1}^{k_{\max}} P_B(i)$ are negligible, while the total success probability $P_S = \sum_{i=1}^{k_{\max}} P_S(i) \simeq 1$. However, as soon as the network throughput gets close to its maximum m, P_B begins to grow, and the system enters the flat region. This point corresponds to the first knee of ξ. Then, P_B grows higher, up to its maximum, corresponding to the largest RACH throughput. From this point on, the system behaviour is driven by p_C and P_{TO}. Indeed, the probability to leave the system shifts from low RACH stages towards higher ones (not shown because of space limitations)

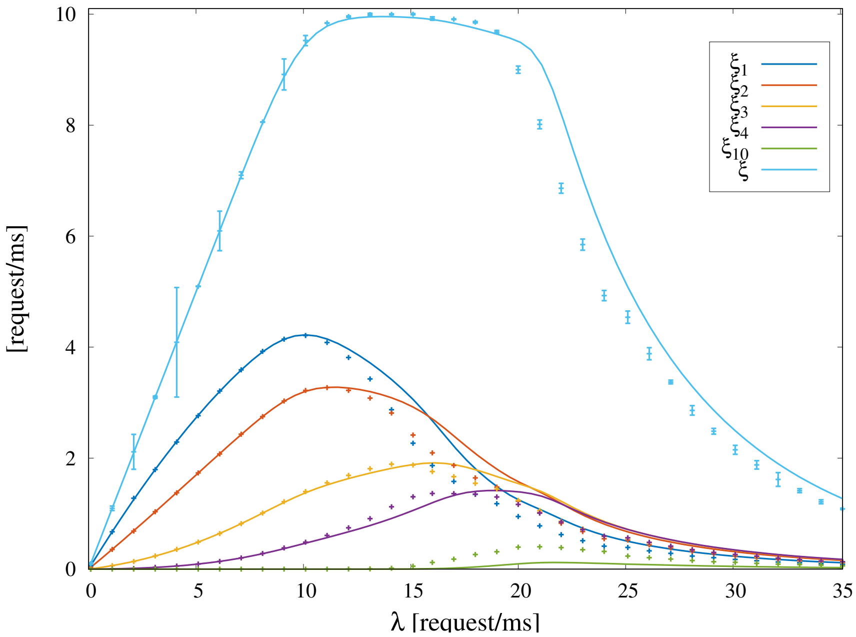

Figure 4.6: Validation of the analytical model through simulation with primary flow only: network throughput $\xi^{(\lambda)}$ and its per-stage components $\xi_i^{(\lambda)}$ (for readability the figure only shows $\xi^{(\lambda)}$, $\xi_1^{(\lambda)}$, $\xi_2^{(\lambda)}$, $\xi_3^{(\lambda)}$, $\xi_4^{(\lambda)}$, and $\xi_{10}^{(\lambda)}$).

since requests, on average, retry several times before leaving the system. Similarly, it can be observed that the stage in which a success occurs shifts to high stage numbers, as shown in Fig. 4.6, where throughput components are illustrated. This same figure also shows the good accuracy achieved by our model in terms of throughput predictions. Indeed, analytical predictions match well the results of the detailed packet-level simulator we developed in Python. We can observe some limited, yet non-negligible, errors only in the rightmost region of ξ, which contains, however, no desirable operational points due to low success probability and, as we will show later, very high latency.

From these initial results, it is already clear that, to obtain a sufficiently good QoS level, it is desirable to keep the system in operational regimes below the point where the RACH saturates, before the beginning of the flat region of ξ.

Impact of transmission rate and packet size. An obvious relation exists among the system throughput (the rate of requests successfully accessing the network, i.e., ξ), the network data rate C, the packet size P_L, and the number of requests that can be processed by the network in a time interval τ (i.e., m). For fixed τ, m only depends on the ratio $\frac{P_L}{C}$. Therefore, to understand the impact of C or P_L on throughput, it is enough to evaluate the impact of m. To this aim, Fig. 4.7-a shows the effect of different values of m on the system throughput, while the rest of the parameters is kept as

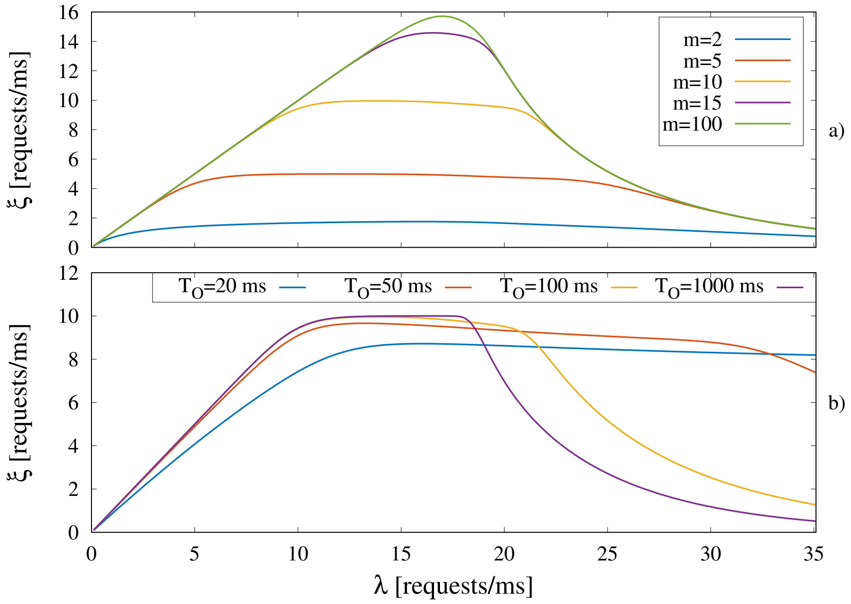

Figure 4.7: Effect of varying model parameters on ξ: *a)* Number of clients served per RAO, m; *b)* Timeout

before. It is worth to point out that, independently of m, the throughput is limited by Θ (i.e., the max rate at which the BS can accept requests), so that high values of m perform practically the same. This can be translated into the following very relevant statement for system design and planning: *B*S capacity increases can lead to (very) small performance improvements.

Impact of timeout. Timeout is a very critical aspect of system design, due to the real-time nature of most of the traffic in SF. Fig. 4.7-b sheds light on the impact of the timeout value on system performance. In particular, we can observe that higher timeout values make MTDs saturate the network sooner. When the network is saturated, increases in λ lead to higher values of p_C, which cause a drastic decrease of ξ. Interestingly, low timeout values impact network throughput also for low input rates, while medium to high values of the timeout only impact the beginning of the breakdown region.

Latency performance. Fig. 4.8 shows how latency is affected by increasing incoming traffic λ. The two pictures summarise this information through box-and-whiskers diagrams, built with the first, 25-th, 50-th, 75-th and 99-th percentiles of the latency of a request, considering the time from its arrival to the moment it leaves the system (with either a success or a failure). In particular, Fig. 4.8-a shows that for values of λ in the range $[0-10]$ the network guarantees a latency lower than 20 ms up to the 75-th percentile, and within

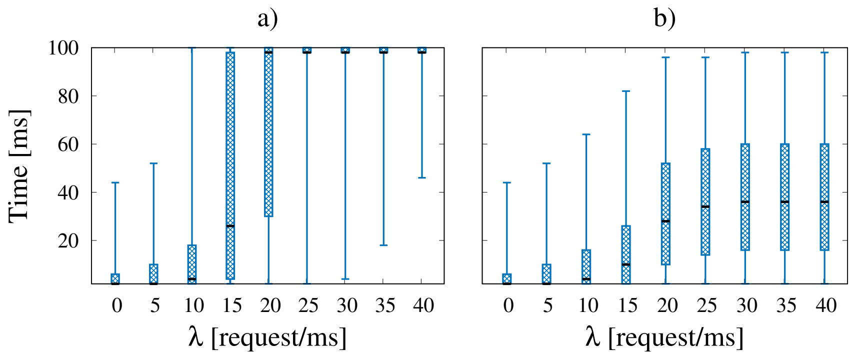

Figure 4.8: Latency distributions computed with our model: lower and higher values correspond respectively 1st and 99-th percentiles; lower and higher ends of the boxes represent the first and third quartile, whilst the median is represented as a black tick. *a)* Distribution of latency of successful and unsuccessful requests, based on (4.15); *b)* Distribution of latency conditioned by a success, based on (4.16).

the timeout (100 ms) up to the 99-th percentile. Note that the range $[0 - 10]$ of λ, is the one for which we saw that the throughput increases linearly. The same kind of results is reported in Fig. 4.8-b, where latency percentiles are conditioned to a success. For higher values of λ, latency significantly grows in the breakdown region, which is, therefore, an undesirable region also from the point of view of latency guarantees.

Sustainable cell population. The key question that an SF network designer has to face is how many cells are necessary to serve a given population of MTD, while providing a predefined QoS level. Our model answers this question by computing the mapping between KPIs and number of MTDs. Let us focus on a single cell operated with the default realistic parameters considered in this section. Fig. 4.9 shows the maximum number of MTDs that can access the network (in the vertical axis) when the 99-th percentile of latency, conditioned to a success, is guaranteed (the value that labels the curves in the figure), as function of the guaranteed total success probability P_S (in the horizontal axis). That is, the curves provide the greatest value of n that guarantees a latency with a 99-th percentile lower than a threshold (the curve label) and a success probability higher than another threshold (the abscissa). As a possible example, we see that one cell is able to handle (roughly) 2100 MTDs, guaranteeing latencies smaller than 90 ms for 99% of the requests, with $P_S \geq 0.6$ (this can be a condition which is representative of a massive scenario, which is however not critical, due to the low success probability value). However, when it comes to serving MTDs with high success probability (say above 90%), and low latency (say below 50 ms at the 99-th percentile of distribution), only a few hundred devices can be connected to a BS. This can be acceptable in a scenario that is critical, but not massive. On the contrary, in

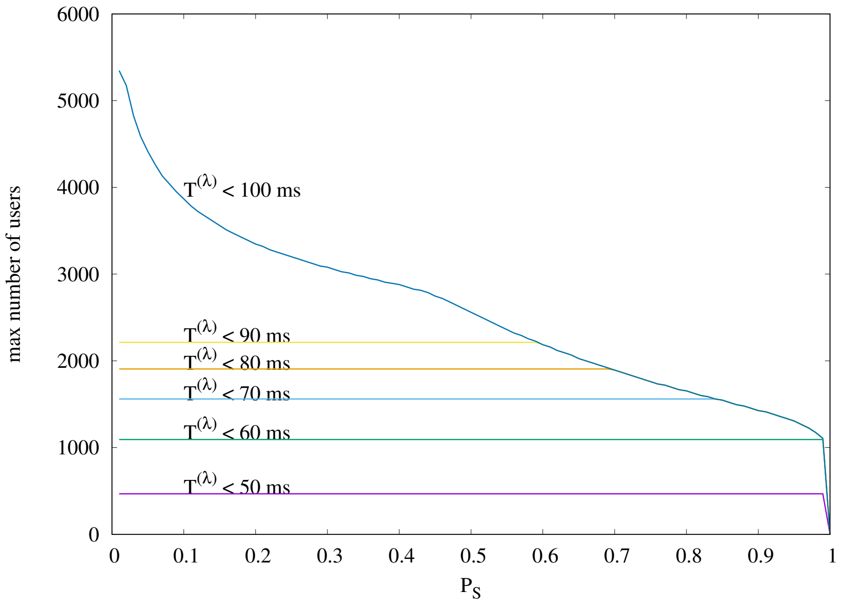

Figure 4.9: Max number of MTDs that can be connected to a BS to guarantee success probability above a threshold and latency below a threshold

a massive and critical context, with high MTD density layouts in the order of tens or even hundreds of users per square meter, this would require deploying ultra-dense BS sets, each BS covering just a few square meters. This is clearly undoable in SF layouts and calls for further technology enhancements, which are out of the scope of this thesis.

Impact of the secondary flow. Fig. 4.10 and Fig. 4.11 provide results for a scenario where there are flows of requests with different nature and requirements (i.e., time-critical and non-time-critical request flows). In particular, the x-axis of Fig. 4.10 represents the aggregate arrival rate of the two flows $(\lambda + \ell)$. The two flows have an equal rate, so that they obtain the same throughput, as long as the timeout probability P_{TO} is negligible. We can observe from a global perspective that the throughput has the same characteristics of the case with requests of only one type. However, by looking separately at the two flows we can observe that a decrease in $\xi^{(\lambda)}$ (due, for instance, to the effects of P_{TO}) favours the delay-tolerant traffic by increasing $\xi^{(\ell)}$.

For what concerns latency, Fig. 4.11 compares the latency distributions experienced by successful requests in the primary, time-critical flow, for various ratios λ/ℓ. The result is that the latency performance of the primary flow is barely dependent on the presence of flow ℓ, although it depends on the aggregate arrival rate. We can thereby conclude that regulating the secondary flow with ACB makes the primary flow experience priority when it comes to

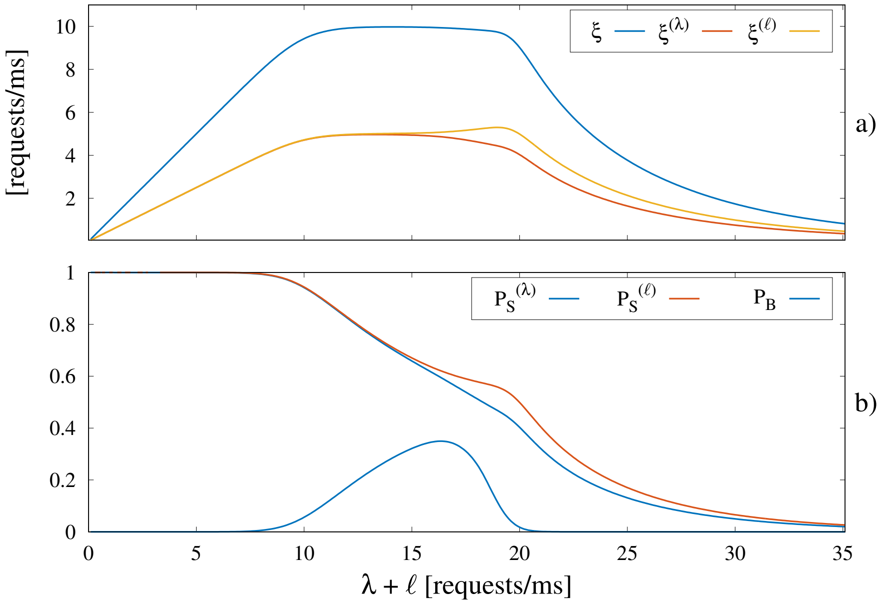

Figure 4.10: Flows with different priorities: *a)* system throughput, *b)* P_S and P_B

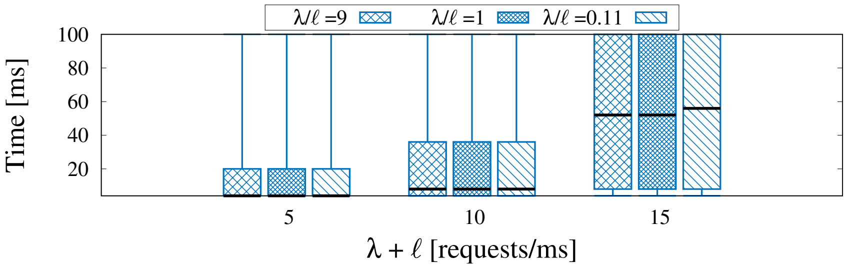

Figure 4.11: Primary flow delay distribution in case of a success in the system with multiple flows

latency guarantees.

We have evaluated the impact of other parameters, although we cannot show those results due to lack of space. The experiments reported here suffice to illustrate the main system features, spot desirable operational points and identify intrinsic limitations in radio access procedures used in 4G/5G networks.

4.6 Related Work

All forecasts predict that the next generation of cellular networks will support, in addition to traditional services, a wide variety of Machine-to-Machine (M2M) services, in the context of the IoT and massive MTC.

The authors of [73] outline the impact of massive M2M communications, and their coexistence with traditional services, on future networks. The paper and its references analyse the issues arising when a high load of M2M traffic must be served, and identify network access mechanisms as possible bottlenecks that may degrade the system performance.

Other investigations study the access mechanisms in LTE and in 5G networks in the case of M2M communications. Examples of such works are, for instance, [1, 74, 75]. All these papers include the performance modeling and analysis of the network access procedures for LTE and 5G, but, although they include many protocol features, (in general) they only focus on access mechanisms, without accounting for blocking at the BS, and for the reciprocal effects of blocking between access mechanisms and BS. The complex interactions of these two different bottlenecks have been highlighted in the case of massive access by using a measurement-based approach [39] and analysis [76].

There exist also some recent studies on enhancing the random access procedure, e.g., by using ACB with power control, thus exploiting the so-called *c*apture effect to partially solve the RACH collision problem [77], or by resolving collisions in the RACH transmissions instead of avoiding them [78]. Such approaches alleviate yet do not solve the problem of massive MTC scenarios like SF, in which the RRC connect phase can fail with non-negligible probability and cause unacceptable latencies due to multiple access retries.

Authors in [79] introduced a performance model for evaluating M2M communications in heterogeneous settings. This model has been used to study the coexistence between M2M and human-to-human communications in the same networks and for evaluating energy saving strategies.

4.7 Conclusions

We have presented and validated a simple, yet accurate, model for the performance analysis and design of cellular networks in smart factory environments characterised by machine-type communications, including the massive and/or mission-critical cases. The model captures many aspects of the dynamics in a cell, such as the different phases of the access procedure, the possible contention preamble collisions and the limited number of uplink grants in the random access response message, the limited number of retrials, the coexistence of different types of traffic (real-time and non-real-time), the use of a timeout for real-time traffic, and the prioritization of different types of traffic flows (e.g., with the ACB technique).

The main merit of the model lies in the valuable insight that it brings on cellular system operations and in the possibility to use it to drive the correct dimensioning of the cellular system in smart factory scenarios. The model also unveils some intrinsic limitations of the class of random access procedures adopted in cellular networks, and can be instrumental for the design of more effective algorithms.

Chapter 5

Data Service Phase: a novel formulation

5.1 Introduction

With the always increasing demand of capacity, the success of network operators depends on the performance of key critical services like the ones offered by dedicated apps for social media, banking/financial transactions, home and factory automation, health monitoring, etc., which are being increasingly ported to smartphones [80, 81]. For such services, as well as for the emerging Internet of Things (IoT) scenario [82], throughput is not necessarily the key metric, while outage (or “blocking probability”) and access delay become fundamental [83], as it is commonly observed during crowded events or when an emergency situation arises [84]. Indeed, those performance figures are affected by the number of devices requesting network access in a cell [85, 86]. Thus, such number risks to become the new network bottleneck, as we discuss in this chapter.

Access to resources in cellular network is basically the result of a two-step operation, which consists in first notifying the network about the user intention to join the network over a random access channel, and second, if the random access request was acknowledged by the base station, in negotiating a data channel [72]. Existing models focus on achievable throughput, and limit the analysis of resource access protocols to the random access operation in legacy cellular networks [85, 87] or in newer cells which support enhanced protocols for machine-type communications or IoT [78,86]. By so doing, they neglect the blocking probability for the data channel negotiation, which is non-negligible when the number of data channels requested by users becomes comparable with the limited number of connections that a base station can handle [72]. A few other models follow the details of signaling channel operation, but do not work in heavily utilized cells, in which the finite amount of resources of the network becomes the bottleneck [1].

In this chapter, we show that the data channel negotiation and utilization can be modeled independently of the random access, and therefore our analysis complements and completes existing studies on the performance of cellular networks. Specifically, we derive a chain of models for the system under study. We show that a queuing network with non-Bernoulli routing is needed to characterize the system in detail, which is not practical and does not lead to convenient analytical expressions. However, with a few approximations validated through simulation, we show that the system can also be modeled with a closed queueing network that admits a product-form solution which is independent of the distribution of service times. From this product-form queuing network we are able to derive closed form expressions for the blocking probability, the throughput and the network service time with multiple users competing for data transmission resources. In addition, we derive convenient recursive expressions for all proposed metrics, which are extremely useful to implement the formulas numerically, and avoid multiplications and divisions with extremely high or extremely low factors. We validate our model against simulation and compare the results with simpler approximations based on well-known approaches such as the Erlang-B formula. Our results show that our closed form expressions are very accurate and outperform alternative methods in terms of robustness, complexity and precision.

5.2 System

Access to service in a cellular network like LTE/LTE-A is based on a random access procedure followed by a signaling exchange needed to synchronize base station and mobile device and assign resources to a connection over which data can be exchanged. The latter is the so-called RRC (Radio Resource Control) connect phase.

This is a general scheme adopted by 3GPP since the early days of cellular data networks, which will be also included in future LTE and 5G releases [72]. Fig. 5.1 offers a high-level view of the cellular system from the point of view of access to transmission resources. The system is characterized by a few blocks that can be analyzed independently. The activity of customers creates a feedback loop between some of such blocks (represented with dashed lines and shaded blocks in the figure), so that a detailed analysis of the behaviour of the system cannot be generalized, since it depends on how connection requests are generated and on how customers return to ask for new transmission resources after a period of inactivity or after a failed attempt.

The random access phase resembles a multi-channel slotted Aloha sys-

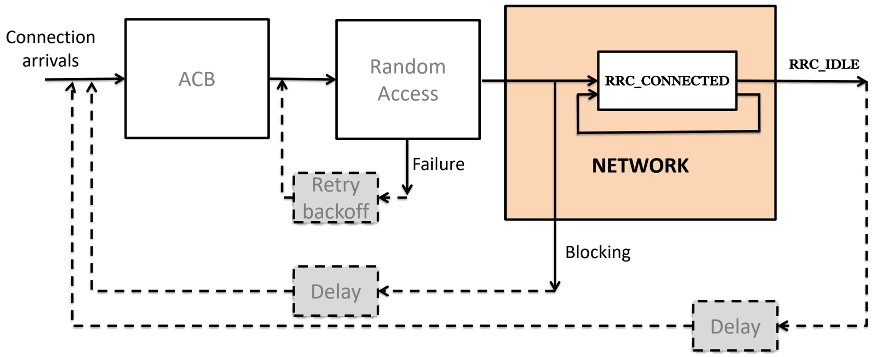

Figure 5.1: High-level system view. Dotted lines and grey-shaded delay/back-off blocks indicate optional system components. Such optional components, jointly with ACB and Random Access, have been widely addressed by the scientific literature and are out of the scope of this work. The response of the NETWORK block to the load offered by the Random Access is analyzed in this chapter, and characterized in terms blocking probability and service offered to the end users to which the base station allots cellular resources.

tem [44] with, optionally, access class barring (ACB [70]). ACB makes access requests a non-persistent process, meaning that a customer ready to request access can be forced to defer its request with some probability. A slot in such system is the interval of time between two random access opportunities (RAOs), regularly announced by the base station. Note that, since in LTE and similar networks, resources are split over time into units called subframes, a RAO includes one or more subframes.

Once the customer attempts random access and succeeds, the following RRC connect phase consists of two parts: access resolution and connection establishment. During access resolution, customers are in `RRC_IDLE` state, and have no resource allocated to transmit and receive data. The ones that have passed the random access phase are then requested by the base station to specify the parameters of their connection request, i.e., they are invited by the base station to proceed with the connection request. More specifically, 3GPP standards specify that in the random access phase, customers announce their willingness to access resources by sending an anonymous orthogonal code during the RAO, picked from a restricted dictionary. If two or more customers pick the same code, the base station decodes it only once and asks whoever has sent that code to specify its connection request parameters in a given set of time-frequency resources. At that point, if two or more customers have used the same code, their requests collide and have to wait for one of the next RAOs to attempt access again. From a logical point of view, the random access phase effectively concludes with the access resolution. This procedure is equivalent to excluding from access grant requests all customers that have used the same orthogonal code during the random access phase, so that random access and access resolution can be jointly analyzed as a regular multi-channel slotted Aloha system. The performance of such system has been widely studied in the literature, so in this chapter we focus on the connection establishment part, which happens in the subsystem indicated as NETWORK in Fig.5.1, and which has been so far oversimplified or completely neglected in the evaluation of cellular performance.

The base station can only decode and acknowledge a limited number of requests per subframe and can only handle a limited number of simultaneous connections. So, only a fraction of non-collided requests involved in the connection establishment procedure receive resources and are promoted to the `RRC_CONNECTED` state [72]. No new connection request can be acknowledged when the limit of connections has been reached. 3GPP standards further specify that, once a connection is established, it remains active until a timeout expires after service completion. After that, it returns to the `RRC_IDLE` state. Therefore, recently served customers keep holding their resources for a short while (decided by the network operator, typically between 1 second and 1 minute) after completing their traffic exchange, which could prevent freshly arrived customers to obtain service even if the actual number of active connections is below the maximum.

In the operation of the system, customers that pass the random access phase are subject to a blocking probability because of the limitation in the number of `RRC_CONNECTED` customers. For what concerns the service rate received by customers, this is a complex function of the scheduler policy implemented at the base station, the quality of wireless links, and the number of active connections. For tractability reasons, here we assume that the base station adopts a processor sharing policy with a single class of users, so that resources are shared equally. Moreover, we assume that all customers can use the same modulation and coding scheme. Such assumptions are realistic for M2M communications and in small cell environments in general.

The system we have described is characterized by two kinds of arrivals: the *e*xogenous flow generated by the random access system and the *e*ndogenous flow generated by customers that return before their `RRC_CONNECTED` timeout expires. Similarly, customers leave the system for two reasons: if blocking occurs over the exogenous arrival flow, and if, after receiving service, customers do not generate traffic before the expiration of the `RRC_CONNECTED` timeout, so that they fall back to the `RRC_IDLE` state.

From the above description, we see that it is important to track both the number of active connections as well as the number of established yet inactive connections, since those values determine blocking probability and service rate. Next, we will show how to model the system and derive closed form expressions for the blocking probability suffered by the exogenous arrival flow and for other key performance indicators such as the average service rate and the throughput of the cellular system.

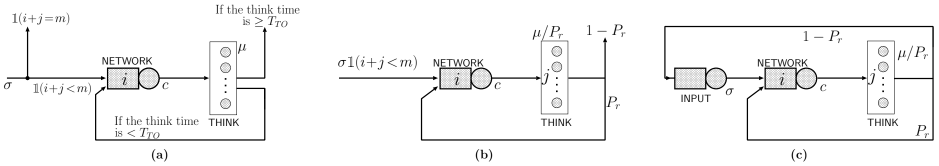

Figure 5.2: Queuing network models for the connection establishment process in a cellular network (NETWORK subsystem in Fig.5.1): network with non-Bernoulli routing (a), open network with population constraint (b), closed network (c)

5.3 A Chain of Queueing Network Models

The system that was described in the previous section can be abstracted into a model where requests for service are submitted by end user terminals. Requests can be refused access to network resources if those are not readily available, and otherwise enter into a service phase. At the end of service, resources remain associated to the end user for a time T_{TO}, before being released. If during this time T_{TO} the end user issues another request, he/she is immediately granted the same resources he/she had before. Otherwise, resources are freed, and a new request by the same user will be accepted conditionally on the availability of free resources in the network.

We can model this behavior with the network of queues in Fig. 5.2(a). Here, a Poisson process with rate σ feeds the queue *NETWORK*, which has one exponential server operating with rate c according to the processor sharing discipline. The assumptions of Poisson arrivals and exponential service times will be relaxed later on. Both σ and c are expressed in services/s. At the end of service, customers enter the infinite-server queue *THINK*, with exponential service at rate μ, expressed in s^{-1} units.

When service at queue *THINK* ends, customers are routed back to queue *NETWORK* if the instance of their service time was shorter than T_{TO}, and they leave the network of queues otherwise (with a service time in *THINK* equal to T_{TO} or longer). At any given time instant, the number of customers at queue *NETWORK* is denoted by i, and the number of customers in queue *THINK* whose received service is shorter than T_{TO} is denoted by j (j thus is *not* the total number of customers at queue *THINK*). Arriving customers are lost when the sum of the numbers i and j is equal to m.

It is important to observe that in the network of queues we just described the state depends on accrued service times, and the customer routing is not governed by Bernoulli choices, but depends on service time instances; hence, this model cannot be solved by using the classical queueing network solution methods (e.g., Markovian analysis, or product-form solution).

We can approximate the behavior of the queuing network model we just described with another queueing network, as in Fig. 5.2(b). In this second model, the infinite-server queue *THINK* has exponential service at rate μ, truncated at T_{TO}. That is, the pdf of the service time at queue *THINK* is

$$f_{T_{\mathrm{THINK}}}(t) = \mu e^{-\mu t}[u(t) - u(t - T_{TO})] + (1 - P_r)\delta(t - T_{TO}) \tag{5.1}$$

which results in an average service time

$$E[T_{\mathrm{THINK}}] = P_r/\mu. \tag{5.2}$$

with

$$1 - P_r = e^{-\mu T_{TO}}, \tag{5.3}$$

where $u(\cdot)$ is the unit step function, and $\delta(\cdot)$ is the Dirac delta function.

In this new network of queues we introduce a Bernoulli choice for customer routing after the service at queue *THINK*, with a probability P_r of re-accessing the network resources, which is equal to the probability of the service at queue *THINK* ending before time T_{TO}.

Arriving customers are lost when the sum of the numbers of customers at queues *NETWORK* (denoted by i) and *THINK* (j) is equal to m.

This modification in the model implies an approximation, which calls for a validation against simulation results, as we will do in a later section of this chapter.

Because of the fact that the total number of customers in the previous model is limited to m, we can transform our open queuing network into a closed queuing network, by adding one more single-server queue (named *INPUT*), with exponential service at rate σ and processor sharing discipline, so as to obtain the model in Fig. 5.2(c). Note that this third queuing network model, with a total customer population of size m, is exactly equivalent to the previous one, and that the exponentially distributed service time of *THINK*, truncated at T_{TO} is still there.

This last model allows the exploitation of a classical queueing network result concerning service time distributions. In particular, the well known BCMP product-form theorem [88] states that the queueing network model of Fig. 5.2(c) has product-form solution whenever the service times of single server queues with processor sharing discipline, or of infinite server queues (the former is the case of queues *INPUT* and *NETWORK*; the latter is the case of *THINK*), has a rational Laplace transform. Probability distributions with rational Laplace transform form a very wide class that includes matrix exponential, coxian, and phase-type distributions (see for instance [89]). This translates into the fact that *any* distribution that can be expressed in phases (e.g., phase-type) is acceptable as a service time distribution in this queuing network while preserving the product-form solution. Hence, by playing with phase-type distributions we can model a very wide class of arrival processes and service time distributions, in either an exact or an approximate manner. For instance, a deterministic service time cannot be exactly represented by using a phase-type distribution, but it can be well approximated by using an Erlang distribution with a large number of phases. A similar thing can be done for a exponential delay, truncated at T_{TO}. As we will show when we address computational issues, the number of phases does not affect the solution complexity, since the performance measures we are interested in only require the specification of the average service times. For the model in Fig. 5.2(c) we can therefore just specify an average delay P_r/μ for the *THINK* queue, and a processor sharing discipline for the *NETWORK* and *INPUT* queues, with no loss of generality. In this manner we can model a very wide class of arrival processes and service time distributions.

For this queuing network model, we can define the state as

$$s_{i,j,k} = (i, j, m - i - j),$$

where i is the number of customers in queue *NETWORK*, j is the number of customers in queue *THINK*, and $k = m - i - j$ is the number of customers in queue *INPUT*. We can then compute the equilibrium probabilities as the product of the state probabilities of the three queues computed for unit arrival rate at the *NETWORK* queue (note that any arrival rate can be used and that the arrival rate at the *THINK* queue is the same as at the *NETWORK* queue, while queue *INPUT* sees $(1 - P_r)$ times the arrival rate of the other queues):

$$\pi_{(i,j,m-i-j)} = \frac{1}{G}\left(\frac{1}{c}\right)^i \frac{1}{j!}\left(\frac{P_r}{\mu}\right)^j \left(\frac{1-P_r}{\sigma}\right)^{m-i-j} \tag{5.4}$$

where G is a normalization constant. In the following, to emphasize the dependency of the normalization constant G with respect to parameter m we will use the notation $G(m)$.

Recursive Equations for Performance Metrics

Dealing with a queuing network that admits a product-form solution allows the use of several different computational algorithms developed for this class of models. In particular, we can apply the extremely effective algorithm known as Mean Value Analysis (MVA) [90], that allows the computation of performance measures through a set of recursive equations. In the following, we provide the MVA equations for the queueing network depicted in Fig. 5.2(c) where m customers cyclically visit the three queues.

For convenience, we label the three queues as $I = 1$, $N = 2$ and $T = 3$, respectively for *INPUT*, *NETWORK* and *THINK*, and in the rest of the chapter we interchangeably use letters and numbers as indexes, depending on the context (e.g., ξ_2 and ξ_N will both refer to the throughput of *NETWORK*). The parameters $S_i, i \in \{1,2\}$, represent the inverse of the service rates for queues *INPUT* and *NETWORK*, whereas Z_3 represents the average delay for the infinite-server queue *THINK* (e.g., $S_1 = 1/\sigma$, $S_2 = 1/c$, and $Z_3 = P_r/\mu$).

We denote by $\mathbf{P}$ the customer routing matrix, where the element $p_{i,j}$ represents the probability that a customer completing service at queue i moves next to queue j for service. From $\mathbf{P}$ we can compute the visit ratio vector $\mathbf{v}$ as $\mathbf{v} = \mathbf{vP}$. The element V_i of $\mathbf{v}$ represents the average relative number of visits of a customer to queue i, and does not depend on the population size.

The performance indexes which summarize the behavior of the queueing network at steady-state are, with population m:

- the queue *utilization* $U_i(m)$, i.e., the fraction of time in which the server of queue i is busy;
- the cumulative *mean sojourn time* $W_i(m)$ experienced by a customer while waiting and subsequently receiving service at queue i;
- the average number of customers (waiting and in service) at a queue, $Q_i(m)$;
- the system throughput $\xi(m)$, and the queue throughput $\xi_i(m)$ (with $\xi_i(m) = V_i\xi(m)$).

These quantities, for $i = 1, 2$, can be computed as follows:

$$
\begin{aligned}
W_i(m) &= S_iV_i(1 + Q_i(m-1)), && (5.5)\\
\xi(m) &= \frac{m}{Z_3V_3 + \sum_{i=1}^{2} S_iV_i(1 + Q_i(m-1))} && (5.6)\\
U_i(m) &= S_iV_i\xi(m), && (5.7)\\
Q_i(m) &= U_i(m)(1 + Q_i(m-1)), && (5.8)
\end{aligned}
$$

The above MVA formulas define a recursive algorithm in terms of the queue lengths $Q_i(m-1)$. The recursion starts from $m = 0$. In particular, we start with $Q_i(0) = 0$, from this we can derive $W_i(1) = S_iV_i$, and then we can derive $\xi(1)$, $U_i(1)$, and $Q_i(1)$. In this manner we can derive the performance indexes for the queueing network with m customers in $O(m)$ steps.

5.4 Performance Metrics

In this section we introduce the set of metrics that characterize the performance of the system under study, and for these metrics we provide a definition in terms of the queueing network model of Fig. 5.2(c).

Blocking probability

The blocking probability at the base station under investigation is reflected in the probability that the sum of customers at queues *NETWORK* and *THINK* is equal to m (or, equivalently, the probability that queue *INPUT* is empty, which occurs when $i+j=m$). Therefore, the blocking probability of exogenous connection requests (proceeding from the Random Access block of Fig. 5.1) is:

$$P_b(m) = \sum_{j=0}^{m} \pi_{(m-j,j,0)} = 1 - U_I(m). \tag{5.9}$$

Note that the sum in the equation above yields a closed form expression for the blocking probability. Indeed, using (5.4) and normalizing the sum of probabilities, it is easy to compute the normalization constant $G(m)$, and hence also derive the blocking probability in closed form, as follows:

$$P_b(m) = \frac{\sum_{j=0}^{m} \frac{1}{j!} \left(\frac{\sigma P_r}{\mu(1-Pr)}\right)^j \left(\frac{\sigma}{c(1-Pr)}\right)^{m-j}}{\sum_{j=0}^{m} \frac{1}{j!} \left(\frac{\sigma P_r}{\mu(1-Pr)}\right)^j \sum_{i=0}^{m-j} \left(\frac{\sigma}{c(1-Pr)}\right)^i}. \tag{5.10}$$

Throughput

The throughput of the base station corresponds to the throughput of queue *NETWORK* (denoted as $\xi_N(m)$), which is equal to the throughput of queue *THINK* (denoted as $\xi_T(m)$), while the throughput of queue *INPUT* is $\xi_I(m) = (1-P_r)\,\xi_N(m)$. It is thus enough to find an expression for the throughput of the *NETWORK* queue, which can be easily obtained by considering that the entire flow entering the queue is served, so that $\sigma(1-P_b(m)) + P_r\xi_N(m) = \xi_N(m)$. Therefore, the following result holds:

$$\xi_N(m) = \sigma \frac{1-P_b(m)}{1-P_r}. \tag{5.11}$$

Average service time

The average time required to serve a request can be mapped into the average time spent at queue *NETWORK*. By using Little's law, we can derive this average time as:

$$W_N(m) = Q_N(m)/\xi_N(m). \tag{5.12}$$

Note that, since the *NETWORK* queue uses a processor sharing policy with no waiting room, the average time spent in the queue can be (equivalently) derived as

$$W_N(m) = \frac{1}{c}\frac{\sum_{i=1}^{m} iP_i(m)}{1-P_0(m)}, \tag{5.13}$$

where $P_i(m)$ is the probability to have i customers under service (for $i = 0, \ldots, m$). Eq. (5.13) simply uses the fact that resources are equally split among active customers, subject to the fact that at least one user is under service. If only a user is under service, $\frac{1}{c}$ is the average service time.

5.5 Recursive Expressions and Approximations

Recursive computation of the blocking probability

Let $A_k(m)$ be the probability of having k customers in the *INPUT* queue when the population of the closed queueing network is m. When the *INPUT* queue is empty, a blocking occurs in the cellular network. From the *A*rrival Theorem [91], we know that

$$A_1(m) = \frac{\xi_I(m)}{\sigma} A_0(m-1), \tag{5.14}$$

By simple inspection of the closed queueing network, the following relations hold:

$$\begin{align} \xi_I(m) &= \sigma(1 - P_b(m)); \tag{5.15} \\ A_0(m-1) &= P_b(m-1); \tag{5.16} \\ \Rightarrow A_1(m) &= (1 - P_b(m))P_b(m-1). \tag{5.17} \end{align}$$

We now look for a relation between $A_1(m)$ and $A_0(m)$. Using the expressions for state probabilities, we obtain the following expression, for $0 \leq k \leq m$:

$$\begin{align} A_k(m) &= \sum_{j=0}^{m-k} \pi_{(m-j-k,j,k)} \\ &= \frac{1}{G(m)} \sum_{j=0}^{m-k} \frac{1}{j!} \left(\frac{\sigma P_r}{\mu(1-Pr)} \right)^j \left(\frac{\sigma}{c(1-Pr)} \right)^{m-j-k}. \tag{5.18} \end{align}$$

From the above expression, evaluated for $k \in \{0, 1\}$, and having identified that $A_0(m) = P_b(m)$, the following relation holds:

$$A_1(m) = \frac{c(1-Pr)}{\sigma} \left(P_b(m) - \frac{1}{G(m)} \frac{1}{m!} \left(\frac{\sigma P_r}{\mu(1-Pr)} \right)^m \right), \tag{5.19}$$

which, once plugged in (5.17), leads to the following recursive expression for the computation of P_b:

$$P_b(m) = \frac{\frac{\alpha(m)}{G(m)} \frac{c(1-P_r)}{\sigma} + P_b(m-1)}{\frac{c(1-P_r)}{\sigma} + P_b(m-1)}, \tag{5.20}$$

where $\alpha(m) = \frac{1}{m!}\left(\frac{\sigma P_r}{\mu(1-P_r)}\right)^m$ can be computed recursively as

$$\alpha(m) = \frac{\sigma P_r}{m\mu(1-Pr)}\alpha(m-1). \tag{5.21}$$

Note that $\frac{\alpha(k)}{G(k)} \in [0,1]$ for all positive values of k, so that the recursive expression of P_b always yields a value between 0 and 1. Moreover, there exists an interesting recursive expression for the computation of $G(m)$ to be used in (5.20):

$$G(m)=G(m-1)+\sum_{j=0}^{m}\frac{1}{j!}\left(\frac{\sigma P_r}{\mu(1-P_r)}\right)^j\left(\frac{\sigma}{c(1-P_r)}\right)^{m-j}.$$

Therefore, $P_b(m)$ can be computed recursively starting with the following initialization: $P_b(0) = 1$, $\alpha(0) = 1$, $G(0) = 1$.

Recursive computation of the service time

The time to serve a request corresponds to the sojourn time in queue *NETWORK*. Eq. (5.13) holds for the *NETWORK* queue with at most m services in progess. However, note that the average number of services in progress, given that at least one customer is under service, is $\frac{\sum_{i=1}^{m} iP_i(m)}{1-P_0(m)}$. This is also given by 1 service in progress plus the average number of services in progress in a system with $m-1$ possible services, that is $1+\sum_{i=1}^{m-1} iP_i(m-1)$. This result holds because the *A*rrival Theorem holds for closed queueing networks with product-form solution [90]. Applying this equality repeatedly leads to the following result:

$$W_N(m) = \frac{1}{c}\sum_{i=0}^{m-1}\prod_{j=1}^{i}\left[1-P_0(m-j)\right]. \tag{5.22}$$

Eq. (5.22) offers the possibility to compute the average time spent at queue *NETWORK* recursively, by analysing systems with at most $m-1$ parallel services, which can be useful for numerical implementations.

Relation between blocking probability and other metrics

There is a simple relation between the probability that the *NETWORK* queue is inactive and the blocking probability, denoted here as $P_0(m)$ and $P_b(m)$, respectively. Therefore, computing $P_b(m)$ is enough to compute also the time spent in *NETWORK*. Specifically, the following result holds:

$$\xi_N(m) = [1 - P_0(m)]c; \tag{5.23}$$

$$\xi_N(m) = \sigma[1 - P_b(m)] + P_r \xi_N(m); \tag{5.24}$$

$$\begin{aligned}\Rightarrow P_b(m) &= 1 - \frac{\xi_N(m)\,(1 - P_r)}{\sigma} \\ &= 1 - \frac{c}{\sigma}(1 - P_r)[1 - P_0(m)]. \end{aligned} \tag{5.25}$$

Moreover, from (5.22), the service time with at most m parallel services can be computed from $P_0(k)$ obtained for $k = 1, 2, \ldots, m-1$, as follows:

$$W_N(m) = \frac{1}{c} \sum_{i=0}^{m-1} \beta(i, m); \tag{5.26}$$

with $\beta(0, m) = 1$ and $\beta(i, m) = \beta(i-1, m)(1 - P_0(m - i))$.

Thus, to characterize the system it is enough to compute the set $\{P_b(k)\}_{0 \le k \le m}$

Approximations

Next, we present a few closed form and iterative approaches to the analysis of the system, based on some simplifications and on the use of well-known formulas, commonly used in cellular system performance analysis and design. The results of these approaches will be compared to those of our model and of simulation in the next section.

Erlang B in closed form. We start with an approximation using the Erlang B formula for an $M/M/k/0$ system with arrival rate σ and $k = \lfloor m(1 - P_r) \rfloor$ servers. In this case, customers are served in parallel with speed c/m each, and the blocking probability can be expressed in closed form as follows:

$$P_b = \text{ErlangB}\left(\frac{\sigma}{c/m}, \lfloor m(1 - P_r) \rfloor\right). \tag{5.27}$$

By using this formula we assume that about mP_r servers are always busy due to returning customers, and service to exogenous requests can be provided by the remaining $m(1 - P_r)$ servers with fixed rate (taken to be the same as the lowest individual service rate that can be experienced at queue *NETWORK*).

Iterative Erlang B. A second approximation based on an $M/M/k/0$ queue can be devised by choosing an arrival rate equal to $\sigma + P_r \xi_N$, and $k = m$ servers. Like before, each server runs at speed c/m. In this case, we account for the maximum number of services in progress, and for both exogenous and endogenous request flows.

This case requires iterations because the queue input and output are coupled by a feedback mechanism. Indeed, note that $\xi = \sigma \frac{1 - P_b}{1 - P_r}$, so that an iterative expression for the loss probability is as follows:

$$P_b = \text{ErlangB}\left(\frac{\sigma}{c/m} \frac{P_r}{1 - P_r}(1 - P_b), m\right). \tag{5.28}$$

Note also that the blocking probability is here approximated as the ratio between throughput and *a*ll arrivals, including returning customers, which in a real system cannot experience blocking.

$M/M/1/k$ **in closed form.** In this case, we approximate the system behavior as an M/M/1 queue with service rate equal to c, that can accept no more than k customers (in service or waiting). The limit on the number of customers is set as $k = \lfloor m(1 - P_r) \rfloor$. The blocking probability is:

$$P_b = \frac{1 - \frac{c}{\sigma}}{1 - \left(\frac{c}{\sigma}\right)^{\lfloor m(1-P_r) \rfloor + 1}}. \tag{5.29}$$

Iterative $M/M/1/k$**.** This case is similar to the second one, with the loss probability formula of an $M/M/1/k$ system with processor speed equal to c and with $k = m$ servers instead of the Erlang B formula with the same number of serves and speed c/m. We use $\xi_N = \sigma \frac{1-P_b}{1-P_r}$ to obtain the following iterative expression:

$$P_b = \frac{1 - \frac{c}{\sigma\left(1+\frac{P_r}{1-P_r}(1-P_b)\right)}}{1 - \left(\frac{c}{\sigma\left(1+\frac{P_r}{1-P_r}(1-P_b)\right)}\right)^{m+1}}. \tag{5.30}$$

5.6 Numerical results

We use a home-grown packet simulator written in python to validate our model and to assess the impact of our main approximations, namely: (*i*) the use of Bernoulli routing to make the queueing network of Fig. 5.2(a) tractable, and (*ii*) the replacement of load dependent service times in the NETWORK station with an exponential distribution with the same average (see Fig. 5.2).

From the expressions derived in Section 5.4, one can notice that blocking probability and throughput, as well as the time spent in *NETWORK*, depend on various parameters, namely the maximum number of services in progress m, the exogenous arrival rate σ, the service rate c, and the pair (μ, P_r) characterizing the *THINK* queue. Note that the parameters (μ, P_r) are equivalent to the pair $(1/\mu, T_{TO})$, that is the average time spent in *THINK* and the `RRC_CONNECTED` timeout. Moreover, results depend on the ratio σ/c rather than on their individual values. Such ratio represents the load offered by the exogenous arrivals only. In the following, we explore how such parameters affect system performance.

Validation and basic results on blocking and throughput. In the system under evaluation, blocking probability and throughput are not equivalent metrics, because of the presence of returning customers that cannot experience blocking. Therefore, we validate by simulation both quantities.

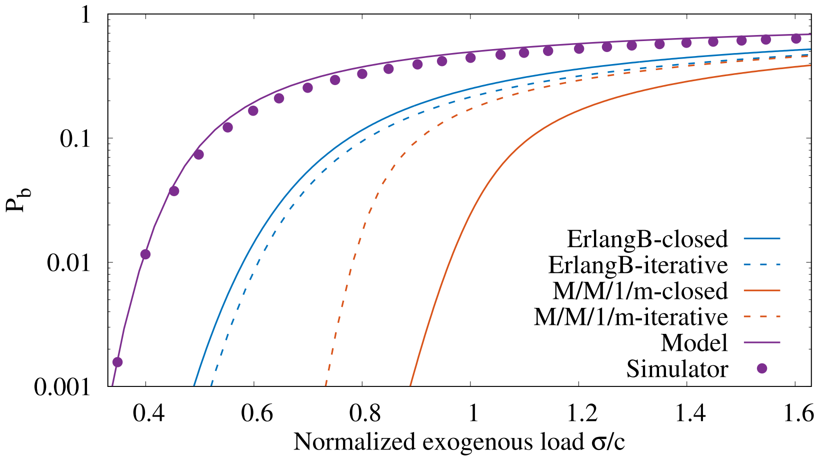

Figure 5.3: Blocking probability computed with our model and with other alternative models, compared to simulation estimates ($m = 50$, $P_r = 0.2$). With small values of m, models that do not account for the presence of the *THINK* queue have low accuracy.

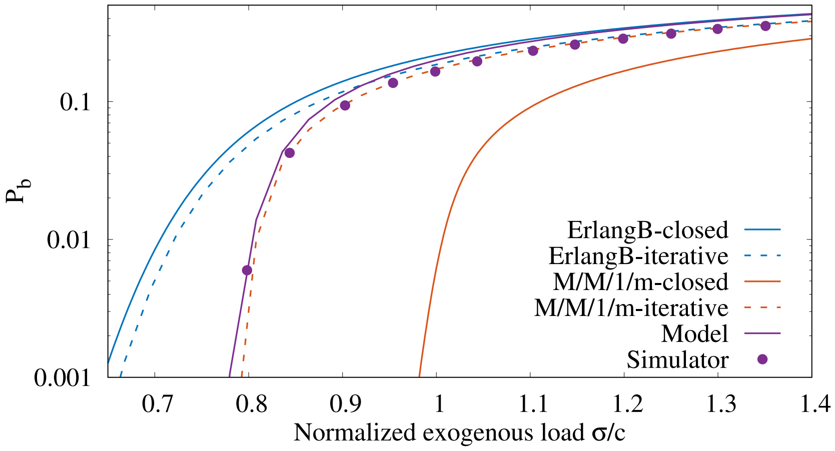

Figure 5.4: Blocking probability computed with our model and with other alternative models, compared to simulation estimates ($m = 200$, $P_r = 0.2$). With high values of m, modeling the *THINK* queue becomes less important, and iterative methods become as good as our model.

In Fig. 5.3 we show the blocking probability of a system with $m = 50$, $1/\mu = 30$ s and $P_r = 0.2$. These values have been selected as representative of reasonable operational conditions of a cellular network in which users wait on average 30 s before issuing a new request after a service is completed (e.g., the time needed to read a web page) and the operator allows maximum 50 users connected (this is a relatively small cell) and set the RRC timeout to $\sim$ 6s (values used by the operators vary from a few seconds to a few tens of

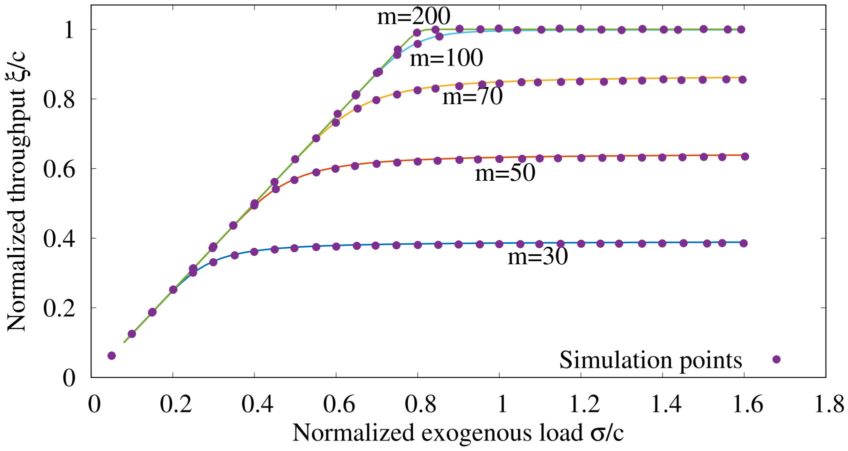

Figure 5.5: System throughput (i.e., throughput of queue *NETWORK*), relative to the queue capacity c, for $\mu^{-1} = 30$ s and $P_r = 0.2$. The system throughput is monotonic in the exogenous arrival rate, and saturates to a value that depends on the maximum number of simultaneous services (the capacity of the queue is constant)

seconds).

In the figure we compare model and simulation, which are quite close. We also report the results computed with the approximate models presented in Section 5.5, which behave quite poorly in all cases. The approximate models do not account for the presence of the infinite-server queue, that affects how and when customers return to service within the `RRC_CONNECTED` timeout. However, for large values of m, as we will comment later, larger fractions of the system population move to the *NETWORK* queue, and the presence of *THINK* becomes less important. Indeed, Fig. 5.4 shows that with $m = 200$ (which is for a large or dense cell, as today's base stations allow ~ 100 `RRC_CONNECTED` users) the approximate models can yield results comparable to our model, especially when using iterative methods. We remark that our model and simulation are very close under all parameter configurations, and that the lower-end of the blocking probability curves can be approximated only with our closed-form model or with iterative methods. Using the latter is less convenient in terms of complexity and because they cannot be used, e.g., to analytically tune the system.

Fig. 5.5 confirms that increasing m without changing μ and P_r has a beneficial impact on throughput, because more customers utilize the server of queue *NETWORK*. The figure also shows that the throughput cannot reach 100% of the *NETWORK* capacity c, unless a high number of parallel services are allowed.

Time spent at queue *NETWORK*. Fig. 5.6 displays the average time spent by customers in the system to complete service of their request, normal-

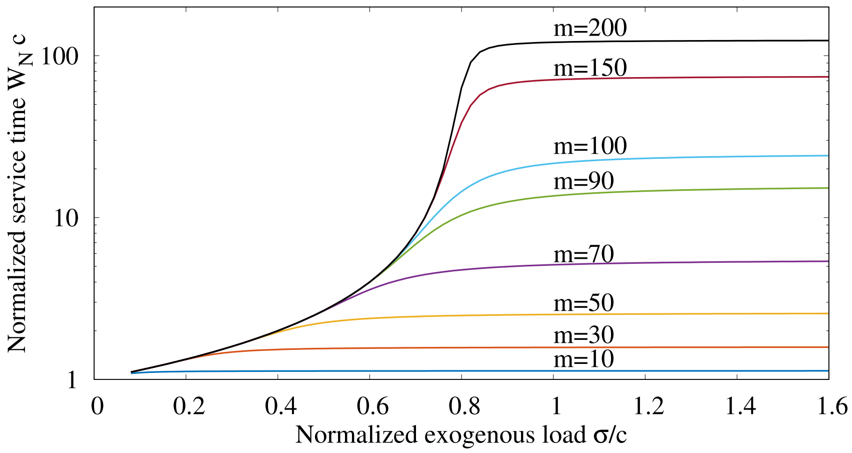

Figure 5.6: The average time spent at queue *NETWORK* increases with the exogenous request arrival rate ($\mu^{-1} = 30$ s, $P_r = 0.2$)

ized to the average service time of a request when served with all *NETWORK* resources, i.e., normalized to $1/c$. Since *NETWORK* uses a processor sharing discipline, the plotted values also represent the average number of simultaneously served requests. With $1/\mu = 30$ s and $P_r = 0.2$, the figure shows that the time necessary to complete a service saturates before the exogenous load reaches 100%, which is due to the presence of recently served customers with allocated resources. In the example, 20% of served customers return to service before the timeout expiration. Meanwhile, queue *NETWORK* keeps service positions ready for them, except they are unused, therefore imposing a restriction on the number of ongoing parallel services. For instance, with a load that saturates the throughput and m sufficiently high to reach the capacity c, using Little's law, the number of customers in *THINK* is $\sim c \cdot (P_r/\mu)$. With $m = 200$ and $1/c = 80$ ms (the time needed to serve a file of 1.5 MB with a 150 Mb/s downlink connection), the number of customers in `RRC_CONNECTED` status with no ongoing transmission is 75, i.e., 37.5% of m.

Impact of m. The importance of m is detailed in Fig. 5.7 for various values of the exogenous load and P_r, with $1/\mu = 30$ s. Increasing m always reduces the blocking probability, especially at high loads, although the price to pay for this improvement is a higher service time (see Fig. 5.6). That is, increasing m allows the system to keep more users busy, although for longer intervals, and reduces the blocking probability experienced by new freshly arrived customers.

Impact of P_r. The way customers returning before the RRC timeout affect blocking probability and throughput is detailed in Fig. 5.8 and Fig. 5.9, respectively. Returning customers have a very negative impact on blocking probability, which is exactly the reason why we cannot analyze a cellular system without considering the RRC timeout and the returning customers,

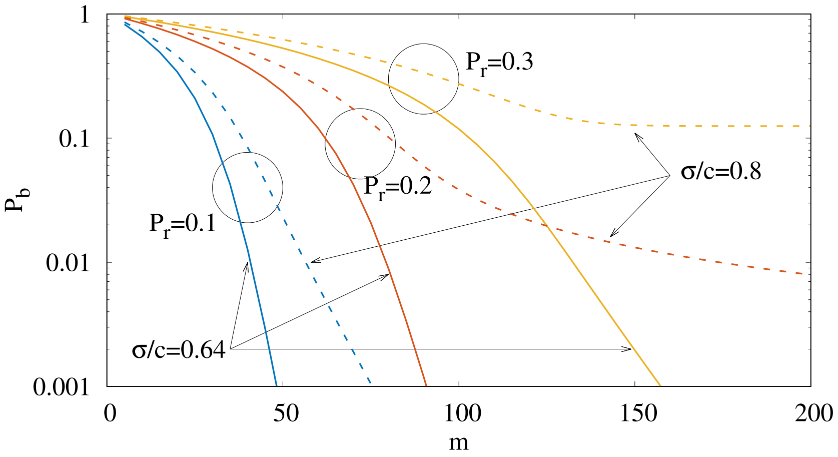

Figure 5.7: Impact of m on blocking probability, with $\mu^{-1} = 30$ s

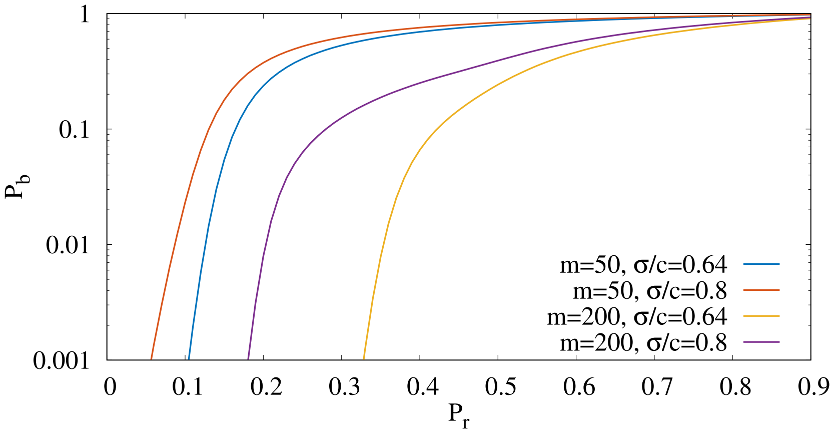

Figure 5.8: Impact of returning customers on blocking probability, with $\mu^{-1} =$ 30 s: high return rates worsen blocking

and the reason why operators need to finely tune the timeout values they use in their base stations. In terms of throughput, it is interesting to notice how reaching capacity c not only depends on m, but also on the fraction of returning customers. Indeed, a very high rate of returning customers heavily degrades the network performance, and the throughput collapses. The optimal value of P_r is hard to figure out, since it depends on other parameters such as m and the exogenous load. The figure shows that a network operator should adjust P_r (i.e., the RRC timeout) depending on cell parameters (m and c) and based on the traffic characteristics (σ and μ).

Relative importance of RRC timeout duration. To better understand the role of the RRC timeout T_{TO}, and considering that P_r is affected

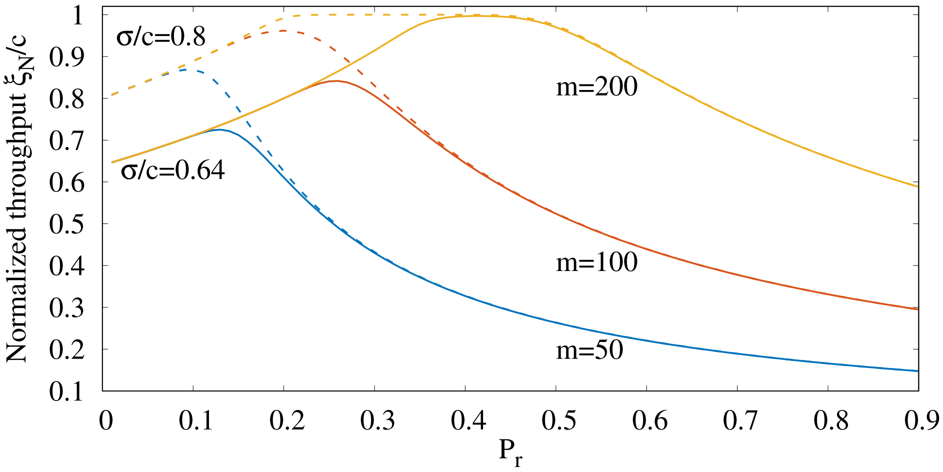

Figure 5.9: Impact of returning customers on throughput: high return rates induce throughput drops ($\mu^{-1} = 30$ s)

by the traffic characteristics through μ, we now plot the normalized network throughput as a function of both $1/\mu$ and T_{TO}.

Fig. 5.10 generalizes the analysis of the behavior of the curves of Fig. 5.9 for the case of intermediate offered loads ($\sigma/c = 0.64$) and $m = 200$. From the figure, we can observe that the base station capacity can be reached for either some optimal values of the timeout, or for very low values of $1/\mu$. In fact, low think times make improbable the return of customers without having to request a new connection through the random access procedure. As shown in Fig. 5.11, for higher loads ($\sigma/c = 0.8$), the impact of T_{TO} is very relevant, as it causes throughput drops at different values, depending on traffic conditions. However, both 3D plots show that for reasonably low average think times (e.g., below $15 - 20$ s), the choice of T_{TO} is much less critical.

These results can be used not only to design a network control mechanism to dynamically steer network configuration to follow user's demand, but also to support the design of machine-type communications in smart factories, in which the think time and the offered load can be imposed rather than estimated, and followed by the network.

5.7 Related work

Outage probability has always been a key performance index in wired and wireless networks, starting with the early days of telephony. Its importance is becoming more and more evident, with the increased dependence of users on their smartphone services. This is why availability is considered one of the key performance indicators for LTE and 5G [86], even if these new technologies

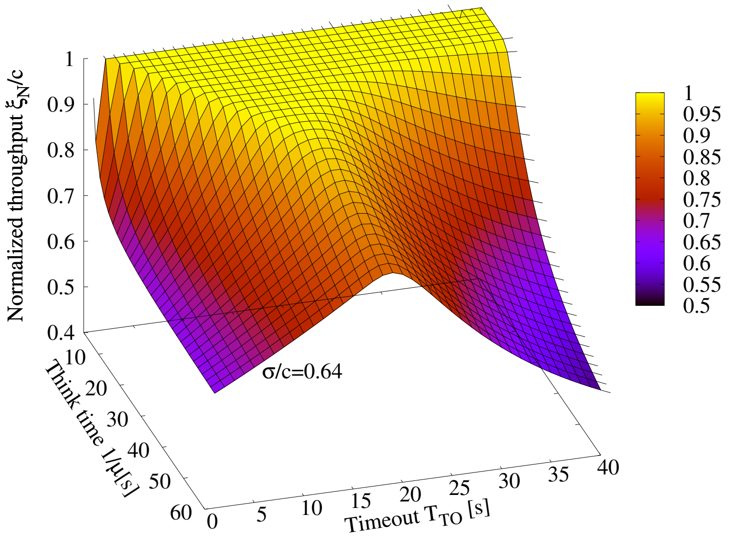

Figure 5.10: Impact of timeout and think time, with $\sigma/c = 0.64$ and $m = 200$

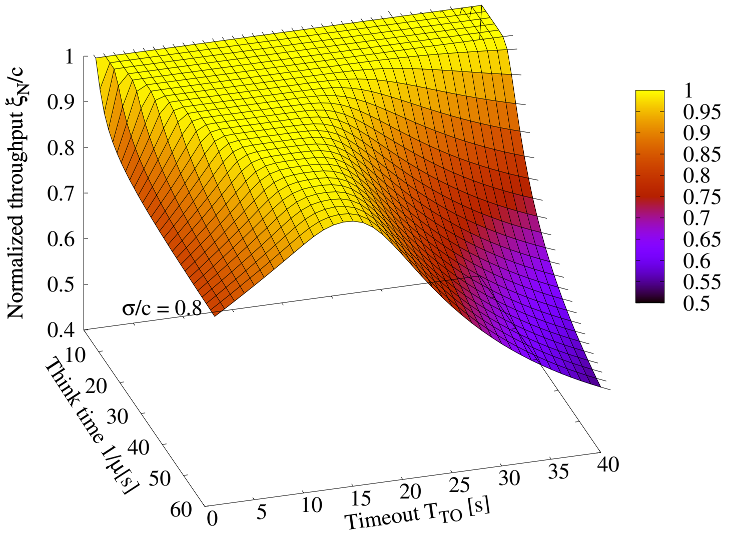

Figure 5.11: Impact of timeout and think time, with $\sigma/c = 0.8$ and $m = 200$

provide much higher capacity and widespread coverage.

In recent works on the performance of cellular networks, such as [92], blocking probability becomes the focus of the analysis, relating outage probability to both cell size and data rates. In addition, this chapter also analyses energy efficiency with respect to blocking probability instead of the more classical throughput.

In LTE networks, the concept of outage probability goes along with the reduction of latency: resources have to be made available to end users in the shortest possible time, even in extremely crowded environments. Guaranteeing high capacity, low latency and service availability are the challenges for next generations of cellular networks. In [84] a detailed evaluation based on real measurements of a popular public event is performed. Measurement results show that the number of dropped connections during such big events increases more than two orders of magnitude. Authors in [86] investigate more in detail the bottlenecks in the network, showing how performance drops in crowded scenarios are due to congestion of both Random Access resources and "network" resources. Again, in [84] and [86], outage probability is identified as one of the main roots of the problem.

Although the relevance of blocking probability has evolved together with cellular network standards, the way it can be computed in an easy and effective manner still remains a research challenge. The system behavior in the `RRC_CONNECTED` state has not been the subject of much research so far. The model we propose in this chapter aims at tackling this problem in a simple and effective way.

5.8 Conclusions

The performance of cellular networks is driven by a set of elements whose effects are rather complex to predict. Characterizing and understanding the behavior and performance of these different elements is a prerequisite for any sound quantitative optimization and management of these networks.

This chapter investigates the characteristics of one of the components of cellular networks which has been overlooked so far, i.e., the system behavior in the `RRC_CONNECTED` state. For this components we presented a simple, yet extremely accurate model, which can be solved in closed form, and for which it is possible to obtain recursive solutions in the maximum permitted number of simultaneous service instances.

The results of our model provide significant insight into the role played by system parameters such as the RRC timeout, and the persistence of the mobile users, showing how it is possible to achieve the best performance. As such, the model we presented in this chapter can be an important tool to assist cellular network configuration and management, and to support network operators in network planning and design.

Chapter 6

Conclusions

In this thesis, we have studied cellular radio access network performance and the data access phase in 3GPP-like systems (i.e., 4G and 5G), in case of HTC and MTC. The results obtained have been used to pinpoint which system components—either alone or while interacting with others—introduce delays, resource misuse or/and system bottleneck. Indeed, in cellular systems, the aggregate behaviour of multiple components might deviate remarkably from any of its single parts.

In Chapter 2 we presented a model to shed light on the key aspects of the typical reaction of a cellular network in crowded HTC environments. Indeed, it has been used to provide useful insight on cellular system operations in very crowded environments, and in the possibility to use it to design the correct dimensioning of the cellular system. As an example, the model allows the assessment of the benefits achievable through the adoption of D2D communications to reduce the congestion on the RACH more effectively than with ACB. Nonetheless, D2D effectiveness is not clear to cellular users, and hence the network provider needs to foster their collaboration. Chapter 3 provides a user-centric approach to content offload over D2D communications, showing an effective strategy of UEs stimulation.

In Chapter 4 we have investigated the challenges of MTC in the extremely demanding scenario of a Smart Factory. However, the merit of this chapter does not only relay on the analytical model provided, which nonetheless unveils some intrinsic limitations in the resource access phase, but also defines an new landscape for MTC. That is, 5GPPP foresees either massive or critical MTC scenarios, while here the two criticality merge in a massive and critical scenario.

The thoughtful study of the resource access phase carried on through this

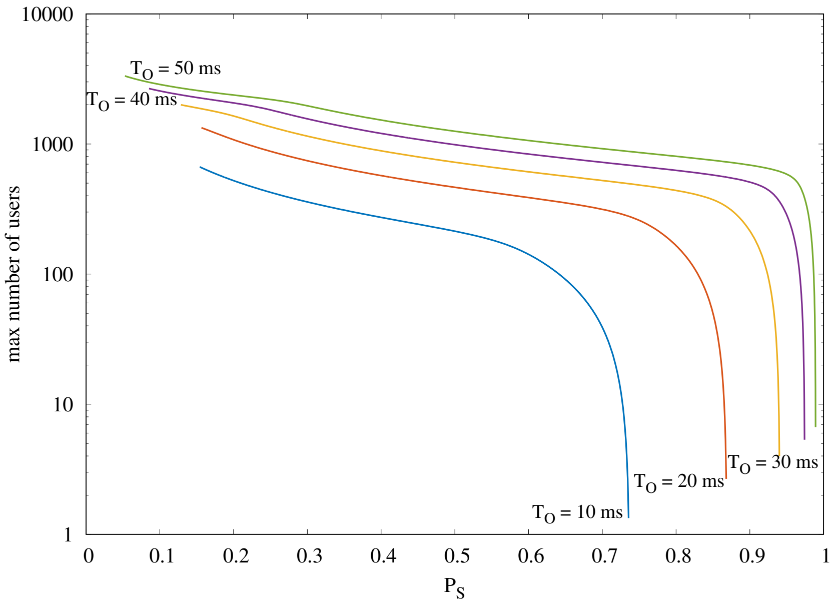

Figure 6.1: Maximum number of MTDs that can be connected to a eNB to guarantee a success probability above a threshold and latency below the timeout.

thesis headed to a novel formulation. Indeed, well-known formulae, like the M/M/1/m and the Erlang-B, do not account for fundamental requirements of cellular networks. As a result, such requirement either need to be additionally modelled or, in the worst case, are overseen. Therefore, the novel formulation of network processor devised in this thesis fills a gap in the literature and provides a valuable tool to support future research: it relates the returning clients with the outage probability, a fundamental measure in case of stringent requirements.

However, the work in this thesis opens up some of unanswered questions. For example, pursuing the ultra-low latency provisioned for Smart Factories, of about 1 to 10 milliseconds, remains a demanding task without a clear path to follow. Indeed, we showed that for so very low latencies cellular networks could not provide many guarantees due to the design of both RAN procedures and data service admission. Figure 6.1 clearly shows that in the scenario investigated in Chapter 4 pursuing application's timeouts lower than 30 ms results in success probability lower than 90% for nearly tens of UEs. Therefore, approaching latency of milliseconds with stringent guarantee of success is still a far result that requires to understand how and where to improve network performance. Therefore, for future developments and system improvement such components have to be either redesigned or enhanced, e.g., by introducing D2D for the RA procedure as it has been proposed in Chapter 2.

Furthermore, another interesting line of work that this thesis opens up is related to network slicing. Indeed, Chapter 4 investigates a scenario with two independent traffic flows–one with strict real-time constraint and the other without. Therefore, its easy to rethink that scenario introducing network slicing to manage traffic flows and guarantee them some sort of QoS. First of all, an analytical model would allow to study how the two slices interact in each one of the system components and, in a subsequent step, it would allow to investigate resource sharing policies and scheduling to maximise one or more performance indexes.

The academic relevance of this doctoral thesis is supported by publications in high level international conferences. Moreover, as pointed out from the results a re-think of the RAN and its procedure will be necessary, and this evidence opens up new research perspectives. Hence, contributions of this thesis are valuable to the research community, tracing possible future research directions, and to standardisation bodies.

Bibliography

[1] G. C. Madueno, J. J. Nielsen, D. M. Kim, N. K. Pratas, C. Stefanovic, and P. Popovski, "Assessment of LTE Wireless Access for Monitoring of Energy Distribution in the Smart Grid," *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 3, pp. 675–688, March 2016.

[2] 5GPPP. 5G empowering vertical industries. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2016/02/BROCHURE_5PPP_BAT2_PL.pdf

[3] Cisco Visual Networking Index. Global Mobile Data Traffic Forecast Update, 2016–2021 White Paper, accessed on March 28, 2017. [Online]. Available: https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/mobile-white-paper-c11-520862.html

[4] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[5] 5GPPP. (2015, October) 5G and the Factories of the Future. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Factories-of-the-Future-Vertical-Sector.pdf

[6] E. Dahlman, S. Parkvall, and J. Skold, *4G: LTE/LTE-advanced for mobile broadband.* Academic press, 2013.

[7] S. Sesia, M. Baker, and I. Toufik, *LTE-the UMTS long term evolution: from theory to practice.* John Wiley & Sons, 2009.

[8] J. G. Andrews, S. Buzzi, W. Choi, S. V. Hanly, A. Lozano, A. C. Soong, and J. C. Zhang, "What will 5g be?" *IEEE Journal on selected areas in communications*, vol. 32, no. 6, pp. 1065–1082, 2014.

[9] R. Ratasuk, N. Mangalvedhe, Y. Zhang, M. Robert, and J.-P. Koskinen, "Overview of narrowband IoT in LTE Rel-13," in *Standards for Communications and Networking (CSCN), 2016 IEEE Conference on.* IEEE, 2016, pp. 1–7.

[10] Ericsonn. (2016, June) Ericsson Mobility Report. [Online]. Available: https://www.ericsson.com/mobility-report

[11] N. Hallfors, M. Alhawari, M. A. Jaoude, Y. Kifle, H. Saleh, K. Liao, M. Ismail, and A. Isakovic, "Graphene oxide: Nylon ECG sensors for wearable IoT healthcarenanomaterial and SoC interface," *Analog Integrated Circuits and Signal Processing*, pp. 1–8, 2018.

[12] 5GPPP. (2015, October) 5G and e-Health. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-eHealth-Vertical-Sector.pdf

[13] H. Farhangi, "The path of the smart grid," *IEEE power and energy magazine*, vol. 8, no. 1, 2010.

[14] 5GPPP. (2015, October) 5G and Energy. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White_Paper-on-Energy-Vertical-Sector.pdf

[15] ——. (2015, October) 5G Automotive Vision. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP-White-Paper-on-Automotive-Vertical-Sectors.pdf

[16] ——. (2016, January) 5G and Media & Entertainment. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2016/02/5G-PPP-White-Paper-on-Media-Entertainment-Vertical-Sector.pdf

[17] M. Wollschlaeger, T. Sauter, and J. Jasperneite, "The future of industrial communication: Automation networks in the era of the internet of things and industry 4.0," *IEEE Industrial Electronics Magazine*, vol. 11, no. 1, pp. 17–27, 2017.

[18] L. Bassi, "Industry 4.0: Hope, hype or revolution?" in *Research and Technologies for Society and Industry (RTSI), 2017 IEEE 3rd International Forum on.* IEEE, 2017, pp. 1–6.

[19] 5GPPP. (2017, January) Vision on Software Networks and 5G. [Online]. Available: https://5g-ppp.eu/wp-content/uploads/2014/02/5G-PPP_SoftNets_WG_whitepaper_v20.pdf

[20] I. F. Akyildiz, P. Wang, and S.-C. Lin, "SoftAir: A software defined networking architecture for 5G wireless systems," *Computer Networks*, vol. 85, pp. 1–18, 2015.

[21] H. Kim and N. Feamster, “Improving network management with software defined networking,” *IEEE Communications Magazine*, vol. 51, no. 2, pp. 114–119, 2013.

[22] B. Han, V. Gopalakrishnan, L. Ji, and S. Lee, “Network function virtualization: Challenges and opportunities for innovations,” *IEEE Communications Magazine*, vol. 53, no. 2, pp. 90–97, 2015.

[23] P. Rost, C. Mannweiler, D. S. Michalopoulos, C. Sartori, V. Sciancalepore, N. Sastry, O. Holland, S. Tayade, B. Han, D. Bega *et al.*, “Network slicing to enable scalability and flexibility in 5g mobile networks,” *IEEE Communications magazine*, vol. 55, no. 5, pp. 72–79, 2017.

[24] “Evolved Universal Terrestrial Radio Access (E-UTRA); Medium Access Control (MAC) protocol specification,” 3GPP, TS 36.321 Release 13 V13.1.0, April 2016.

[25] Y. Y. Mihov, K. M. Kassev, and B. P. Tsankov, “Analysis and performance evaluation of the DRX mechanism for power saving in LTE,” in *Electrical and Electronics Engineers in Israel (IEEEI), 2010 IEEE 26th Convention of.* IEEE, 2010, pp. 000 520–000 524.

[26] A. T. Koc, S. C. Jha, R. Vannithamby, and M. Torlak, “Device power saving and latency optimization in LTE-A networks through DRX configuration,” *IEEE Transactions on wireless communications*, vol. 13, no. 5, pp. 2614–2625, 2014.

[27] Y.-P. Yu and K.-T. Feng, “
cycles adjustment scheme for 3GPP LTE systems,” in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th.* IEEE, 2012, pp. 1–5.

[28] L. Zhou, H. Xu, H. Tian, Y. Gao, L. Du, and L. Chen, “Performance analysis of power saving mechanism with adjustable DRX cycles in 3GPP LTE,” in *Vehicular Technology Conference, 2008. VTC 2008-Fall. IEEE 68th.* IEEE, 2008, pp. 1–5.

[29] H.-W. Ferng and T.-H. Wang, “Exploring Flexibility of DRX in LTE/LTE-A: Design of Dynamic and Adjustable DRX,” *IEEE Transactions on Mobile Computing*, vol. 17, no. 1, pp. 99–112, 2018.

[30] N. M. Balasubramanya, L. Lampe, G. Vos, and S. Bennett, “DRX with quick sleeping: A novel mechanism for energy-efficient IoT using LTE/LTE-A,” *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 398–407, 2016.

[31] C.-H. Wei, R.-G. Cheng, and S.-L. Tsao, "Modeling and estimation of one-shot random access for finite-user multichannel slotted ALOHA systems," *IEEE Communications Letters*, vol. 16, no. 8, pp. 1196–1199, 2012.

[32] D. Davis and S. Gronemeyer, "Performance of slotted ALOHA random access with delay capture and randomized time of arrival," *IEEE Transactions on Communications*, vol. 28, no. 5, pp. 703–710, 1980.

[33] T. P. de Andrade, C. A. Astudillo, and N. L. da Fonseca, "The impact of massive machine type communication devices on the access probability of human-to-human users in LTE networks," in *Communications (LATINCOM), 2014 IEEE Latin-America Conference on.* IEEE, 2014, pp. 1–6.

[34] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "On the Accurate Performance Evaluation of the LTE-A Random Access Procedure and the Access Class Barring Scheme," *IEEE Transactions on Wireless Communications*, vol. 16, no. 12, pp. 7785–7799, 2017.

[35] ——, "Performance analysis of access class barring for handling massive M2M traffic in LTE-A networks," in *Proc. of IEEE ICC*, May 2016.

[36] 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," *IEEE Transactions on Vehicular Technology*, vol. 63, no. 5, pp. 2467–2472, 2014.

[37] M. Vilgelm, H. M. Gürsu, W. Kellerer, and M. Reisslein, "LATMAPA: Load-adaptive throughput-maximizing preamble allocation for prioritization in 5G random access," *IEEE Access*, vol. 5, pp. 1103–1116, 2017.

[38] F. Morvari and A. Ghasemi, "Priority-based adaptive access barring for m2m communications in lte networks using learning automata," *International Journal of Communication Systems*, vol. 30, no. 16, 2017.

[39] M. Zubair Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "A First Look at Cellular Network Performance During Crowded Events," in *Proc. of the ACM SIGMETRICS*, ser. SIGMETRICS '13. New York, NY, USA: ACM, 2013, pp. 17–28.

[40] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2010-2015," Tech. Rep., February 2011. [Online]. Available: http://www.cisco.com/en/US/solutions/collateral/ns341/ns525/ns537/ns705/ns827/white_paper/c11-520862.pdf

[41] W. Mohr, "5G empowering vertical industries," Cisco, Tech. Rep., April 2016. [Online]. Available: https://ec.europa.eu/digital-single-market/en/blog/5g-empowering-vertical-industries-0

[42] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Communications Surveys & Tutorials*, 2014.

[43] A. Asadi, V. Mancuso, and R. Gupta, "An SDR-based Experimental Study of Outband D2D Communications," in *Proc. of IEEE INFOCOM*, Apr. 2016.

[44] M. Ajmone Marsan, D. Roffinella, and A. Murru, "ALOHA and CSMA protocols for multichannel broadcast networks," in *Proc. of Canadian Commun. Energy Conf.*, Montreal, P.Q., Canada, Oct. 1982.

[45] J. J. Nielsen, D. M. Kim, G. C. Madueno, N. K. Pratas, and P. Popovski, "A Tractable Model of the LTE Access Reservation Procedure for Machine-Type Communications," in *Proc. of IEEE GLOBECOM*, 2015.

[46] "Study on RAN Improvements for Machine-type Communications," 3GPP, TR 37.868 Release 11 V11.0.0, Septemeber 2011.

[47] L. A. Andres Laya and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *IEEE Communications Surveys & Tutorials*, vol. 16, no. 1, pp. 4,16, April 2011.

[48] O. Arouk and A. Ksentini, "General Model for RACH Procedure Performance Analysis," *IEEE Communications Letters*, vol. 20, no. 2, pp. 372–375, Feb 2016.

[49] T. P. de Andrade, C. A. Astudillo, and N. L. da Fonseca, "Random access mechanism for RAN overload control in LTE/LTE-A networks," in *Communications (ICC), 2015 IEEE International Conference on.* IEEE, 2015, pp. 5979–5984.

[50] Y.-C. Yuan-Chi Pang, G.-Y. Lin, and H.-Y. Wei, "Context-Aware Dynamic Resource Allocation for Cellular M2M Communications," *IEEE Internet of Things Journal*, vol. 3, no. 3, pp. 318–326, 2016.

[51] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *IEEE Communications Surveys and Tutorials*, vol. 16, no. 4, pp. 1801–1819, 2014.

[52] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, and A. Iera, "When D2D communication improves group oriented services in beyond 4G networks," *Wireless Networks*, pp. 1–15, 2014.

[53] L. Militano, M. Condoluci, G. Araniti, A. Molinaro, A. Iera, and F. H. P. Fitzek, “Wi-Fi cooperation or D2D-based multicast content distribution in LTE-A: A comparative analysis,” in *Proc. of IEEE International Conference on Communications, ICC*, 2014, pp. 296–301.

[54] T. Peng, Q. Lu, H. Wang, S. Xu, and W. Wang, “Interference avoidance mechanisms in the hybrid cellular and device-to-device systems,” in *Proc. of the IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, 2009, pp. 617–621.

[55] C. H. Yu, K. Doppler, C. Ribeiro, S. Xu, and O. Tirkkonen, “Performance impact of fading interference to Device-to-Device communication underlaying cellular networks,” in *Proc. of the IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, PIMRC*, 2009, pp. 858–862.

[56] A. Asadi and V. Mancuso, “Energy efficient opportunistic uplink packet forwarding in hybrid wireless networks,” in *Prof. of the fourth international conference on Future energy systems (e-Energy ’13)*, 2013, pp. 261–262.

[57] Q. Wang and B. Rengarajan, “Recouping opportunistic gain in dense base station layouts through energy-aware user cooperation,” in *Prof. IEEE Inter. Symposium on World of Wireless Mobile and Multimedia Networks (WoWMoM)*, 2013, pp. 1–9.

[58] W. Alliance, “Wi-Fi Peer-to-Peer (P2P) Specification v1. 1,” *WI-FI ALLIANCE SPECIFICATION*, vol. 1, pp. 1–159, 2010.

[59] A. Chen, D. Lee, and P. Sinha, “Optimizing Multicast Performance in Large-Scale WLANs,” in *Proc. of the 27th IEEE Int. Conf. on Distributed Computing Systems (ICDCS 2007)*. Toronto, Ontario, Canada: IEEE Computer Society, 2007.

[60] J. Huang, F. Qian, A. Gerber, Z. Mao, S. Sen, and O. Spatscheck, “A Close Examination of Performance and Power Characteristics of 4G LTE Networks,” in *Proc. of the 10th Int. Conference on Mobile Systems, Applications, and Services (MobiSys ’12)*. ACM, 2012.

[61] W. Xu, L. Liang, H. Zhang, S. Jin, J. C. Li, and M. Lei, “Performance enhanced transmission in device-to-device communications: Beamforming or interference cancellation?” in *Proc. of 2012 IEEE Global Communications Conference, (GLOBECOM)*, 2012, pp. 4296–4301.

[62] R. Zhang, X. Cheng, L. Yang, and B. Jiao, “Interference-aware graph based resource sharing for device-to-device communications underlaying cellular networks,” in *Prof. of 2013 IEEE Wireless Communications and Networking Conference (WCNC)*, 2013, pp. 140–145.

[63] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, "Device-to-device communication in 5G cellular networks: challenges, solutions, and future directions," *IEEE Communications Magazine*, vol. 52, no. 5, pp. 86–92, 2014.

[64] M. Du, X. Wang, D. Wang, and Y. Wang, "Device-to-Device Dynamic Clustering Algorithm In Multicast Communication," in *Proc. of IEEE 11th International Conference on Dependable, Autonomic and Secure Computing*, 2013.

[65] B. Zhou, H. Hu, S.-Q. Huang, and H.-H. Chen, "Intracluster Device-to-Device Relay Algorithm With Optimal Resource Utilization," *IEEE T. Vehicular Technology*, vol. 62, no. 5, pp. 2315–2326, 2013.

[66] M. Condoluci, L. Militano, G. Araniti, A. Molinaro, and A. Iera, "Multicasting in LTE-A networks enhanced by device-to-device communications," in *Workshops Proceedings of the Global Communications Conference, GLOBECOM*, 2013, pp. 567–572.

[67] F. Pantisano, M. Bennis, W. Saad, and M. Debbah, "Spectrum Leasing as an Incentive Towards Uplink Macrocell and Femtocell Cooperation," *IEEESelected Areas in Communications*, vol. 30, no. 3, 2012.

[68] Y. Zhang, L. Song, W. Saad, Z. Dawy, and Z. Han, "Contract-Based Incentive Mechanisms for Device-to-Device Communications in Cellular Networks," *IEEESelected Areas in Communications*, vol. PP, no. 99, 2015, to appear.

[69] M. Maternia and S. E. El Ayoubi (editors), "5G PPP use cases and performance evaluation models," 5G-PPP, Tech. Rep., Apr. 2016.

[70] I. Leyva-Mayorga, L. Tello-Oquendo, V. Pla, J. Martinez-Bauset, and V. Casares-Giner, "Performance Analysis of Access Class Barring for Handling Massive M2M Traffic in LTE-A Networks," in *IEEE International Conference on Communications (ICC)*, 2016.

[71] D.-W. Seo, "Explicit Formulae for Characteristics of Finite-Capacity M/D/1 Queues," *ETRI Journal*, vol. 36, no. 4, pp. 609–616, 2014.

[72] E. Dahlman, S. Parkvall, and J. Skold, *4G, LTE-Advanced Pro and The Road to 5G, Third Edition*, 3rd ed. Academic Press, 2016.

[73] A. Biral, M. Centenaro, A. Zanella, L. Vangelista, and M. Zorzi, "The challenges of M2M massive access in wireless cellular networks," *Digital Communications and Networks*, vol. 1, no. 1, pp. 1 – 19, 2015.

[74] O. Arouk, A. Ksentini, and T. Taleb, "Performance Analysis of RACH Procedure with Beta Traffic-Activated Machine-Type-Communication," in *Proc. of GLOBECOM*, 2015.

[75] S. Cherkaoui, I. Keskes, H. Rivano, and R. Stanica, "LTE-A random access channel capacity evaluation for M2M communications," in *Proc. of Wireless Days*, 2016.

[76] P. Castagno, V. Mancuso, M. Sereno, and M. Ajmone Marsan, "Why Your Smartphone doesn't Work in Very Crowded Environments," in *Proc. of WoWMoM*, 2017.

[77] Z. Alavikia and A. Ghasemi, "A multiple power level random access method for M2M communications in LTE-A network," *Transactions on Emerging Telecommunications Technologies*, 2016.

[78] M. S. Ali, E. Hossain, and D. I. Kim, "LTE/LTE-A Random Access for Massive Machine-Type Communications in Smart Cities," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 76–83, January 2017.

[79] D. Niyato, P. Wang, and D. I. Kim, "Performance modeling and analysis of heterogeneous machine type communications," *IEEE Transactions on Wireless Communications*, vol. 13, no. 5, pp. 2836–2849, 2014.

[80] A. Q. Gill, D. Bunker, and P. Seltsikas, "Moving Forward: Emerging Themes in Financial Services Technologies Adoption," *Communications of the Association for Information Systems*, 2015.

[81] C. Torres-Huitzil and A. Alvarez-Landero, "Accelerometer-Based Human Activity Recognition in Smartphones for Healthcare Services," *Mobile Health, Springer Series in Bio-/Neuroimformatics*, vol. 5, pp. 147–169, 2015.

[82] J. Gubbi, R. Buyya, S. Marusic, and M. Palaniswami, "Internet of Things (IoT): A Vision, Architectural Elements, and Future Directions," *Future generation computer systems*, vol. 29, no. 7, 2013.

[83] D.-H. Shin, "Measuring the Quality of Smartphones: Development of a Customer Satisfaction Index for Smart Services," *International Journal of Mobile Communications*, vol. 12, no. 4, pp. 311–327, 2014.

[84] M. Z. Shafiq, L. Ji, A. X. Liu, J. Pang, S. Venkataraman, and J. Wang, "Characterizing and Optimizing Cellular Network Performance During Crowded Events," *IEEE/ACM Trans. Netw.*, vol. 24, no. 3, pp. 1308–1321, Jun. 2016.

[85] Y. J. Choi, S. Park, and S. Bahk, “Multichannel Random Access in OFDMA Wireless Networks,” *IEEE Journal on Selected Areas in Communications*, vol. 24, no. 3, pp. 603–613, Mar. 2006.

[86] P. Castagno, V. Mancuso, M. Sereno, and M. Ajmone Marsan, “Why Your Smartphone Doesn’t Work in Very Crowded Environments,” in *Proc. of WoWMoM*, 2017.

[87] C. Úbeda, S. Pedraza, M. Regueira, and J. Romero, “LTE FDD Physical Random Access Channel Dimensioning and Planning,” in *Proc. of IEEE VTC Fall*, 2012.

[88] F. Baskett, K. M. Chandy, R. R. Muntz, and F. G. Palacios, “Open, Closed, and Mixed Networks of Queues with Different Classes of Customers,” *Journal of ACM*, vol. 22, no. 2, pp. 248–260, Apr. 1975.

[89] S. Asmussen and M. Bladt, “Renewal Theory and Queueing Algorithms for Matrix-Exponential Distributions,” in *Matrix-analytic Methods in Stochastic Models,* Lecture Notes in Pure and Applied Mathematics, vol. 183, 1997, pp. 313–341.

[90] M. Reiser and S. S. Lavenberg, “Mean-Value Analysis of Closed Multichain Queuing Networks,” *Journal of ACM*, vol. 27, no. 2, pp. 313–322, Apr. 1980.

[91] K. C. Sevcik and I. Mitrani, “The Distribution of Queuing Network States at Input and Output Instants,” *J. ACM*, vol. 28, no. 2, 1981.

[92] S. Batabyal and S. S. Das, “Distance Dependent Call Blocking Probability, and Area Erlang Efficiency of Cellular Networks,” in *Proc. of IEEE VTC Spring*, May 2012.