UNIVERSITÀ
DEGLI STUDI
FIRENZE

DOTTORATO DI RICERCA IN
INFORMATICA, SISTEMI E
TELECOMUNICAZIONI

CICLO XXVIII

COORDINATORE Prof. Luigi Chisci

Access and Resource Allocation Strategies for Machine-to-Machine Communications

Settore Scientifico Disciplinare ING-INF/03

Dottorando
Dott. Giovanni Rigazzi

Coordinatore
Prof. Luigi Chisci

Tutori
Prof. Romano Fantacci

Prof. Petar Popovski

Ing. Francesco Chiti

Ing. Tommaso Pecorella

Ing. Camillo Carlini

Anni 2012 / 2015

Access and Resource Allocation Strategies for Machine-to-Machine Communications

by

Giovanni Rigazzi

Submitted to the Department of Information Engineering
on November 30, 2015, in partial fulfillment of the
requirements for the degree of
DOCTOR OF PHILOSOPHY

Abstract

Machine-to-Machine (M2M) communications represent the cornerstone technology enabling the pervasive deployment of automated applications, connecting billions of devices or objects without the need for human intervention. Within the past few years, the number of M2M-based services has dramatically grown, spurred by remarkable benefits provided by the deployment of smart and cost-effective applications in a wide range of areas, including remote sensing, health monitoring and Intelligent Transportation Systems (ITS). Therefore, great research effort has been invested to the design of novel solutions or to the optimization of existing communication systems in order to accommodate M2M traffic with diverse QoS requirements.

This thesis introduces novel approaches to enhance M2M communications based on the adoption of Device-to-Device (D2D) connectivity among the devices, in conjunction with existing techniques widely implemented, such as clustering, packet aggregation and trunking. In the context of cellular systems, D2D communications are an attractive and intelligent solution to reduce the power consumption and the latency due to the short-

range connectivity employed by two devices operating within the licensed spectrum. Furthermore, the cellular operator coordinates peer discovery and link establishment and provides the same security mechanisms adopted in conventional cellular communications.

In the first approach, we envision a centralized scheme to allocate cellular resources in a heterogeneous scenario, where the machine devices can exploit D2D connectivity, realizing a multi-hop D2D network overlaying an LTE-A system. The main challenges involving the D2D link establishment are covered, such as proximity discovery and device pairing, and a routing mechanism is proposed in order to efficiently forward packets to a collector device through multi-hop transmissions.

To address the RAN overload issue originated from the mass access, a cellular user can collect and combine the traffic generated by machine devices through low power D2D connections and then forward the aggregated data to the attached base station, along with the traffic originated from higher layers. As a result, enormous benefits can be obtained in terms of energy efficiency and throughput, as the cellular user acts as a relay and is responsible of the data transmission on the trunked cellular uplink. We further enhance this mechanism by introducing a backoff mechanism to mitigate collisions occurring during the access reservation phase, performed by the machine devices before transmitting a packet. In addition, we evaluate the advantages in terms of trunking gains provided by a multi-cluster configuration, where multiple cellular users forward traffic sent by devices belonging to different clusters, without generating interference.

In the context of vehicular networking, M2M communi-

cations aim to enable the internet-working among connected vehicles, which can autonomously exchange information and make quick decisions in critical situations. The main challenge associated with Vehicular Ad-hoc Networks (VANET) is how to provide reliable connectivity between vehicles in scenarios with highly dynamic topologies and unpredictable channel conditions. We choose to overcome this issue by developing a clustering algorithm based on the mobility correlation degree of the vehicles in a typical highway scenario, along with a relaying scheme aiming at supporting the communication between different clusters. The proposed protocol is then evaluated by implementing a realistic mobility model, taking into account interactions between approaching vehicles and assuming that the communications are established according to the IEEE 802.11p/WAVE standard.

Thesis Supervisor: Romano Fantacci, Full Professor

Thesis Supervisor: Petar Popovski, Full Professor

Thesis Supervisor: Francesco Chiti, Assistant Professor

Thesis Supervisor: Tommaso Pecorella, Assistant Professor

Thesis Supervisor: Camillo Carlini, Telecom Italia

List of Publications[1]

Journal Papers

[1] G. Rigazzi, N. K. Pratas, P. Popovski, and R. Fantacci, "M2M Traffic Offloading onto Trunked Cellular Uplink via D2D Connections," *Submitted to IEEE Trans. on Wireless Comm.*, Sept. 2015.

[2] F. Chiti, R. Fantacci, R. Mastandrea, G. Rigazzi, A. Suarez Sarmiento, and E. Macias Lopez, "A Distributed Clustering Scheme with Self Nomination: Proposal and Application to Critical Monitoring," *Wireless Networks*, Springer US, August 2014

Conference Papers

[1] G. Rigazzi, N. K. Pratas, P. Popovski, and R. Fantacci, "Aggregation and Trunking of M2M Traffic via D2D Connections," *IEEE ICC 2015 - Mobile and Wireless Net-*

[1]The numbering associated with publications does not refer to the Bibliography section of this thesis.

working Symposium (ICC'15), London, United Kingdom, June 2015

[2] G. Rigazzi, F. Chiti, R. Fantacci, and C. Carlini, "Multi-hop D2D networking and resource management scheme for M2M communications over LTE-A systems," *IEEE IWCMC*, Cyprus, August 2014

[3] T. Pecorella, R. Viti, G. Rigazzi, and C. Carlini, "Application driven SON," *IEEE IWCMC*, Cyprus, August 2014

[4] F. Chiti, G. Rigazzi, and R. Fantacci, "A Mobility Driven Joint Clustering and Relay Selection for IEEE 802.11p/WAVE Vehicular Networks," *IEEE ICC 2014*, Sydney, Australia, June 2014

[5] A. Tassi, G. Rigazzi, C. Khirallah, D. Vukobratovic, F. Chiti, J. Thompson, and R. Fantacci, "Reliable Multicast in LTE-A: An Energy Efficient Cross-layer Application of Network Coding," *IEEE ICT 2013*, Casablanca, Morocco, May 2013

Book Chapters

[1] F. Chiti, R. Fantacci, D. Giuli, F. Paganelli, G. Rigazzi, "Vehicular Networking in 5G," book chapter in *5G Mobile Communications*, Springer

Acknowledgements

First of all, I would like to thank Prof. Romano Fantacci for giving me the opportunity to conduct the doctoral studies and for the valuable advice provided in many circumstances in the past three years.

I wish to thank all the guys of the former LART laboratory in Florence. Among all, special thanks to Andrea, who continuously encouraged and helped me find the strength to complete my Ph.D.

I would also thank Prof. Petar Popovski for the priceless and rewarding working experience I had within his research group during my visiting period at Aalborg university. His insightful guidance and constant support not only gave me great motivations, but also significantly contributed to my growth in many aspects. Thanks also to Nuno for the enormous patience and the fundamental help provided during and after the visiting period in Denmark, and to all the guys I met in AAU for the unforgettable days I spent there.

Last but not least, I wish to thank my family, who has always believed in me and strongly supported throughout this experience.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Machine-to-Machine Communications

Machine-to-machine (M2M) or machine-type (MT) communications refer to the technology that enables a system of wired and wireless networks, usually distributed across the world, for collecting data from machines-type devices (MTD), such as sensors and meters, that are massively and densely deployed, and for transmitting events captured by low-end devices to high-end applications. It is anticipated that M2M communications will expand to 2.1 billion connections by 2021, realizing the so-called Internet of Things (IoT), where physical objects foster quick and direct integration between computing systems and physical world [1]. This, in turn, will not only generate remarkable revenue growth and new opportunities for network operators and service providers, but will also lead to the exponential increase in resource demands necessary to accommodate significantly different service requirements, ranging from low latency and high reliability as in mission-critical applications, or high scalability and service availability for smart-metering services.

Unlike human-centric communications, where remote servers pro-

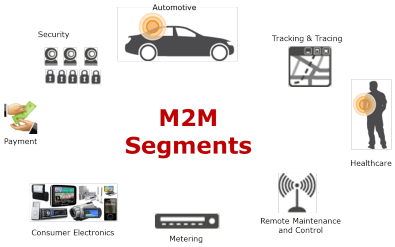


Figure 1-1: Overview of potential M2M applications.

vide services to the mobile users and the majority of the data traffic is carried over the downlink channel, in M2M communications a wide range of heterogeneous devices can automatically connect to other devices or to machine-type servers and exchange data, according to the specific M2M application and service. Fig. 1-1 illustrates the most relevant M2M applications and related scenarios. For instance, smart meters have been largely deployed to efficiently monitor energy and utilities, or in asset tracking and logistics, whereas healthcare applications and Intelligent Transportation Systems (ITS) can employ wireless sensors for data reporting and critical event detection. In the context of M2M, car-to-car (C2C) or vehicular communications enable a wide range of innovative and attractive applications, such as accident prevention, assisted driving and infotainment. Moreover, asset security and protection can be enhanced by adopting remote surveillance systems, capable of detecting relevant events and generating alarms accordingly.

1.1.1 Key Features of M2M Communications

Despite the vast range of M2M applications exhibiting different traffic patterns, M2M communications present common characteristics and key features, which can be classified as follows:

- Massive number of devices

Several M2M applications typically involve the deployment of a very large number of devices, which is commonly referred to as massive M2M scenario [2]. As an example, smart metering is being considered a very effective cellular M2M solution to improve the efficiency of power distribution and reduce the waste of energy. In such a scenario, a high number of densely deployed devices can autonomously and periodically exchange data with the grid infrastructure, thus generating congestions and affecting the performance of the channel. This not only leads to a huge waste of resources and increased network delay, but also to the performance degradation of H2H communications.

- Small data burst transmissions
  In numerous applications, a machine may transmit at low frequency, e.g., one transmission every few minutes or hours, a small amount of data. Such a traffic leads to high signaling overhead compared to the amount of of information data, as the network needs to activate the necessary signaling procedures involving the Radio Resource Control (RRC) layer, to set up and release the radio bearers at the end of the burst. To reduce the amount of signalization required, new promising solutions aim to apply *connectionless* transmissions in 4G or 5G networks, where packet size less than 300 bytes can be conveyed without involving any radio bearer establishment [3].

- Energy efficiency
  One of the crucial requirements in the development of M2M applications is to deploy highly energy-efficient devices, able to operate for several years without the need for battery replacement. For instance, in environment monitoring, MTDs are typically located in remote areas and energy harvesting solutions cannot be applied due to low-cost or low-complexity constraints. For this reason, the support to a low-power consumption mode must be provided by the cellular system, e.g., increasing the duration of the DRX cycle [4] or disabling frequent

signaling transmissions, such as tracking area updates [5].

- Low cost/complexity
  Due to the large number of devices expected to be deployed in the next few years, low complexity design becomes of paramount importance for reducing manufacturing cost of the single device. This motivates the efforts of wireless industry to develop ultra-low complexity devices equipped with low-power single antennas and half-duplex RF transceivers, avoiding sophisticated receiver equalization techniques and limiting the data rates supported, which considerably reduce the cost required for processing power and memory [6].

- Security
  Security is a fundamental requirement in M2M communications due to the potentially high risk of physical and remote attacks. Although authentication and data integrity functions can be provided by existing cellular systems, several applications may require additional security for the data, being the devices often unattended and vulnerable to software manipulation. As a result, new solutions have been proposed to overcome these issues, such as embedding subscriber identity module (SIM) cards or utilizing physical-layer security techniques with RF fingerpriting, which allow to detect tampering or position variations [7].

1.1.2 M2M Emerging Technologies

Driven by attractive opportunities to optimize capital and maintenance costs, manufacturing companies have been gradually started implementing M2M services, e.g., sensor and actuator devices, mostly equipped with low-cost cellular interfaces, for business critical functions, such as telemetry and automation process control. Although GSM-based systems can represent a mature and convenient solution, capable of providing ubiquitous coverage and global connectivity as

well a high level of security, the large number of simultaneous connections within a cell may generate overload situations and then saturate access and core network. Therefore, alternative technologies have been proposed and investigated by wireless industry and academia. In the following, we briefly illustrate the most promising solutions introduced to support M2M communications and evaluate pros and cons of each approach.

IEEE 802.11ah

IEEE 802.11 (WiFi) systems have significantly evolved in the last few years, becoming a popular wireless connectivity solution for home and office environments, due to cost-effective and easy deployment on unlicensed bands. Significant enhancements have been also developed with the goal of reducing the power consumption and increasing the overall energy efficiency. Nevertheless, the lack of mobility and QoS support as well as interference mitigation mechanisms makes WiFi systems not adequate to satisfy M2M communication requirements. To overcome these challenges, IEEE has mainly focused on the standardization of a novel technology, called IEEE 802.11ah, based on the IEEE 802.11ac standard, but capable of supporting long-range connections of extremely dense M2M deployment, also guaranteeing high spectral and energy efficiency [8]. In addition, IEEE 802.11ah is not backward compatible with the other members of the IEEE 802.11 family. As a result, new access schemes and physical layer solutions may be also developed to further increase the throughput and limit the power consumption.

Unlicensed Low-Power Wide-Area Technology (LPWA)

Low-Power Widea-Area (LPWA) represents a new category of proprietary M2M technologies operating on unlincesed bands and primarly supporting applications generating small data bursts with low data rate, such as smart meters transmitting small amounts of traffic on a daily basis [9]. This technology proved to be highly effective in

large-scale M2M installation, with a small number of access points involved and in diverse coverage conditions. However, long-range transmissions on unlicenced spectrum are limited by technical regulations, while the antenna gains of an MTD and a BS can largely differ, thus causing link budget asymmetry between downlink and uplink. Moreover, scalability may be considerably affected by interference originated from the huge number of devices with radio interfaces operating on the same spectrum, such as Wi-Fi, Zigbee and IEEE 802.15.

Cellular M2M

To date, cellular M2M communications over LTE/LTE-A networks have been considered the most mature and cost-effective technology to handle the large amount of traffic in highly dense M2M deployments [10]. Besides the intrinsic benefits in terms of low installation and management costs due to the existing infrastructure, M2M communications can take advantage of broad coverage, scalability and reliable security protocols. Nevertheless, such systems are designed and optimized for H2H communications and cannot efficiently accommodate traffic with diverse patterns and requirements. As a consequence, numerous studies have been carried out by RAN working groups, mostly focusing on adapting current LTE systems to M2M communications. For instance, LTE Release 12 and 13 introduce a new UE category tailored to M2M communications, called LTE Cat.0, which aims at providing higher coverage and reducing the cost and the complexity of the hardware, by enabling transmissions on narrower bandwidths [11]. As the massive number of devices requesting access poses the major concern to the access network, huge research efforts have been invested to mitigate the overload issue. Some of the proposed techniques are illustrated in more details in Chapter 2.2. Furthermore, power consumption can be drastically reduced by limiting the signalization overhead due to small and infrequent data burst transmissions, or by extending the discontinuous reception (DRX) cycle, at the expense of extra delay [4].

Although the deployment of a vast number of low-range small cells may represent a valid option to support high volumes of traffic and reduce the transmit power, potential increase in interference and high operative costs significantly limit the use of this solution. Therefore, an alternative approach consists of adopting a hybrid model [12], where devices are connected using a specific radio technology within a cluster and can send data to a certain cluster head, which then forwards the aggregated traffic to the eNB. This solution can reduce the number of devices attached to the same eNB, thus preventing the occurrence of congestions in the RAN.

Another approach to relieve network congestions follows the Cognitive Radio (CR) paradigm [13]. One technique allows the MTDs to sense the licensed spectrum and transmit whenever a CU is not active. To this end, cognitive devices attach to a dedicated eNB station, which is responsible of coordination with the other eNB stations handling the traffic originated from H2H communications. As a result, spectral efficiency can be increased at the expense of scalability and QoS support as the number of MTDs becomes higher. An alternative solution assumes that smart eNodeBs (S-eNB) are in charge of simultaneously sensing and monitoring both licensed and unlicensed bands, such as 5 GHz and TV white space bands [14]. This allows to increase the number of MTDs supported as well as to enhance spectral efficiency, as long as the required level of system complexity is fulfilled.

1.1.3 Network Architecture for M2M Communications

The need to provide global integration among numerous solutions for M2M applications represents the main motivation for designing and developing a standard end-to-end M2M communication architecture, taking into account the diverse characteristic features associated with the MTDs.

Fig. 1-2 shows a high-level architecture of a cellular-based M2M

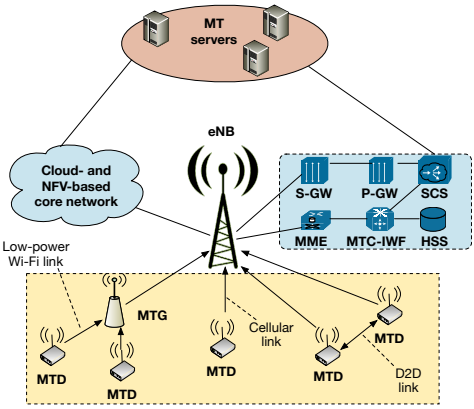


Figure 1-2: Future network architecture to support M2M communications.

communication system, where we can identify three different layers or domains. The lowest level contains MTDs equipped with cellular interfaces, along with MT gateways (MTG), deployed to forward traffic originated from devices belonging to unlicensed low-power networks, also referred to as *capillary networks* [15]. In addition, MTDs may also exchange data through Device-to-Device (D2D) connectivity, as proposed in latest releases of 3GPP Long Term Evolution (LTE) standard. To allow the communication between MTDs and MT servers, a cellular network is then employed, mainly consisting of a radio access network (RAN) and a core network (CN) or evolved packet core (EPC) in LTE systems. The former includes all the functions to manage the air interfaces, whereas the latter controls the packet delivery, providing a wide variety of services, such as authentication and traffic monitoring. An EPC embraces all the network elements handling both M2M and Human-to-Human (H2H) communications. Serving gateway (SGW) and packet data network gateway (PGW) manage the packet forwarding through the network by means of bearer channels established with the end users. Mobility-

related functions are then performed by the mobility management entity (MME), while the home subscriber server (HSS) stores information on subscribed users. Finally, service capability servers (SCS) and MTC-inter-working function (MTC-IWF) provide services for M2M applications deployed in external networks.

This architecture also reflects the predicted evolution of current cellular networks towards the next 5G systems [14]. In this context, network function virtualization (NFV) will be one of the key M2M enablers due to the enormous benefits provided by the deployment of virtual machines managing network functions, thus reducing the use of expensive network infrastructures. Based on the resource demands, NFV can efficiently allocate virtual hardware instances, while higher layer functions performed by the eNB can be implemented by a network cloud, which further lowers the operational maintenance costs.

1.2 Focus of the Work

In this thesis, we concentrate on the analysis of the impact of M2M communications within traditional cellular systems and propose novel solutions to accommodate traffic originated from MTDs without affecting the QoS of H2H communications. The majority of the works on this topic mainly focuses on either optimizing existing cellular systems or devising new schemes to adapt the RAN to the massive number of access requests occurring in large-scale M2M deployments. Nonetheless, M2M traffic is mostly sporadic and packets are small in size, thus generating data burst transmissions with short durations. As a result, predicting the incoming traffic and optimizing the system accordingly is considered a quite challenging task.

Our approach aims at enabling cellular M2M communications by leveraging D2D connectivity between MTDs and nearby CUs. Following the principle of *capillary networks*, where MTDs communicate through short-range communication technologies and send data to a gateway device, we elaborate a packet aggregation and trunking

mechanism with the goal of increasing the system throughput and improving the energy efficiency. In the context of M2M communications, traffic aggregation represents an attractive solution to mitigate the small-packet burst transmission issue, along with a trunking mechanism capable of carrying data belonging to different flows. Once M2M packets are aggregated and combined with data locally generated by the CU, the aggregated traffic is sent to the base station (BS), which can then dispatch each packet to the predefined destination.

D2D connectivity can also be conveniently exploited to realize a D2D multi-hop network overlaying a cellular network. Through network-assisted procedures, D2D pairs can be located and D2D links can be enabled utilizing reserved cellular resources, thus avoiding additional interference and preventing the degradation of H2H communication QoS. Furthermore, network assistance is of paramount importance for the multi-hop communication as the centralized resource allocation policy is designed to assign more resources to devices acting as relays, with the aim of mitigating potential bottlenecks in the multi-hop overlay network.

Peer-to-peer communications are a key element in VANETs. In highly dynamic scenarios, communications are typically established among nodes within the same coverage and a cellular infrastructure cannot provide the required connectivity due to unpredictable channel conditions. To help improve communications, a clustering scheme represents an advantageous solution due to the reduced number of connections needed to exchange data. Once a cluster head is elected, communications are only performed between cluster members and cluster head, which can also forward traffic to other clusters via inter-cluster communications. Therefore, significant benefits in terms of energy efficiency and latency reduction can be guaranteed by employing a clustering scheme in VANETs.

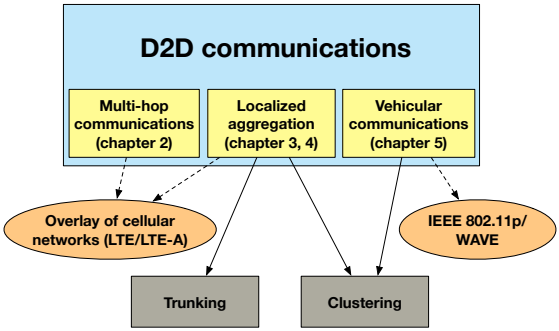


Figure 1-3: Block diagram illustrating the contributions of the thesis as well as paradigms and technologies adopted.

1.3 Thesis Contributions

As mentioned above, this work focuses on enabling M2M communications by conveniently applying D2D communications and by adopting well-know paradigms, such as clustering, aggregation and trunking. To help comprehend each contribution of this thesis, Fig. 1-3 illustrates the main concepts employed in the following chapters, along with the technology solutions utilized. In the following, we also provide a brief description of the organization of the thesis.

1. D2D communications represent a key enabler for M2M communications within cellular systems. In chapter 2, we design a novel protocol to set up a multi-hop network by leveraging D2D connectivity among nearby devices. We choose to employ D2D overlay communications, thus avoiding additional interference with traditional cellular H2H communications and reserving a given amount of resources for the communications between D2D pairs. Furthermore, our scheme allows eNBs to identify clusters of potential D2D users, whereas the D2D discovery phase is handled by enabling the broadcast of beacon signals. The multi-hop communication is then performed by building a

connected graph and allocating resources according to the role of the D2D users in the network.

2. Another solution exploiting low-power D2D connectivity is introduced in chapter 3. We follow the paradigm of traffic aggregation and trunking and develop an innovative scheme to convey M2M traffic to an eNB without overloading the RAN. First, MTDs transmit packets through D2D links established with a close CU, once access is gained by performing a Framed Slotted ALOHA algorithm. Then, MTD packets are aggregated with user traffic generated by higher layers and sent to the eNB via the cellular uplink. This approach guarantees a considerable reduction of the consumed power as the MTDs use short-range D2D links, while RAN congestion is prevented since only the CU is allowed to communicate with the eNB.

3. Chapter 4 describes an enhanced D2D trunking protocol to offload M2M traffic onto trunked cellular uplinks. We mainly focus on improving the access reservation phase by introducing a backoff algorithm, which helps mitigate the collisions generated by devices simultaneously sending an access request. System performance is analyzed by elaborating a Markov chain, which reflects the temporal evolution of an access request performed by an MTD. In addition, we define the trunking gains as the gains in terms of throughput, energy efficiency and delay reduction, obtained by partitioning a certain number of MTDs in the same cell in clusters and providing a trunked link in each cluster, thus allowing a multi-cluster offloading without generating any interference.

4. In chapter 5, we consider vehicular communications in scenarios characterized by highly dynamic topologies, where vehicles communicate through the IEEE 802.11p/WAVE protocol. Our idea is to enable the exchange of critical data among vehicles by employing a cluster formation algorithm based on the relative positions of a node with respect to a potential cluster-head.

This allows to establish a communication link with a single vehicle, which can then forward the collected traffic to other cluster-heads within the same radio coverage.

Chapter 2

Multi-hop D2D Networking and Resource Management Scheme for M2M Communications over LTE-A Systems

2.1 Introduction

Along with the Machine Type Communication (MTC) definition, the Third Generation Partnership Project (3GPP) Release 12 introduces in LTE-A the support to D2D or direct-mode communications, allowing P2P transmissions between devices in proximity [16]. This functionality enables a new generation of devices able to communi-

cate with each other, without any kind of human intervention and without involving the cellular infrastructure for the user plane. Despite the numerous benefits, such as high spectral efficiency, enhanced coverage and low power consumption, D2D communications pose also several challenges. First of all, direct mode communications should be established without impairing traditional communications via BS, i.e., by avoiding possible *interferences* with other devices. Secondly, peer and service discovery functionalities need to be introduced, since a device is typically not aware of other terminals in the proximity. In [17] and [18], for example, two different solutions are proposed to accommodate cellular and D2D communications in Frequency Division Duplexing (FDD) and Time Division Duplexing (TDD) systems, respectively. Following the paradigm of Multi-hop Cellular Networks (MCN) [19], D2D can also provide more flexibility in a cellular network, as direct links enable the communication among remote terminals using multiple hops.

In this chapter, we first examine major issues introduced by the deployment of a massive number of machines in an LTE-A cellular network. We then describe a multi-hop networking scheme consisting of opportunistic D2D communications, where all the terminals are capable of selecting the best route to reach a given destination, without relying on a centralized routing algorithm. Finally, a network-aware resource allocation algorithm is applied in order to handle diverse bandwidth requests dictated by terminals taking part in the multi-hop D2D network.

The remainder of the chapter is organized as follows: in Section 2.2 and 2.3, we analyze the key aspects of MTC and D2D communication, respectively, pointing out the main issues related to the realization of a *multi-hop* network as overlay of a cellular network. Then, in Section 2.4 we describe the system model and propose a novel approach to enable multi-hop communication via D2D communications. Conclusions are finally drawn in the Section 2.5.

2.2 Key aspects of MTCs

The design of M2M communications in LTE-A networks presents specific issues deriving from the coexistence of human and machine type communications. As shown in Tab. 2.1, M2M communications can be classified in three different categories, depending on the application scenario, each characterised by different features and specific requirements in terms of latency, packet loss, traffic volume and scalability. Furthermore, for each application class, a potential set of technical solutions is presented, according to the specific technology area.

2.2.1 Channel access

A critical aspect of M2M communications consists in guaranteeing the *simultaneous* access of MTDs and reducing the probability of congestion due to the massive number of devices demanding resources. Unlike human-based communication, an MTD may demand network access at fixed times or upon external triggers, depending on the specific application or service provided. For instance, a *smart grid* needs to collect information sent by sensor devices in order to process and provide suitable decisions to improve the efficiency of electricity generation and distribution. As a result, a massive number of access requests are expected and each device may transmit or receive data with a different degree of regularity.

To handle the network access phase, LTE-A systems provide two different solutions based on the Random Access (RA) technique: *i)* a contention-based approach, where access grants are assigned only after a contention resolution and collision events may frequently occur, and *ii)* a contention-free solution, where semi-static access resources are allocated to a specific class of devices with low latency requirements. Since in an M2M scenario a high number of devices are expected to require access at the same time, the former approach turns out to be more appropriate. Furthermore, it is worth pointing out that in LTE-A the RA phase can occur not only during the network association or in case of link outage, but also when the devices are

Table 2.1: M2M application scenarios and technological solutions.

Application Scenario		Environmental monitoring	Domotic and Industrial	Military and bio-medical
Main features	Cost	Low	Medium	High
	Density	High	Medium	Low
	Mobility	Static	Low	Medium-High
	Battery Duration	Medium	Medium	High
Requirements	Latency	Medium	Medium	Very Low
	Packet Loss	Medium	Low	Very low
	Traffic volume	Small	Moderate	Moderate
	Scalability	High	Medium	Medium
Technology Areas	Channel access	ACB, CERA	ACB, CERA, Resource separation	CERA, Dynamic allocation
	Resource management	Shared	Orthogonal, Shared	Orthogonal
	Proximity discovery	Non-assisted	Network-assisted	Network-assisted

not synchronized or scheduling request resources are missing on the uplink control channel.

Devices attempting to access the network utilize the RACH channel by choosing a preamble which is transmitted on a free RA slot. Accordingly, whenever multiple devices select and transmit the same preamble on the same slot, a collision occurs and a backoff algorithm is activated to resolve the contention. As described in [20], an LTE-A system with 10 of 64 preambles reserved for contention-free operations can provide 200 access opportunities every second, assuming a gap of 5 ms between two access attempts. Nevertheless, the presence of a random backoff algorithm can strongly affect the performance of the random access technique, limiting the maximum number of successful channel access attempts.

In an effort to overcome the aforementioned problem, several solutions are proposed in the literature and can be classified as follows [21]:

1. *Access class barring* (ACB) based schemes: these solutions aim at minimizing congestion occurrences by mapping each device to an appropriate access class (AC) with a certain barring time duration, according to the type of application considered.

2. Methods relying on the separation of RACH resources: reserving a certain number of preambles to H2H communications can help mitigate access overloading caused by M2M communications, even though this implies a reduced amount of resources available for MTDs.

3. Dynamic allocation of RACH resources: an efficient technique to alleviate access congestions can be to offer additional resources every time the system is close to the saturation. Nevertheless, providing further resources to MTDs implies a reduced amount of resources for data transmission.

4. *Coded-Expanded Random Access* (CERA) schemes: a promising solution to increase the number of devices allowed to access

the network consists in grouping RA slots in *virtual* frames representing the access opportunities. Furthermore, the preamble sequences are replaced by codewords that are transmitted over the PRACH. However, this method turns out to be less energy efficient than the previous ones, since it requires higher energy to perform the coded expanded access.

As previously stated, Tab. 2.1 presents for each scenario a potential set of technology solutions, according to the specific application in use. For instance, environmental monitoring may consist of sensor devices with heterogeneous traffic patterns. An ACB scheme can then reduce potential congestions by mapping each device into an appropriate AC. On the other hand, military applications may sometimes demand more resources and a dynamic allocation approach can better operate in this context.

2.2.2 Radio resource allocation for M2M communications

Once the devices have obtained access to the network, available resources are to be efficiently distributed according to the specific M2M application involved. First of all, to allow the simultaneous presence of H2H and M2M communications, there is the need to design an efficient resource sharing mechanism. Moreover, each type of MTD can be characterized by different QoS requirements and fulfilling all the needs is still an open challenge.

Unlike H2H communications, where most of the requested services generate multimedia traffic and introduce timing constraints, M2M communications have different QoS requirements, being the transmitted packet rate not uniform (e.g., a sensor device may perform a measurement and generate a packet with a certain rate). For this reason, applying a radio resource allocation approach with the aim of satisfying delay or jitter requirements turns out to be an inadequate solution. In addition, M2M traffic is typically transmitted by using a small amount of resources that can be generated by a

massive number of devices within the same LTE-A cell. Accordingly, several MTDs may use a certain number of resource blocks (RBs) in the same subframe and how to arrange every transmission can be a complicated task.

Existing approaches that simultaneously allocate resources for M2M and H2H communications can be summarized in *orthogonal* and *shared channel* allocation methods [22, 23]. In the former approach, interference is minimized by transmitting human and machine traffic data on orthogonal channels, even though a decreased spectral efficiency is expected. As an example, authors in [24] design a M2M contention-free allocation scheme based on a pool of resources periodically reoccurring in time, which is dimensioned so as to provide a certain reliability in the delivery of a report subject to a deadline. Conversely, shared channel allocation maximizes the spectral efficiency by allowing the resource sharing between M2M and H2H communications at the expense of a higher interference level. Therefore, several solutions introduced in literature rely on the Cognitive Radio (CR) paradigm, opportunistically employing licensed bands to transmit M2M traffic without affecting H2H cellular communications.

2.3 D2D communications

Among all the features recently introduced in LTE Rel. 12, D2D communications represent the fundamental technology to enable the P2P data exchange between mobile users, without accessing to the cellular network. This not only allows to reduce the consumed power and the delay due to a D2D pairing procedure performed by two CUs, but also to extend the overall network capacity and offload the cellular network via a direct link, thus alleviating the RAN congestion. Nevertheless, enabling direct connectivity also leads to higher interference within a cellular network if the resources are shared between cellular and D2D communication modes, as in the case of D2D communications underlaying cellular systems. In the following, we briefly discuss two of the major challenges related to D2D communications

and describe promising solutions proposed in literature.

2.3.1 Radio resource allocation for D2D communications

Similarly to the M2M-H2H resource allocation problem, D2D communications need to be performed by sharing resources with traditional cellular communications, where CUs transmit data towards the eNB. Resource allocation modes can be classified as follows [18]:

- *Cellular mode*: D2D user traffic is routed toward the BS which is in charge of delivering it to the receiving device.
- *Orthogonal sharing mode*: A certain amount of resources is reserved for CUs and the remaining portion is utilized to support D2D communications.
- *Non orthogonal sharing mode*: Both D2D and cellular communications are established by using the same resources at the same time, thus causing interference.

It is also worth pointing out that how to select the most suitable mode is still an open issue and several techniques have been proposed that take into account the received signal strength on the D2D link, the distance between the devices and the potential interference produced [25, 26].

2.3.2 Proximity discovery

To establish a direct communication, a D2D user needs to obtain information on D2D candidates, i.e., devices able to act as peers and to pair with the user requiring communication. Similarly to traditional cellular communications, a suitable search procedure is necessary in order to achieve time-frequency synchronization. As described in [27], peer discovery can be performed by adopting two different approaches:

1. *Non-assisted approach.* A device is in charge to send a beacon message as reference signal for the receiver synchronization. Since no coordination is provided by a central unit, the peer discovery phase needs to be designed with the aim of selecting the device transmitting the beacon message. Furthermore, the beaconing procedure leads to a higher energy consumption and thus to a shorter battery life.

2. *Network assisted approach.* The cellular infrastructure can help to increase the energy efficiency by supporting the discovery phase, e.g., by selecting the best D2D candidate users.

In the latter case, the network entity can also decide when the D2D communication shall be established, depending on whether a-priori or a-posteriori criterion is used.

A-priori-based techniques enable the P2P communication if the candidate devices are not already exchanging data through the eNB. Specifically, before establishing a communication between two users, the network identifies potential D2D candidates by means of a client-server procedure, where the D2D client intending to activate a D2D communication must send a request to the network and wait for the beacon message transmitted by a D2D server. On the other hand, an a-posteriori approach allows the seamless transition from a traditional cellular communication to the direct communication. In more detail, whenever a session is activated between two users and the devices are in the proximity of each other, the network can trigger the switch from the cellular mode to the D2D mode. Once a D2D candidate *pair* is selected, the D2D server can send a beacon message directly to the D2D client, allowing the network to acquire information on the channel quality experienced by the beacon receiver.

2.4 Proposed Protocol

In this section, an integrated framework is proposed to support an opportunistic networking scheme as overlay of an LTE-A system. One of

the major challenges is how to extend D2D communications towards the multi-hop capability widely adopted in ad-hoc networks. To this end, network-assisted solutions can support network establishment and maintenance by identifying potential D2D users and providing more resources to nodes acting as relay, depending on the network configuration.

2.4.1 System Model

Consider a single LTE-A cell network scenario where H2H and M2M communications share the same resources. We assume that each device is characterized by different degree of mobility. Fig. 2-1 shows a possible scenario where LTE-A devices are represented by both mobile phones and vehicles served by the eNB. We assume all the devices capable of operating both in cellular mode, i.e., by communicating with the outside through the eNB, and in P2P mode, i.e., by directly exchanging data traffic, without involving the eNB. In such a scenario, a typical application may be the delivery of temperature and humidity samples by generic devices (*sensing nodes*) to a specific collector device which gathers and processes the information received. Furthermore, due to transmit power constraints, we assume no direct communication between sensing nodes and the collector node, resulting in the need of a multi-hop communication scheme. In other words, D2D users located in a certain area should be selected as *relay* nodes to forward data to the destination and the *relay* selection can be performed by the eNB or by the D2D users, according to a centralized or distributed approach, respectively. An explicit routing scheme is also applied: this entails that the *relay* selection is not opportunistically performed, but a routing algorithm selects the best route toward a destination.

Besides, we consider an *overlay* resource allocation, where resources dedicated to cellular and D2D users are orthogonal either in time or in frequency, avoiding co-channel interference between the two systems. This also implies that a certain amount of resources is reserved for D2D traffic and cannot be utilized by CUs. For the sake

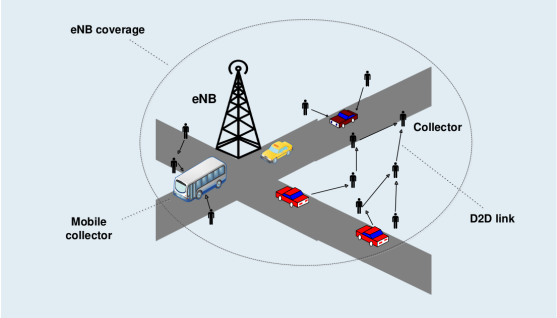


Figure 2-1: A typical scenario with opportunistic D2D communications overlaying an LTE-A network

of simplicity, we then assume that D2D communications can be performed only during the uplink frame. Accordingly, only the BSs will be affected by the interference introduced by D2D users: a CU will not receive interference as during the uplink phase every CU will be transmitting traffic data over the uplink channel. Finally, we consider a TDD half-duplex communication scheme, thus enabling downlink and uplink phases in the same carrier frequency but in different time intervals.

2.4.2 Network configuration and management

Our approach consists of three distinct phases: (i) a D2D cluster discovery phase, where the eNB identifies potential D2D users in a certain area, while taking into account possible services to be accomplished, (ii) a graph formation phase, dealing with the selection of the best routes towards the collector node, and (iii) a resource allocation algorithm aware of the D2D network configuration previously established.

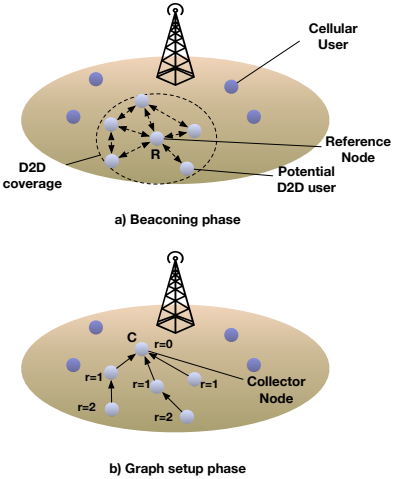


Figure 2-2: Illustration of the discovery and graph formation phases.

D2D cluster discovery

As mentioned in the previous section, a network assisted approach is a valid solution to support the discovery phase as the selection of the D2D candidates can be easily performed by the network entity. Fig. 2-2a shows a single cell scenario where a cluster consisting of D2D users is identified. We assume that the eNB is able to process the requirement and to identify a cluster of potential D2D users, including collector and sensing nodes. Once a D2D cluster is discovered, specific authentication procedures are applied in order to allow direct communication between devices in LTE-A networks. To analyze the D2D connectivity, a beacon broadcasting phase is necessary, since every device in the cluster needs to obtain channel state information (CSI) about the other cluster members. Such an operation is executed by allowing each node in the cluster to broadcast in turn a beacon message. By overhearing the beacon signal, the quality of all the D2D links can be achieved and then exploited to build con-

nectivity graph. Further, by processing the beacon message, a D2D user can also identify the collector node located in the cluster and obtain information related to the pattern of data traffic associated with potential services.

Graph construction

Once the eNB has obtained CSI of all the D2D nodes in the cluster, network topology can be derived according to a graph-theoretic approach. Specifically, a network is considered *connected* if all the node pairs can be connected by means of multi-hop path. In the case of wireless communications, connectivity is affected by channel conditions, so that end-to-end communication performance decreases as the number of hops in a path increases. Efficient multi-hop communications can be analyzed by means of a connectivity graph, where each vertex represents a D2D user and each edge denotes a direct link between two cluster members. We assume that an edge exists only when the Signal-to-Noise-Ratio (SNR) is higher than a certain threshold in order to avoid links with the worst channel conditions to take part in the graph construction. We also consider the resulting graph as a *directed acyclic graph* (DAG), where no loops are present. Traditional solutions to build and maintain the connectivity graph are based on routing schemes that select the best route from a source node to a destination node, according to the minimum number of hops or to the cost to reach an intermediate node.

To this end, Routing Protocol for Low-power and Lossy networks (RPL) [28] is proposed as possible solution to compute a convenient route toward a given destination. Because of its simplicity and effectiveness, RPL is widely adopted for Wireless Sensor Networks (WSN) and represents a valid candidate for our networking scheme. Following the paradigm of gradient routing protocols, RPL defines a Destination Oriented Directed Acyclic Graph (DODAG) by means of an objective function in combination with a set of constraints or metrics. Each node is assigned a *height* or *rank*, represented by a scalar value that increases as the distance from a collector node increases.

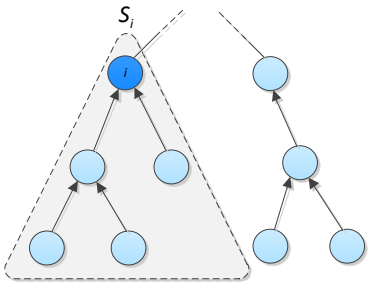


Figure 2-3: An example of RPL graph construction

Figure 2-2b shows the network structure build by an RPL protocol. A cumulative cost function is used to compute the distance according to the selected metric, e.g., hop count, node residual battery or a certain combination of metrics. Finally, by exchanging control messages, every node chooses as next-hop the neighbor characterized by the largest gradient. Moreover, the loop-free characteristic of directed acyclic graphs is preserved since each node always has a higher rank than all its parent nodes.

Network-aware resource allocation algorithm

In such a network configuration, a serious drawback is represented by potential bottlenecks involving nodes with a low rank value[1]. As previously described, the rank defines the relative position of a node in the DODAG with respect to the root, i.e., the collector node. Accordingly, nodes characterized by low ranks are generally close to the collector node and thus more prone to bottlenecks, due to the increasing amount of traffic to forward to destination. An effective resource allocation strategy should take into account the rank of each node, providing more resources to the nodes closest to the collector.

[1]This problem is commonly addressed by means of the *min-cut analysis*.

As shown in Figure 2-2b, D2D users with rank 1 are directly connected to the collector C and are in charge of delivering to C both traffic locally generated and data transmitted by nodes with higher ranks. By notifying the eNB with its own rank, every node can be allocated a sufficient number of RBs to allow the multi-hop communication. To this end, a *naive* resource allocation strategy can then assign resources according to a proportional fair approach, by setting scheduling weights for each node to $\frac{r_i}{\sum_j r_j}$, where r_i is the estimated rate related to node i. Furthermore, by exploiting information about the connectivity graph, the overall rate R_i associated to each vertex i in the graph is derived as follows:

$$R_i = r_i + \sum_{j \in \mathcal{S}_i} r_j \tag{2.1}$$

where $\mathcal{S}_i$ is the sub-tree of i, as highlighted in Figure 2-3.

Once the eNB has obtained the overall rates, the resource allocation algorithm assigns to each vertex a weight equal to $\frac{R_i}{\sum_j R_j}$, allowing a more accurate and efficient resource management.

2.5 Conclusions

In this chapter we proposed a novel approach to enabling wireless networking functionalities in cellular M2M communications. A multi-hop scheme is investigated in order to minimize additional power consumption and interference. Further, a possible integrated framework comprising network and service discovery, as well as a resource allocation strategy, is introduced to seamlessly deliver data traffic. By exploiting D2D communications overlaying LTE-A systems, CUs can directly exchanging data without affecting normal cellular communications.

Chapter 3

Aggregation and Trunking of M2M traffic via D2D Connections

3.1 Introduction

One of the main issues associated with cellular M2M communications is how to provide connectivity to a large number of devices densely deployed within a cell and how to avoid the RAN overloading due to simultaneous access requests. Following the principle of capillary networks, we introduce a novel technique based on the concept of localized packet aggregation, where CUs help to transmit machine traffic through D2D connectivity established with nearby devices.

In literature, packet aggregation techniques have been proposed in diverse network architectures and systems. As an example, data aggregation and compression are extensively applied in clustered Wireless Sensor Networks (WSN), in order to reduce the energy consumption and increase the battery life time of the sensor nodes [29]. In the

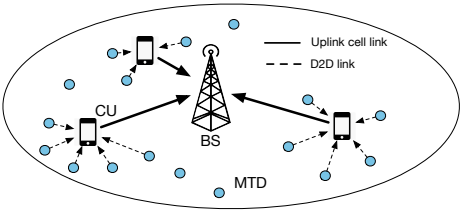


Figure 3-1: MTDs and CUs in a single cell network.

context of M2M, a random access scheme is investigated in [30], where sensor devices and data collectors are randomly deployed within a cell, while an approach to group a number of machines into a *swarm* to alleviate the RAN overload and reduce the number of connections between the devices and the BS is presented in [31]. This also allows to aggregate traffic packets originated from low-power MTDs and dispatch them to a BS, preventing the network access to overload and accommodating devices characterized by poor communication links [12]. However, relay-based schemes implicitly assume the presence of helper nodes, which not only increase the network operational expenditure, but also consume some of the radio resources to transmit the aggregated traffic to the BS, thus leading to the degradation of the network performance [32].

Inspired by the paradigm of *Trunked Radio Systems* [33], where limited radio resources are shared among a large set of users, we propose an access protocol that can potentially mitigate access overload and exploit the benefits of D2D. As shown in Fig. 3-1, packets generated by several MTDs are collected by a nearby CU through D2D links, which is in charge of aggregating and delivering the traffic to a BS. Our protocol consists of an access reservation phase, where the machines contend for access, and a Time Division Multiple Access (TDMA) scheme, where the time is divided in slots and each previously granted device is allocated a Time Slot (TS). After aggregating

the packets and adding its own data, the mobile device transmits to the BS by adapting the power and the transmission rate to the channel conditions as well as the actual amount of data that need to be sent. We show that there is a fundamental trade-off between latency and power required for the uplink transmission in an M2M scenario consisting of a large number of machines. We compare our technique with a traditional cellular access system, where the machines access directly to the BS, and we thereby demonstrate the power benefits of the trunking scheme.

The remainder of this chapter is organized as follows. Section 3.2 describes the adopted system model, while the proposed approach is introduced in Section 3.3. In Section 3.4 a detailed analysis of the protocol is conducted and the numerical results are presented in Section 3.5. Conclusions are drawn in Section 3.6.

3.2 System Model

Our system consists of a single cell in which a BS B provides cellular access to a subscribed user U, and an M2M network composed by several low-power MTDs connected to U via D2D communication. The time is divided in frames with duration $L \cdot T$, where L is the fixed number of slots in the frame and T is the slot duration. D2D communication takes place in a subset of TSs from a frame allocated to the user U, such that it can be characterized as overlay/in-band D2D [34]. As shown in Fig. 3-2, the MTDs are located in a proximity area around U, which corresponds to the maximum D2D coverage range, such that a D2D communication session can be established between U and each nearby MTD. For simplicity, we suppose that all the machines are at the same distance x_m from U and transmit at fixed power P_m, whereas the distance U-B is fixed to x_U. The arrival process for machine-type data at the MTDs is modeled according to the Poisson distribution with arrival rate of λ packets per second. We assume that each MTD can generate at most one packet of D_m bits per frame, such that the packet arrivals to the population

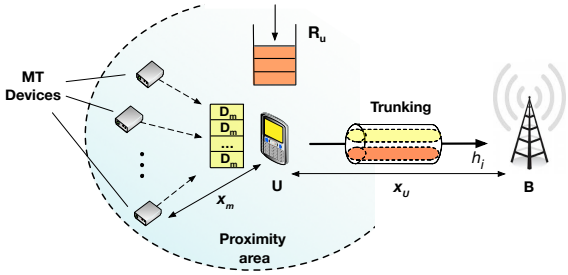


Figure 3-2: Illustration of the system model.

of MTDs are distributed at different MTDs. To evaluate the performance of the trunking scheme, we also suppose that the cellular uplink is constantly used by U to transmit its own data at a constant bit rate of R_u. In the sequel, U is assumed to be in charge of: *i*) gathering the machine-type traffic received through $D2D$ links, *ii*) aggregating the MTDs data with its own packets, and *iii*) forwarding the aggregated traffic to B via the cellular uplink. Furthermore, all the packets must be transmitted within a time interval equal to the frame duration $L \cdot T$, which represents the packet deadline.

All devices are equipped with a single antenna at both the transmitter and receiver side and U has full and instantaneous channel state information (CSI) of the link towards B[1], allowing U to perform uplink power control. We assume that all links are characterized by a block fading channel, where the channel state remains constant over the frame period and the fading realization follows a Rayleigh distribution with Probability Density Function (PDF), f_h, defined

[1]Note that the absence of full CSI leads to the degradation of the uplink performance and consequent increase in the power consumption. The study of this kind of scenario is left for future work.

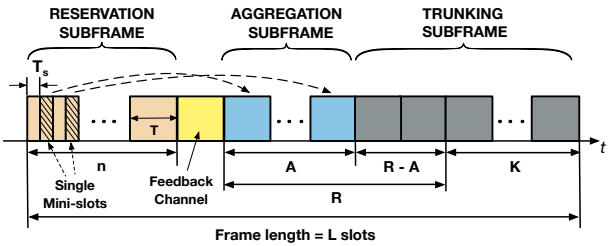


Figure 3-3: Frame structure of the proposed MAC protocol.

as,

$$f_h(u) = \frac{1}{\overline{h}} \exp\left(-\frac{u}{\overline{h}}\right), \tag{3.1}$$

where $\overline{h}$, the mean channel gain associated with the small scale fading, is assumed to be $\overline{h} = 1$. We define the SNR from the j^{th} transmitter, γ_j, as,

$$\gamma_j = \frac{P_j h_j x_j^{-\alpha} K_D}{\sigma^2}, \tag{3.2}$$

where P_j is the transmission power of the j^{th} node, x_j is the distance between the receiver and the j^{th} node, α is the path loss exponent, K_D is the path loss constant and σ^2 is the noise variance. The list with all the mathematical symbols used in the chapter is given in Table 3.1.

3.3 Proposed Framework

Both D2D and uplink cellular communication take place within a TDMA frame of length L, as illustrated in Fig. 3-3. One-shot transmission scheme is employed, such that a packet that is not successfully received is dropped, i.e. not present in the next frame. The frame

is divided into three orthogonal phases: reservation, aggregation and trunking, described in the sequel.

Reservation

The first portion of the frame is dedicated to an access reservation phase that allows the MTDs to use random access and indicate their need for data transmission. A TS is divided into a fixed number of reservation mini-slots R of duration T_s, such that n, the number of TSs allocated during this phase, is:

$$n = \lceil R \cdot \frac{T_s}{T} \rceil, \tag{3.3}$$

where R is the number of reservation mini-slots. The access reservation procedure uses *Framed Slotted ALOHA* [35], where an MTD with a data packet to transmit, selects randomly and uniformly one of the available reservation mini-slots to transmit its reservation token. When multiple MTDs select to transmit in the same mini-slot, we assume that a destructive collision occurs, which leads to none of the reservation tokens being detected in that mini-slot. At the end of this phase, a mini-slot is declared as reserved only if a single MTD has selected it and the associated transmission is strong enough to be detected by the receiver. In other words, even if there is a single MTD that selects the mini-slot, the channel can be in outage such that the reservation token is not decoded and U does not grant data access to the corresponding MTD. The next slot in the frame, as depicted in Figure 3-3, is reserved for the transmission of the necessary feedback to the MTDs through a channel with very robust coding, so that any errors occur with very low probability.

Aggregation

The aggregation phase, as depicted in Figure 3-3, is where the D2D communication between the MTDs that were granted access and U takes place. The number of slots given for data transmission to the

Sym	Definition	Sym	Definition
λ	MTD arrival rate	D_m	MTD packet payload
λ_f	Avg no. MTDs per frame	D_u	User packet payload
Λ_i	No. MTDs in frame i	h	Channel power gain
p_d	Prob of successful D2D TX	$P_{U,i}$	U-B req. TX power
P_O	Outage probability in U-B	T_s	Mini-slot duration
L	Frame length	T	Slot duration
K	No. trunking slots	μ	TCI cutoff parameter
R	No. mini-slots	Γ_m	m-U SNR threshold
A	No. scheduled D2D slots	$R_{i,a}$	Aggregated data rate
n	No. access reservation TSs	W	System bandwidth
P_{mtd}	MTD TX power TSs	K_D	Path loss factor
x_j	Receiver-node j distance	σ^2	Noise variance
γ_j	Receiver-node j SNR	a	No. succ. mini-slots

Table 3.1: Mathematical Notation

MTDs is $A \leq R$, where A corresponds to the number of successful reservations, excluding the ones that would lead to outage in the data transmission part. The system has R preallocated slots for MTD data transmission, such that when A is known, $R - A$ slots are returned back to U for uplink transmission. Once all the scheduled machines have completed the D2D transmission, U aggregates the machine packets with its own data.

Trunking

The last phase corresponds to the trunking phase, where the cellular uplink is used as a trunk to deliver both the received machine traffic and the user data to B. The device U uses power control in the uplink, based on the Channel State Information at the Transmitter (CSIT) as well as the total amount of aggregated data to transmit, as described in Sec. 3.4.3. Finally, we note that the number of slots in the trunking phase is equal to $K + (R - A)$, where K is the fixed

number of reserved trunking slots, and $R-A$ is variable and depends on the number of slots not utilized in the aggregation phase. In conclusion, a single TDMA frame consists of a fixed number of slots L, divided as follows:

$$L = n + R + K + 1. \tag{3.4}$$

3.4 Analysis

3.4.1 Reservation

A successful aggregation occurs when (a) the reservation slot has been selected by a single MTD and (b) the channel MTD-U is sufficiently strong not to lead to outage. Given R mini-slots, the probability to have a successfully reserved mini-slots is

$$\Pr\left(A = a|R\right) = \sum_{s=1}^{R} \Pr\left(A = a|s\right) \Pr\left(S = s|R\right), \tag{3.5}$$

where $\Pr\left(S = s|R\right)$ is the probability of having s single mini-slots:

$$\Pr(S = s|R) = \sum_{m=0}^{\infty} \Pr(S = s|m, R)\ \Pr(\Lambda_i = m), \tag{3.6}$$

where $\Pr(\Lambda_i = m)$ is the probability of having m MTD arrivals in the i^{th} frame. Let λ be the MTD arrival rate per second and $\lambda_f = E[\Lambda_i] = \lambda LT$ be the average number of MTDs transmitting at the beginning of i^{th} frame. Then Λ_i, assuming stationarity and independence from the channel states, is given by a Poisson distributed process with arrival rate λ_f and associated probability mass function,

$$\Pr(\Lambda_i = m) = \frac{\lambda_f^m e^{-\lambda_f}}{m!}, m \geq 0. \tag{3.7}$$

In (3.5), the probability $\Pr(A = a|s)$ to have a successful reservations given s single mini-slots is

$$\Pr(A = a|s) = \binom{s}{a} p_d (1 - p_d)^{s-a}, \tag{3.8}$$

where p_d is the probability that the reservation token is successfully decoded. Assuming Γ_M as the minimum SNR required for the transmission from an MTD to be decodable, then,

$$p_d = 1 - \Pr(\gamma_m < \Gamma_m) = e^{-\Gamma_m \frac{\sigma^2}{P_{mtd} \bar{h}_i x_m^{-\alpha} K_D}}. \tag{3.9}$$

To compute $\Pr(S = s|m, R)$, we follow the approach described in [36]. The process of distributing MTDs into mini-slots can be modeled as a bins and balls problem, with distinguishable bins and balls. In this context, the bins and the balls are represented by the mini-slots and the MTDs, respectively, and the number of single mini-slots corresponds to the number of bins with occupancy number equal to 1. By taking into account the results presented in [36], the probability of having exactly s out of R mini-slots given m contending MTDs is:

$$\Pr(S = s|m, R) = \frac{\binom{R}{s} \prod_{k=0}^{s-1} (m - k) G(R - s, m - s)}{R^m}, \tag{3.10}$$

where[2]

$$G(u, v) = u^v + \sum_{t=1}^{v} (-1)^t \prod_{j}^{t-1} [(v - j)(u - j)](u - t)^{v-t} \frac{1}{t!}. \tag{3.11}$$

To ease the computation of (3.6), we provide the following approximation. Assuming that the probability of an MTD choosing a random mini-slot out of R mini-slots is $1/R$, the mean number of contending MTDs per mini-slot is equal to λ_f / R. Thus, the probability of a

[2]We define $\binom{n}{k} = 0$ for $n < k$.

single MTD selecting a mini-slot can be approximated as,

$$p_s \approx \frac{\lambda_f}{R} e^{-\frac{\lambda_f}{R}}. \tag{3.12}$$

Now, by assuming the amount of MTDs in each mini-slot to be independent, the probability that s out of R mini-slots are single is,

$$\Pr(S = s|R) \approx \binom{R}{s} p_s^s (1 - p_s)^{R-s}, \tag{3.13}$$

where $(1-p_s)^{R-s}$ is the probability that none of the remaining $R-s$ mini-slots is single, and $\binom{R}{s}$ is the number of ways this slot can be selected. This approximation becomes tighter with the increase in the number of arrivals in the frame.

3.4.2 Aggregation

After the access reservation phase, each of the a accepted MTDs are allowed to transmit via the D2D link to U, through their own dedicated data slot. As a result, the aggregated data rate in the i-th frame R_i becomes,

$$R_{i,a} = \frac{D_u + aD_m}{T(K + R - a)}, \tag{3.14}$$

where D_m is the MTD packet payload, while D_u is the user payload, generated every frame time period LT with rate R_u as $D_u = LTR_u$.

3.4.3 Trunking

As mentioned in Sec. 3.3, we consider a cellular uplink with adaptive rate allocation, where the user transmitter is able to dynamically adjust the power depending on the data rate and on the channel conditions. Assuming AWGN channel and capacity-achieving codes, the data rate in the i-th frame R_i is related to the transmitting power

$P_{U,i}$ by the Shannon's capacity equation:

$$R_{i,a} = W \log_2(1 + \frac{P_{U,i} h_i x_j^{-\alpha} K_D}{\sigma^2}), \tag{3.15}$$

from which $P_{U,i}$ is obtained as,

$$P_{U,i} = \left(2^{\frac{R_{i,a}}{W}} - 1\right) \frac{\sigma^2}{h_i x_j^{-\alpha} K_D}, \tag{3.16}$$

which is the transmit power needed to sustain the data rate R_i over the channel during the i-th frame. To compensate for the variation of the channel between U and B, we consider a Truncated Channel Inversion (TCI) policy [37], where the channel fading is inverted only if the fade depth is above a given cutoff value μ. Therefore, the average transmit power required to sustain the data rate $R_{i,a}$ is derived as follows:

$$\begin{aligned} \mathbb{E}[P_{U,i}|R_{i,a}] &= \int_{\mu}^{\infty} P_{U,i}\ f_h(x)dx \\ &= \left(2^{\frac{R_{i,a}}{W}} - 1\right) \frac{\sigma^2}{\overline{h}_i x_U^{-\alpha} K_D} E_1\left(\frac{\mu}{\overline{h}}\right), \end{aligned} \tag{3.17}$$

where μ is the cutoff parameter representing the minimum value of channel fade depth that can be compensated and $E_1(\cdot)$ is the exponential integral function. Accordingly, the related outage probability is given by:

$$P_O = \Pr(h < \mu) = \int_0^{\mu} f_h(x)dx = 1 - \exp\left(-\frac{\mu}{\overline{h}}\right). \tag{3.18}$$

Note that, to guarantee a certain outage probability, we can conveniently select the value of μ, as follows,

$$\mu = -\bar{h} \log(1 - P_O). \tag{3.19}$$

3.5 Numerical Results

3.5.1 Performance Metrics

We now provide the analytical expressions for the performance metrics which are examined in the next section. As first metric, we consider the expected number of MTDs served per second, defined as:

$$\mathbb{E}[N] = \frac{1}{L \cdot T} \sum_{a=0}^{R} \Pr\left(A = a|R\right), \tag{3.20}$$

where p_d denotes the probability of the data slot being decodable by the receiver and is obtained from (3.9), while $\Pr\left(A = a|R\right)$ is the probability of having successfully reserved a mini-slots given R mini-slots, obtained from (3.5).

Next, the average transmit power per served machine $\mathbb{E}[P_m]$ is considered, which consists of the power needed to reserve a TS P_{res} plus the power to transmit the data packet P_{agg} and the power to send the machine packet through the trunk uplink channel $\mathbb{E}[P_{tr}]$:

$$\mathbb{E}[P_m] = P_{res} + P_{agg} + \mathbb{E}[P_{tr}] = 2P_{mtd} + \mathbb{E}[P_{tr}], \tag{3.21}$$

where $P_{res} = P_{agg} = P_m$. Depending on whether the MTDs are connected to U or B, we assume $P_{mtd} = P_{m,U}$ or $P_{mtd} = P_{m,B}$ respectively, while $\mathbb{E}[P_{tr}]$ is given by,

$$\mathbb{E}[P_{tr}] = \sum_{a=0}^{R} \mathbb{E}[P_{U,i}|R_{i,a}] \Pr\left(A = a|R\right). \tag{3.22}$$

Finally, we define the probability of a user being served as the prob-

Parameter	Value	Parameter	Value
P_O	0.01	T_s	0.1 ms
T_s	0.1 ms	σ^2	-97 dBm
R_u	100 Kbps	W	180 KHz
K_D	-30 dB	D_m	100 bits
Γ_m	-3 dB	x_m	10 m
T	1 ms	$\overline{h}$	1
x_U	200 m	α	3
$P_{m,B}$	18 dBm	$P_{m,U}$	-20 dBm

Table 3.2: Simulation Parameters

ability of being the only one to select a given mini-slot, given m arrivals, and of the reservation token being detected in the reservation phase and the trunking link not being in outage:

$$P_S = (1 - P_O) \sum_{m=0}^{\infty} p_d \left(1 - \frac{1}{R}\right)^{m-1} \Pr(\Lambda_i = m). \tag{3.23}$$

3.5.2 Simulation Results

To numerically analyze the performance of our protocol, we developed a MATLAB-based simulator, implementing the functionalities described in Sec. 3.2 and assuming the parameters shown in Table 3.2. Simulations are based on a Monte Carlo approach, where each point corresponds to the average value of 10^5 iterations. We first examine the performance metrics by increasing λ, in order to observe the impact of different network access volumes on the system. Furthermore, we analyze the average power spent to serve a single machine and compare our approach with a typical cellular system, where the

distance between MTDs and B is fixed to x_U and all the machines communicate with B, using a fixed power level $P_m = P_{m,B}$ and without relying on the trunk link. It is also worth pointing out that in our system the latency is only given by the frame length L, which depends on the parameters R and K, as shown in (3.4).

Fig. 3-4 shows the average number of machines served $\mathbb{E}[N]$ as a function of λ. Taking as reference the curve with parameters $K = 1$ and $R = 10$, we can observe that the number of selected single mini-slots increases and reaches its maximum value at $\lambda \approx 800$. Then, the impact of the collisions becomes evident and $\mathbb{E}[N]$ decreases. Moreover, choosing $R = 10$ corresponds to assigning $n = 1$ slot to the access subframe, which increases the protocol overhead as well as the system latency. We note that a higher K significantly affects the system performance, since a longer frame results in a larger number of contending machines due to the aforementioned Poisson arrivals assumption. This, in turn, leads to an increased number of collisions in the reservation subframe and a consequent decrease of $\mathbb{E}[N]$. In addition, it is worth noticing that the expected value of the total power transmitted over the U-B link is directly related to $\mathbb{E}[N]$, as the more machine packets are aggregated, the higher power is required to sustain the resulting data rate. Therefore, the maximum required transmit power can be obtained by considering values of λ which maximize $\mathbb{E}[N]$, according to R and K.

The trade-off between latency and transmit power is illustrated in Fig. 3-5. First of all, we notice a significant difference in terms of consumed power between our scheme and a system without trunk link, due to the fact that in the latter case the MTDs transmit at fixed power and no dynamic power allocation is allowed. Secondly, if we consider the trunk link approach with $\lambda = 250$ and choose $R = 20$, the power consumption is approximately reduced by 50% with respect to the case of $R = 10$. In other words, by increasing R, not only the number of served machines increases, but also the number of available trunking slots grows, since we assumed the system able to reuse the portion of the frame not occupied by the D2D transmissions,

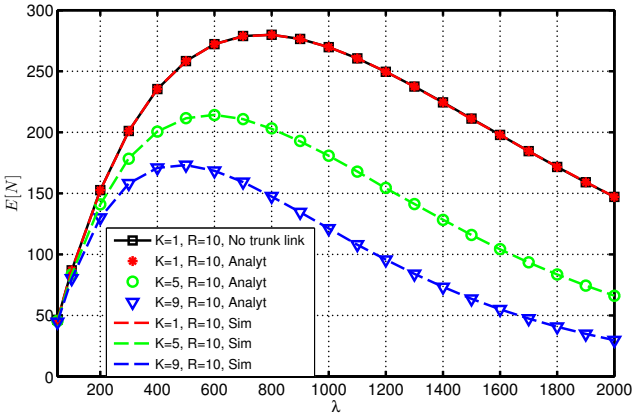


Figure 3-4: Average number of served MTDs for different K and R.

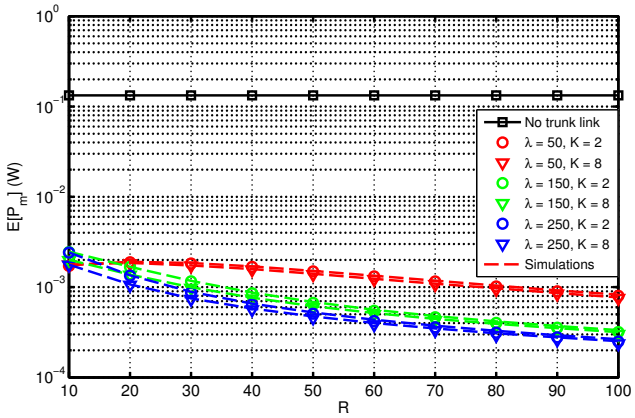


Figure 3-5: Average transmit power per served MTD for different λ and K.

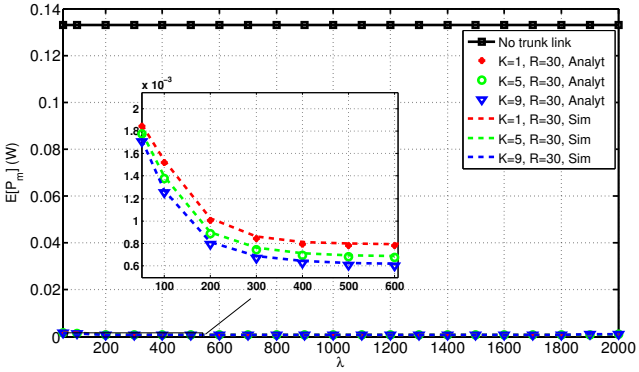


Figure 3-6: Average transmit power per served MTD for different K and R.

i.e., $R - A$. In contrast, the latency of the system becomes higher, as we need to assign more slots to the reservation subframe, and this also implies a significant overhead increment due to the required reservation mini-slots.

Fig. 3-6 shows the mean transmit power spent per MTD as a function of λ. As previously noticed, the power spent to transmit a single machine packet in the case of trunk link not available is fixed and does not depend on the number of contending MTDs. On the other hand, if the machines utilize the trunk link provided by U, $\mathbb{E}[P_{tr}]$ decreases as λ becomes higher and turns out to be considerably lower than the case of trunking not enabled. To evaluate the impact of different number of trunk slots, we also included a zoomed in section illustrating $\mathbb{E}[P_{tr}]$ in the case of low load access conditions. More specifically, we can notice how the power decreases as λ increases, since a higher number of machines are allowed to transmit, and the minimum value corresponds to the value of λ maximizing $\mathbb{E}[N]$.

Finally, the outage probability of an MTD requiring access is shown in Fig. 3-7. In case of low access load (i.e., $\lambda < 100$), the outage

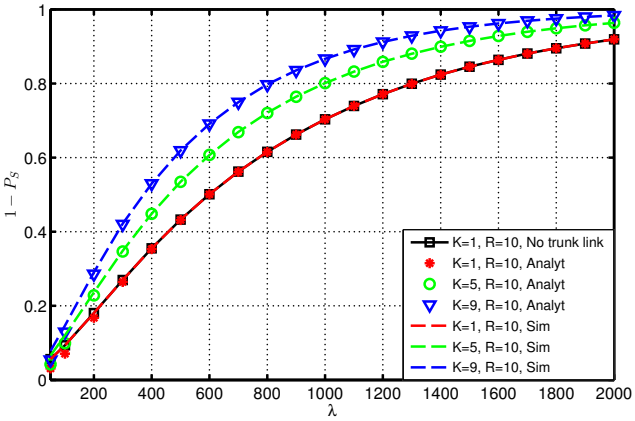


Figure 3-7: Machine outage probability for different K and R.

probability is below 0.2 and a higher K does not affect significantly the system performance. By increasing λ, the outage probability becomes higher due to the large number of access attempts causing collisions in the reservation phase. As expected, for higher access load (i.e., $\lambda > 400$), the outage probability is clearly higher if we assume $K > 1$ and this also implies a lower number of machines served with respect to the case of a more energy-consuming system characterized by $K = 1$.

3.6 Conclusions

We proposed a new solution to alleviate the impact of the massive number of MTDs on a cellular system by exploiting D2D communications. Our TDMA-based MAC scheme guarantees the simultaneous delivery of packets generated by the MTDs and by a user device, without involving any relay or helper node. The analytical and simulation results show that there is a fundamental trade-off between

the latency and the transmit power in the case of Poisson-distributed machine arrivals. By increasing the frame length, we can achieve a remarkable reduction of the transmit power, even though the system latency increases and a low delay packet requirement cannot be fulfilled. Moreover, our scheme can significantly reduce the average amount of power spent to sustain an MTD in comparison with a system without trunk link, guaranteeing the same user throughput.

Chapter 4

M2M Traffic Offloading onto Trunked Cellular Uplink via D2D Connections

4.1 Introduction

The main idea described in this chapter is to mitigate the RAN overload by adopting a trunking-based scheme [38], where MTDs are grouped into local clusters lead by a CU, as shown in Fig. 4-1. These CUs then gather MTT through short-range D2D links and deliver it to the attached BS, without any user QoS degradation. Unlike other relay-based solutions that utilize unlicensed short range radio [12], we use the same cellular interface in a direct communication mode, thus exploiting licensed spectrum reliability, security and easy integration with the wide-area cellular connection [34]. Furthermore, we provide a direct comparison between this approach and a traditional machine

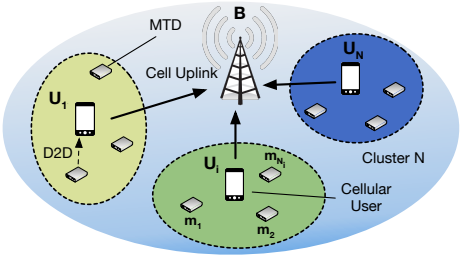


Figure 4-1: Single cell network scenario: MTDs associated with CUs for data offloading.

cellular access, in terms of throughput, average power consumption and latency per MTD packet delivered. As first proposed in [38], the trunking scheme allows to: *i*) forward the MTT without relying on any deployed relay or helper node, thus reducing network design and maintenance costs, *ii*) avoid interference with other systems by transmitting over the operator's licensed spectrum, and *iii*) enhance the energy efficiency, as the MTC data are sent over low power D2D links.

To handle the simultaneous machine access, we design a contention-based MAC protocol, which coordinates the D2D links established between a set of close MTDs and a target CU. More specifically, the machines employ a Reservation Framed S-ALOHA (RFS-ALOHA) [39] protocol to reserve a slot for data transmission, whenever a new machine packet arrival occurs. Although CSMA/CA protocols are largely adopted in mesh networks, ALOHA-based protocols are particularly suitable for MTDs characterized by very low computational complexity in single-receiver topologies as considered in this chapter. As the terminals do not perform collision detection and idle listening, energy consumption and signaling overhead are considerably reduced.

The main contributions of this chapter, which go beyond the work presented in [38], are the following:

- An enhanced trunking protocol that aggregates the MTT via

access reservation, incorporating a fixed-window backoff algorithm to shape the incoming MTT, and then trunks the aggregated traffic to the cellular BS;

- An analytical model of the proposed protocol based on a discrete time Markov chain, which captures the stochastic behavior of the MT data aggregation, within a finite population setting;
- An evaluation of the trunking gains in terms of throughput, latency and energy efficiency, achieved when employing multiple CUs and partitioning the set of machines into multiple clusters.

The rest of the chapter is organized as follows. Section 4.2 illustrates recent approaches that address the RAN congestion. In Section 4.3, the system model is presented, explaining the network topology, the channel model and the traffic model. The proposed protocol is introduced in Section 4.4, where we describe each phase involving the offloading of the machine packets onto the CU. In Section 4.5, our analytical model is illustrated and numerical results are examined in Section 4.6. Finally, Section 4.7 concludes the chapter.

4.2 Related work on RAN overload mitigation

To address the RAN congestion due to massive MTT issue in LTE/LTE-A systems, access-class-barring (ACB) and extended access barring (EAB) schemes have been proposed and investigated within 3GPP [40–42]. In the EAB, the MTDs are classified as delay tolerant and are randomly assigned to a class group by the network operator, being able to access the network only if their respective class is not currently blocked by the EAB mechanism. This is equivalent to blocking the incoming traffic with a certain probability. Following the ACB/EAB approach, the Fast Adaptive S-ALOHA (FASA) [43] protocol tracks

the network status by monitoring the access attempts and exploits this information to support highly synchronous traffic originated from event-driven MT communications. To provide QoS and solve the RAN overload, authors in [44] have proposed the Prioritized Random Access with Dynamic Access barring (PRADA) scheme, where physical random access channel (PRACH) resources are preallocated for different MTD classes, with class-dependent backoff procedures, and massive access requests are mitigated by employing dynamic access barring. However, system reliability is greatly affected by the presence of synchronous machine arrivals due to the recurrent activation of the EAB algorithm.

Hierarchical schemes aim at alleviating the RAN congestion by grouping the MTDs or by employing relay nodes to forward the MTT. In [45], specific network improvements are introduced for scenarios where MTDs act as gateways for capillary networks of heterogeneous devices. The work in [12] proposes an energy efficient client relay scheme based on IEEE 802.16 networks, where MTC data packets are assumed sufficiently short to be sent during the random access requests, thus reducing control (or signaling) overhead. The authors focused on a typical smart metering MTC scenario and have analyzed protocol performance in the case of power outage (i.e. where a large number of devices report failure almost simultaneously). However, the deployment of relay nodes implies higher operational costs and may also reduce the amount of available radio resources due to the additional relay links to the BS. In [46] MTDs are grouped according to similar mobility patterns, with the aim of reducing the signaling overhead associated with mobility management, and only a selected MTD is in charge of performing the signaling on behalf of its group, without increasing the overall energy consumption.

Cognitive Radio (CR) has been considered an attractive technology to accommodate MTT. The deployment of cluster head (CH) nodes to collect MTT, exploiting an underlay Cognitive Radio (CR) technology in cellular networks, was proposed in [13]. Specifically, a hierarchical network structure is envisioned, where MTDs connect

to CH nodes via secondary CR links, which then offload the MTT onto the BS, leveraging traditional cellular links. As a result, the number of cellular accesses is expected to be drastically reduced if the underlay MTC network is optimally dimensioned. The Cognitive M2M paradigm is conveniently applied in [47], where the authors introduce a centralized cognitive MAC protocol and a specific frame structure to foster the coexistence with the primary network, able to handle both periodic and event-driven machine data. Furthermore, an optimization framework is applied based on the tradeoff between primary network protection and QoS requirements of the secondary network, while the performance is evaluated in a smart metering scenario. In [48] the authors consider MTC relaying with dynamic spectrum access in white spaces for both links (MTD-relay and relay-BS) and analyze the end-to-end delay when the relay node is assigned a higher access probability. Nevertheless, MTDs are typically low-cost and low-complexity and cannot support advanced functionalities to perform spectrum sensing and data collection.

Exploiting the CR paradigm, *cooperative* M2M communications assume autonomous grouping of the MTDs into clusters and choose relay nodes to convey the MTT to the BS, without interfering with the primary cellular system. In [49], MTDs are assumed to be capable of forming coalitions and acting as relays with the aim of preventing network congestion and increasing energy efficiency. Specifically, a non-transferable utility (NTU) coalitional game is formulated to model the coalition formation for relay transmission and the performance is compared with the direct transmission, as the number of MTDs joining the coalition increases. An opportunistic transmission protocol based on CR technology and cooperative communications between MTDs is presented in [50]. The MTDs can autonomously perform spectrum sensing and relay selection in large-scale MTC networks, guaranteeing statistical QoS, and key network properties, i.e., network connectivity, degree distribution and network diameter, are investigated by means of social network analysis.

MTC communications may also take advantage of proximity dis-

covery features provided by D2D communication mode, defined in 3GPP Rel.12 [34]. Novel solutions have been recently proposed to offload user data traffic exploiting direct connectivity in licensed bands [51], e.g., LTE-A, and in the unlicensed Wi-Fi spectrum [52]. The work in [31] groups the MTDs into swarms either logically, according to similar characteristics and QoS requirements, or physically, leveraging the D2D connectivity. In [53] the authors provide a comparative study of power control strategies applicable to D2D communications, although not focusing on machine type D2D. Inspired by the group-based MTC operation, [54] proposes an uplink scheduling algorithm to increase the overall system throughput, where a selected MTD can relay traffic originated from devices belonging to the same group through short-range D2D communications. Network energy efficiency enhancement is addressed in [55], where a multi-hop hierarchical scheme is applied to aggregate MTT generated by wireless devices, which can act as a transmitter or as an aggregator, and an energy density optimization framework is developed by using stochastic geometry tools. Authors in [56] and [57] analyze the impact of underlay low rate MTC communications within cellular systems, assuming D2D connectivity between MTDs and CU and Successive Interference Cancellation (SIC).

4.3 System Model

Here we describe the network setup, the wireless channel model and the traffic model. A list with all the mathematical symbols used in the chapter is shown in Table 4.1.

4.3.1 System Topology

Consider the single cell network illustrated in Fig. 4-1, where a set of N CUs U_i, $\forall i = 1, \ldots, N$, are attached to the BS B. We assume that U_i is located in the proximity of a number M_i of homogeneous low-power MTDs $m_j, \forall j = 1, \ldots, M_i$, such that the entire network can be

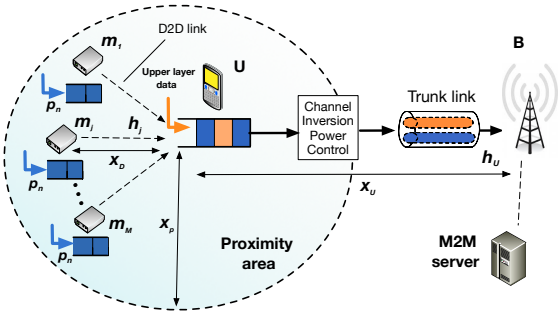


Figure 4-2: Structure of a cluster consisting of MTDs deployed around the generic CU U.

partitioned in N non-overlapping clusters, being U_i the cluster-head of the cluster i. For notation simplicity, we refer to the generic CU U by omitting the index i.

As shown in Fig. 4-2, we assume that a 2-hop network can be realized between the MTDs and B. The m_j–U links are established by employing a D2D overlay/in-band scheme [34], where D2D and cellular links occupy separate radio resources of the licensed spectrum and there is no mutual interference.

We assume that the cellular network manages the D2D connections, by providing synchronization, peer discovery and authentication/security functionality. A D2D proximity area is represented by a circle with center U and radius x_p, such that each MTD lying within this region may establish a D2D connection with U and join its cluster. We then assume that U is located at distance $x_U \gg x_p$ from B and communicates to it via the cellular link U–B. For analytical simplicity, we assume that all the nodes are static and that each MTD, belonging to the same cluster, is located at the same distance x_D from U.

Symbol	Parameter
N	Total number of CUs
M_i	Total number of MTDs in cluster i
p_n	MTD activation probability
Γ_m	SNR threshold of m–U links
L	Frame length
T	Slot duration (s)
T_s	Mini-slot duration (s)
Q	Number of trunking slots
R	Number of mini-slots
w	Reservation subframe duration in no. slots (see (4.3)))
R_u	User data rate (bps)
D_m	MTD packet size (bits)
x_D	Distance between MTDs and U (m)
x_U	Distance between U and B (m)
$\bar{h}$	Average channel gain
σ^2	Noise variance
α	Path loss exponent
$P_{m,U}$	MTD transmit power (m-U link)
$P_{m,B}$	MTD transmit power (m-B link)
P_U	CU transmit power
p_d	Probability of successful D2D TX (see (4.8)))
p_t	Probability of successful slot reservation (see (4.11))
p_r	Probability of reservation token (re)transmission (see (4.17))
π_I	Steady-state prob of MTD being idle (see (4.12))
$\pi_{i,j}$	Steady-state prob of counter being j and BO stage i
K	Retransmission cutoff parameter
W	Max backoff window size
$R_{ag}(a)$	Resulting data rate given a aggregated MTDs (see (4.25))
P_O	U–B channel outage probability (see (4.29))
$E[\tau]$	Average system throughput (see (4.31))
P_S	Probability of successful reservation (see (4.32))
$E[P]$	Avg power spent to deliver a packet to B (see (4.33))
$E[D]$	Avg delay for a successfully delivered pkt (s) (see (4.37))

Table 4.1: Mathematical Notation

4.3.2 Wireless Channel Model

We assume the channel gains to be characterized by the path-loss and the small-scale Rayleigh fading, ignoring any contribution from the large-scale fading. Let us denote the small-scale fading contribution by h, having the following Probability Density Function (PDF):

$$f_h(u) = \frac{1}{E[h]} \exp\left(-\frac{u}{E[h]}\right), \tag{4.1}$$

where the mean value is assumed to be $E[h] = 1$. We assume block fading, such that the channel state remains constant over a frame period L. The SNR for the i^{th} wireless link can be expressed as:

$$\gamma_i = \frac{P_i h_i x_i^{\alpha}}{\sigma^2}, \tag{4.2}$$

where, α is the path loss exponent, σ^2 is the noise variance, and P_i is the transmit power of the i^{th} node. Finally, we assume that the machines transmit with fixed power $P_{m,U}$, whereas U has full channel state information (CSI) of the link toward B and can thus apply uplink power control.

4.3.3 Traffic Model

We assume the system time to be divided into fixed length frames of L non-overlapping TSs. Two types of traffic are delivered to B via U in each frame: (i) data originated from the U upper layers with constant bit rate R_U, and (ii) packets received from the MTDs through the D2D links. These data flows are processed and combined at U, and the resulting aggregate data flow sent to B, which then retrieves and forwards the MTT to an external MTC server. We assume that U is equipped with a transmit buffer, sufficiently large to accommodate both traffic types.

At the beginning of a frame, each *idle* machine, i.e., not in a process of retransmission, generates a single packet of size D_m bits and

attempts access with probability p_n. If at the first try the machine does not complete successfully the access attempt, then it will backoff and attempt retransmission in the following frames. The machine will continue to reattempt access until it is either successful or it reaches the maximum number of allowed retransmissions K, after which the packet is dropped. We assume that each machine can only hold one packet in its own buffer; therefore a new packet arrival will only occur after the previous one has either been transmitted or dropped.

4.4 Proposed Protocol

We now describe each step involving the delivery of MTD packets to B during a frame. In what follows, we assume no centralized admission control algorithm, e.g., ACB/EAB, applied to limit the number of machines requesting access. In other words, each MTD in the proximity area is allowed to offload data onto U every time a new packet is generated. We also introduce the baseline and the multi-cluster configuration, which will be used to evaluate the trunking gains in Sec. 4.5.5.

4.4.1 Machine Access Reservation

An Immediate First Transmission (IFT) scheme is assumed, where an incoming MTD packet is transmitted in the next frame with probability equal to one. Each MTD indicates its need for a data transmission, by randomly selecting one of the reservation mini-slots within the reservation subframe. Each TS is divided into a fixed number of reservation mini-slots R of duration T_s, such that the reservation subframe has w mini-slots, where

$$w \triangleq \lceil R \cdot T_s / T \rceil. \tag{4.3}$$

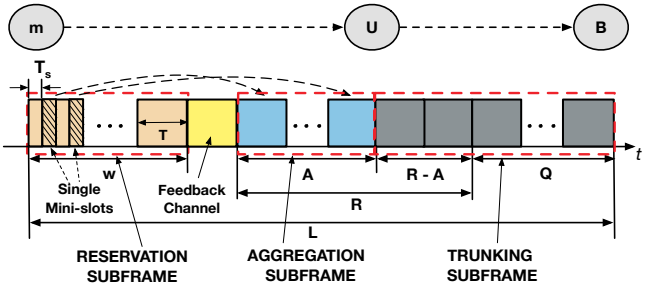
The access reservation scheme is based on the *Framed S-ALOHA* protocol [35], where an MTD selects randomly and uniformly one of the

R available reservation mini-slots to transmit a reservation token. In this sense, Framed S-ALOHA is similar to a traditional multi-channel S-ALOHA system with a single random access channel, such as in the case of the LTE access reservation protocol [58]. In our setting, the mini-slots replace the orthogonal random access preambles used in LTE for activity detection. In order to model a reservation failure events, we introduce the following assumptions:

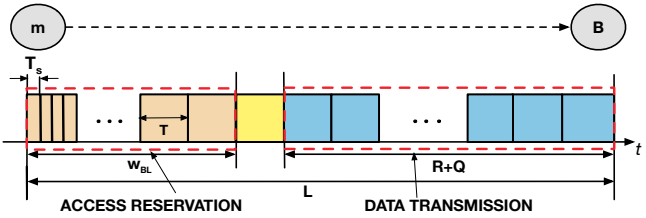
1. Multiple transmissions during the same mini-slot lead to a destructive collision, i.e., capture effect is not considered in our model.

2. A token transmitted over a single selected mini-slot may not be decoded due to poor channel condition. We model this event by p_d - the probability of successfully transmitting a token on a non-collided mini-slot .

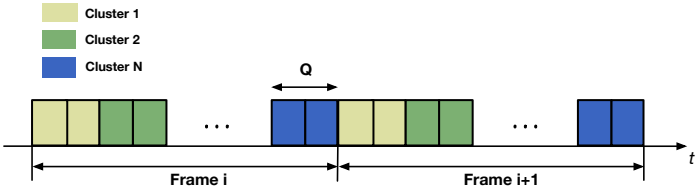
The next slot in the frame, as shown in Figure 4-3a, is reserved for sending feedback to the MTDs. The feedback message is assumed to be very robust, with negligible probability of error. To mitigate the access congestion on U, MTDs employ a Fixed-Window Backoff (FWB) algorithm, where congestion events are prevented by simply applying a backoff window of fixed size W and by limiting the number of retransmissions to K. After a failed token transmission, a backoff counter is uniformly selected in the range $[0, W-1]$ and is decremented by one at the end of each frame. Every time the counter reaches 0, the reservation token is retransmitted in the beginning of the next frame. If the number of unsuccessful access attempts exceeds K, the buffered data packet gets dropped and the MTD returns to the idle state.



(a)



(b)



(c)

Figure 4-3: Frame structure for single-cluster scenarios with DSA (a), baseline scheme (b), and periodic trunking slot allocation for multi-cluster scenarios with FSA (c) ($Q = 2$). For the sake of clarity, we also include each node (m, U and B) involved in the communication.

4.4.2 Aggregation and Trunking

In the data aggregation phase, see Figure 4-3a, the D2D communication between the granted access MTDs and U takes place[1]. Let us denote the total number of slots assigned to the MTDs by $A \leq R$, which corresponds to the number of MTDs holding a reservation token. Assuming that R slots are always preallocated for MTD data transmission, we consider two different slots allocation policies:

1. *Fixed slot allocation* (FSA) policy, where the R slots are solely employed for the data aggregation phase.
2. *Dynamic slot allocation* (DSA) policy, where $R - A$ not used slots are returned back to U for the uplink transmission occurring in the subsequent subframe.

Once all scheduled MTDs have completed the D2D transmission, U aggregates the MTC packets with the data coming from the upper layers. The last subframe is devoted to trunking, where the cellular uplink is used as a trunk to deliver both the received MTT and the user data to B. The device U applies power control in the uplink, based on the Channel State Information at the Transmitter (CSIT) as well as the amount of data to transmit. Moreover, the number of slots contained in the trunking subframe is equal to $Q+(R-A)\cdot I_D$, where Q is the fixed number of reserved trunking slots, $R - A$ is variable and dependent on the number of slots not utilized in the previous subframe, and I_D is the policy indicator function, which is equal to 1 or 0 depending on whether DSA or FSA is enabled, respectively. In conclusion, a single TDMA frame consists of a fixed number of slots L, divided as follows:

$$L = w + R + Q + 1, \tag{4.4}$$

[1]When the MTD is able to successfully reserve access, i.e. when the MTD is the only one to select the reservation mini-slot and the channel is good enough for the reservation transmission to be decoded, then the subsequent D2D data transmission is also successfully decoded due to the assumption of block fading over the frame duration.

where one slot is added to account for the feedback channel, as shown in Fig. 4-3a.

4.4.3 Baseline and multi-cluster configuration

To evaluate the proposed protocol, we consider a direct access baseline scheme, whose frame structure is illustrated in Fig. 4-3b. Specifically, we assume that the number of slots for data transmission is $R + Q$, while the access reservation subframe consists of $w_{BL} = \lceil (R + Q) \cdot T_s / T \rceil$ slots, thus assigning all the available resources to the MTDs and neglecting the CU traffic. In other words, the MTDs select one of the $R + Q$ available mini-slots and transmit a packet to B over an assigned data slot, without relying on the CU support. We also point out that L is fixed and equal to the trunking protocol frame length, thus allowing a fair comparison between the two schemes when no CU traffic is served.

Furthermore, to provide a full characterization of the trunking gains we investigate the protocol performance in a multi-cluster scenario, for which a periodic trunking slot allocation is considered as shown in Fig. 4-3c. To avoid co-channel interference, we need to design the system allowing the coexistence among all the trunking devices operating on the same carrier frequency[2]. For this reason, the communication between CUs and BS is temporally separated exploiting the TDMA-based frame structure, i.e., each cellular link is preassigned an orthogonal channel for data transmission. More specifically, all the CUs are synchronized and can perform the trunking operation only during the preassigned Q trunking slots, which are distributed along the entire frame, as illustrated in Fig. 4-3c. This implies that FSA is the only suitable policy, as DSA would permit cellular transmission on not occupied D2D slots, thus generating co-channel interference in the trunking link. Finally, we assume that in the multi-cluster scenario the reservation and aggregation phases

[2]We assume the interference from the MTDs on the other clusters to be negligible, due to the very low transmission power.

take place in a channel orthogonal to the one where the trunking phase takes place. In other words, the reservation and aggregation phases do not suffer co-channel interference generated by the high power trunking transmissions from the other clusters. Moreover, the co-channel interference generated from the reservation and aggregation transmissions from nearby clusters is assumed to be negligible due to the low power transmissions.

4.5 Analysis

4.5.1 Access Reservation

The access reservation procedure can be modeled as shown in Fig. 4-4(a). We denote m as the number of machine packet arrivals observed by U in a frame,

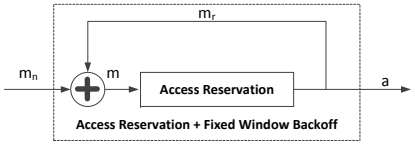
$$m = m_n + m_r \tag{4.5}$$

which is the sum of the arrivals of newly activated MTDs m_n and the backlogged MTDs m_r. We denote a as the number of successfully reserved TSs. Given R available mini-slots and m contending MTDs, the probability of having $a \leq R$ successfully reserved TSs in the aggregation subframe is computed as follows:

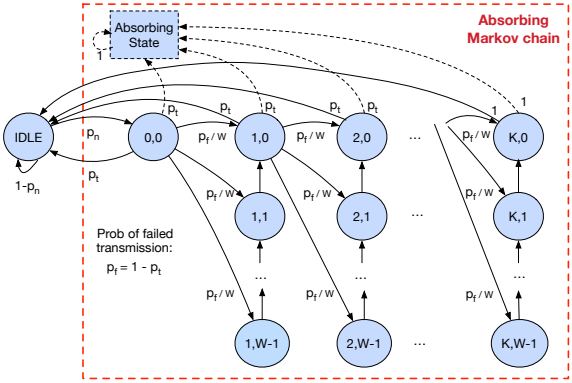
$$\begin{aligned} \zeta(a, m) &\triangleq \Pr\left(A = a | R, m\right) \\ &= \sum_{s=a}^{R} \Pr\left(A = a | s\right) \Pr\left(S = s | R, m\right), \end{aligned} \tag{4.6}$$

where $\Pr\left(A = a|s\right)$ is the probability that a successful reservations occur given s single mini-slots, i.e.,

$$\Pr\left(A = a|s\right) = \beta(s, a, p_d), \tag{4.7}$$



(a)



(b)

Figure 4-4: Analytical model: (a) Combination of new arrivals and backlogged traffic triggering the access reservation procedure; (b) MTD state transition diagram for the Markov chain of the FWB algorithm. An absorbing Markov chain is also obtained by considering $p_n = 0$ and the related state transitions are represented by dashed arrows all converging on the absorbing state.

where $\beta(x, y, z) = \binom{y}{x} z^x (1-z)^{y-x}$ denotes the binomial identity[3] and p_d is the probability that the reservation token is successfully decoded. The latter can be computed assuming Γ_m as the minimum

[3]We define $\binom{n}{k} = 0$ for $n < k$.

SNR required to successfully decode a token, as follows:

$$p_d \triangleq 1 - \Pr(\gamma_m < \Gamma_m) = \exp\left(-\Gamma_m \frac{\sigma^2}{P_{m,U}\ \bar{h}\ x_D^{-\alpha}}\right). \tag{4.8}$$

Furthermore, the probability of having exactly s single mini-slots selected, given R mini-slots and m contending MTDs, is given by [36]:

$$\Pr(S = s|m, R) = \frac{\binom{R}{s} \prod_{k=0}^{s-1}(m-k) G(R-s, m-s)}{R^m}, \tag{4.9}$$

where

$$G(u, v) \triangleq u^v + \sum_{t=1}^{v}(-1)^t \prod_{j}^{t-1}[(v-j)(u-j)] \frac{(u-t)^{v-t}}{t!}. \tag{4.10}$$

It is also worth noticing that $\zeta_{a,m} = 0$ if $a = R < m$, since R single mini-slots cannot be selected if the number of contending machines m is higher than R.

4.5.2 Fixed Window Backoff

We now characterize the Fixed Window Backoff (FWB) scheme, where the backlogged state of each user is modeled by a Markov chain as depicted in Fig. 4-4(b). Similar to [58], we start by introducing p_t the probability of a user successfully completing the access reservation procedure,

$$p_t = \sum_{n=0}^{M-1} \beta(n, M-1, p_r) \sum_{k=0}^{n} \frac{(k+1)\zeta(k+1, n+1)}{n+1}, \tag{4.11}$$

where M is the total number of MTDs in the network, p_r is the probability of a user attempting (re)transmission and the inner sum corresponds to the normalized throughput conditioned on the simul-

taneous transmission of $n+1$ MTDs[4]. We note that although the Markov chain models the state of a single MTD, the impact of the other MTDs in the MTD access is captured within p_t and p_r.

Assuming the system to be stationary, we introduce the probability that an MTD is idle as π_I, and the probability that its backoff counter value is j during the ith backoff stage as $\pi_{i,j}$. Assuming that the probability of an MTD in idle state becoming active is given by p_n, then π_I is defined as,

$$\pi_I = (1-p_n)\pi_I + \sum_{i=0}^{K-1} p_t \pi_{i,0} + \pi_{K,0}, \tag{4.12}$$

where K is the maximum number of allowed retransmissions and $\pi_{0,0} = p_n \pi_I$. $\pi_{i,j}$ is defined as,

$$\pi_{i,j} = \frac{1-p_t}{W}\pi_{i-1,0} + \pi_{i,j+1}, \tag{4.13}$$

for $0 < i \leq K$, $0 \leq j < W-1$, where W is the window size. Noting that

$$\pi_{i,W-1} = \frac{1-p_t}{W}\pi_{i-1,0},$$

then $\pi_{i,j}$ can be rewritten as,

$$\pi_{i,j} = \frac{W-j}{W}(1-p_t)^i p_n \pi_I. \tag{4.14}$$

To obtain π_I, we recur to the following normalization condition,

$$\pi_I + \pi_{0,0} + \sum_{i=1}^{K}\sum_{j=0}^{W-1} \pi_{i,j} = 1. \tag{4.15}$$

[4]The $n+1$ MTDs term comes from considering n concurrent MTDs to the MTD being observed.

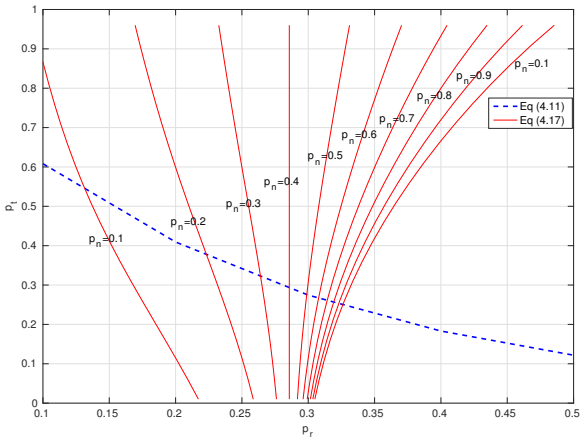


Figure 4-5: Intersections between equations (4.11) and (4.17), $M = 40$, $p_d = 0.9$, $R = 10$.

After manipulation, π_I is given by:

$$\pi_I = \left[1 + p_n \left(1 + \frac{(W+1)(p_t - 1)\left[(1-p_t)^K - 1\right]}{2p_t}\right)\right]^{-1}. \quad (4.16)$$

The packet (re)transmission probability p_r, is obtained from,

$$p_r = \sum_{i=0}^{K} \pi_{i,0} = \frac{1-(1-p_t)^{K+1}}{p_t} p_n \pi_I. \quad (4.17)$$

The valid solutions of p_r, which satisfy both (4.11) and (4.17) can be obtained by numerically finding the intersection points between these two functions. As shown in Fig. 4-5, as p_n increases, the (4.11) intersects the (4.17) at a single point, which can be interpreted as the system stability point. We now compute the average access delay

$E[D_A]$, experienced by an MTD that has successfully reserved access. We start by altering the state transition model in Fig. 4-4(b), by introducing an absorbing state. The resulting absorbing Markov chain is represented by the red dashed box, where the dashed arrows denote all the state transitions toward the single absorbing state. The $E[D_A]$ is computed by averaging the number of state transitions required to reach the absorbing state, when starting from the initial state $(0,0)$. Let us first introduce $D_{i,j}$ as the time in access frames, for an absorbing process starting from the initial backoff counter j in the i^{th} retransmission to reach the absorbing state. The corresponding expectation $E[D_{i,j}]$ is computed as,

$$E[D_{i,j}] = E[1 + D_{i,j-1}] = E[j + D_{i,0}] = E[D_{i,0}] + j, \tag{4.18}$$

where j indicates the number of frames the backoff counter needs to reach the state $(i,0)$. The mean time required to reach the absorbing state starting from the state $(0,0)$ is,

$$\begin{aligned} E[D_{0,0}] &= p_t \cdot 1 + (1-p_t)\frac{1}{W}E\left[\sum_{j=0}^{W-1}(1+D_{i,j})\right] \\ &\overset{(a)}{=} p_t + (1-p_t)\left[E[D_{1,0}] + \frac{W+1}{2}\right] \end{aligned} \tag{4.19}$$

where (a) comes from,

$$\begin{aligned} \frac{1}{W}E\left[\sum_{j=0}^{W-1}(1+D_{i+1,j})\right] &= \frac{1}{W}\sum_{j=0}^{W-1}(1+E[D_{i+1,j}]) \\ &= \frac{1}{W}\sum_{j=0}^{W-1}(1+E[D_{i+1,0}]+j) \\ &= E[D_{i+1,0}] + \frac{W+1}{2}. \end{aligned}$$

The corresponding mean time required to reach the absorbing state starting from the state $(i, 0)$ is

$$E[D_{i,0}] = p_t + (1 - p_t)\left[E[D_{i+1,0}] + \frac{W+1}{2}\right] \tag{4.20}$$

for $0 \leq i < K$. Setting the boundary condition $E[D_{K,0}] = 1$, then $E[D_{0,0}]$ yields,

$$E[D_{0,0}] = 1 + \frac{W+1}{2} \cdot \frac{(1-p_t) - (1-p_t)^{K+1}}{p_t}, \tag{4.21}$$

which includes the delay contribution of devices that were able to reserve access successfully, denoted as $E[D_A]$, and devices that were dropped due to exceeding the maximum allowed K retransmission attempts, $E[D_D]$. We can separate these two contributions as follows,

$$E[D_{0,0}] = (1 - (1-p_t)^K)E[D_A] + (1-p_t)^K E[D_D]. \tag{4.22}$$

The average delay experienced by a packet successfully delivered to U, becomes,

$$E[D_A] = \frac{E[D_{0,0}] - (1-p_t)^{K+1}E[D_D]}{1 - (1-p_t)^{K+1}}. \tag{4.23}$$

Thus, the average delay experienced by a dropped packet $E[D_D]$ is defined as,

$$E[D_D] = 1 + K \cdot E\left[\frac{1}{W}\sum_{j=0}^{W-1}(1+j)\right] = 1 + K \cdot \frac{W+1}{2}. \tag{4.24}$$

Fig. 4-6 shows the average access delay (left y-axis) and the average throughput (right y-axis) as the backoff window increases and for different values of M. We observe that lower delays can be obtained with a smaller backoff window W. However, small values of W imply

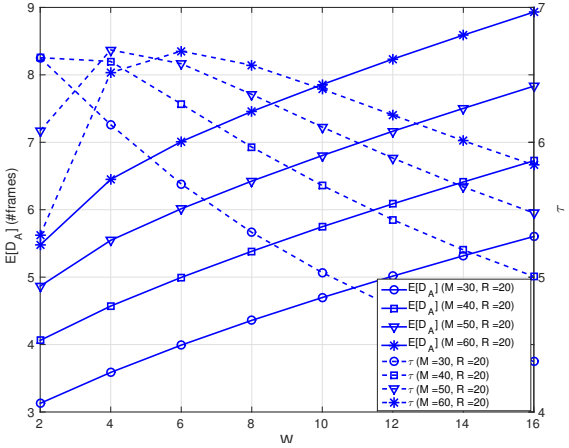


Figure 4-6: $E[D_A]$ (left y-axis) and τ (right y-axis) vs W for different values of M and R ($K = 8$, $p_n = 0.8$).

lower throughput when M is higher than R (e.g., $M = 40, 50$), due to an increase in the number of collisions.

4.5.3 Aggregation and Trunking

At the end of the access reservation phase, each of the a accepted MTDs is scheduled by U and can transmit over a dedicated data slot in the aggregation subframe. As a result, the aggregated data rate in a frame, given a MTD packet receptions, R_{ag}, can be expressed as follows:

$$R_{ag}(a) = \frac{D_u + aD_m}{T(Q + I_D(R - a))}, \tag{4.25}$$

where D_m is the MTDs packet payload and D_u is the user payload, which is generated in every frame period LT with rate R_u, i.e., $D_u = LTR_u$.

As mentioned in Sec. 4.3.2, an adaptive rate allocation scheme is assumed to take place in the link between U and B, such that the transmit power can be dynamically adjusted by U, depending on the data rate R_{ag} and on the channel conditions. Assuming AWGN channel and capacity-achieving codes, the data rate R_{ag} and the transmit power P_U are coupled by the Shannon's formula as follows:

$$R_{ag}(a) = B_W \log_2(1 + \frac{P_U h_U x_U^{-\alpha}}{\sigma^2}), \tag{4.26}$$

where B_W is the system bandwidth. Hence, we can obtain P_U as follows:

$$P_U = \left(2^{\frac{R_{ag}(a)}{B_W}} - 1\right) \frac{\sigma^2}{h_U x_U^{-\alpha}}, \tag{4.27}$$

which is the transmit power needed to sustain the data rate R_{ag}. To compensate for the variation of the channel conditions, we consider a Truncated Channel Inversion (TCI) policy [37], where the channel fading is inverted only if the fade depth is above a given cutoff value μ. Therefore, the average transmit power needed to sustain R_{ag} is derived as follows:

$$\begin{aligned}\mathbb{E}[P_U|R_{ag}(a)] &= \int_{\mu}^{\infty} P_U \ f_h(x)dx \\ &= \left(2^{\frac{R_{ag}(a)}{B_W}} - 1\right) \frac{\sigma^2}{E[h_U]x_U^{-\alpha}} E_1\left(\frac{\mu}{E[h]}\right),\end{aligned} \tag{4.28}$$

where μ is the cutoff parameter, representing the minimum value of channel fade depth that can be compensated, and $E_1(\,\cdot\,)$ is the exponential integral function. Accordingly, the related outage probability P_O is given by:

$$P_O \triangleq \Pr(h < \mu) = \int_0^{\mu} f_h(x)dx = 1 - \exp\left(-\frac{\mu}{E[h]}\right), \tag{4.29}$$

which denotes the probability that the channel gain h is smaller than μ, and, consequently, the probability that U's transmission is suspended during the frame i. To further guarantee a certain outage probability P_O, we can conveniently select the value of μ as follows:

$$\mu = -\bar{h}\log(1 - P_O). \tag{4.30}$$

4.5.4 Performance Metrics

We now provide the analytical expressions for the different considered performance metrics.

Throughput and Success Probability

The throughput of our scheme $E[\tau]$ corresponds to the number of MTD packets aggregated per frame, which is given by,

$$E[\tau] = \sum_{n=0}^{M}\sum_{j=0}^{n} j\zeta(j,n)\beta(n,M,p_r), \tag{4.31}$$

where p_r includes the probability of the MTD being either in the first transmission or in the sub-sequent retransmissions; and the success probability given K allowed retransmissions is given by,

$$P_S = 1 - (1 - p_t)^{K+1}. \tag{4.32}$$

Transmission Power

The mean total power $E[P]$ spent to send an MTD packet to B is,

$$E[P] = E[P_R] + E[P_A] + E[P_U], \tag{4.33}$$

where P_R accounts for the power used by the MTD until it succeeds in the reservation phase,

$$E[P_R] = P_{m,U} \sum_{i=1}^{K+1} i \cdot p_t (1-p_t)^{i-1} \tag{4.34}$$
$$= P_{m,U}(1-p_t)^K \frac{Kp_t^2 + p_t^2 - Kp_t - 1 + 1/(1-p_t)^K}{p_t},$$

with $P_{m,U}$ denoting the fixed transmission power used in the reservation attempt, and P_A denotes the power used by the MTD to transmit its packet in the aggregation phase, i.e.,

$$E[P_A] = P_{m,U}. \tag{4.35}$$

Finally, P_U denotes the power employed by U to transmit the aggregated traffic divided by the number of MTD packets sent:

$$E[P_U] = \sum_{m=1}^{M} \left[\sum_{a=1}^{\min(m,R)} \frac{\mathbb{E}[P_U | R_{ag}(a)]}{a} \zeta(a,m)) \right] \beta(m, M, p_r). \tag{4.36}$$

Latency

The mean latency experienced by a packet generated by an MTD and successfully delivered to B is computed as follows:

$$E[D] = TL \cdot \left(E[D_A] + E[D_U] \right), \tag{4.37}$$

where $E[D_A]$ denotes the delay, in number of frames, experienced by an MTD packet successfully delivered to U, and $E[D_U]$ represents the delay in the trunking due to the link outage, which is given by the expected value of a random variable following the geometric distribution with parameter $1 - P_O$, i.e.,

$$E[D_U] = \frac{1}{1-P_O} - 1, \tag{4.38}$$

where the subtracted one frame is due to the trunking transmission occurring within the same frame where the MTD packet aggregation took place, whenever the trunking link is not in outage.

4.5.5 Trunking Gains

We conclude this section by modeling the performance of our protocol on a multi-cluster scenario, where multiple CUs camped on the same cell can employ the trunking mechanism in order to forward the MTT toward the BS. To this end, we define the trunking gains in terms of throughput, energy efficiency and latency obtained by partitioning the set of M MTDs into N clusters of size M_i[5] with respect to the case of direct access, i.e., no trunking device is employed.

We define the overall throughput obtained by employing N different and independent trunk links as:

$$E_N[\tau] = \sum_{i=1}^{N} \sum_{n=0}^{M_i} \sum_{j=0}^{n} j\zeta(j,n)\beta(n,M,p_r). \tag{4.39}$$

Then, the trunking gain associated with the throughput $\mathcal{G}_\tau(N)$ is given by:

$$\mathcal{G}_\tau(N) \triangleq \frac{E_N[\tau]}{E_{BL}[\tau]}, \tag{4.40}$$

where $E_{BL}[\tau]$, the throughput of the baseline access scheme, is given by,

$$E_{BL}[\tau] = \sum_{n=0}^{M} \sum_{j=0}^{n} j\Pr\left(A = a | R+Q, m\right) \beta(n,M,p_r),$$

where all the MTDs perform access reservation and data transmission toward B according to the frame structure shown in Fig. 4-3b.

To obtain the trunking gain $\mathcal{G}_P$ associated with the consumed

[5]For simplicity, we assume homogeneous clusters with $M_i = \frac{M}{N}, \forall i = 1, ..., N$ MTDs.

power, we first need to define the energy efficiency η (bits/W):

$$\eta \triangleq \frac{D_m}{E[P]}, \tag{4.41}$$

which expresses the overall amount of bits delivered to the BS per Watt. Therefore, we define $\mathcal{G}_P$ as the ratio between the energy efficiency of a trunking scheme with $N \geq 1$ and the baseline scheme, i.e.,

$$\mathcal{G}_P(N) \triangleq \frac{\eta_N}{\eta_{BL}} = \frac{E_{BL}[P]}{E_N[P]}, \tag{4.42}$$

where $E_N[P]$ represents the average power consumed to transmit an MTD packet, given N active trunk links, and $E_{BL}[P] = E[P_R^{BL}] + P_{m,B}$ indicates the average power spent to deliver an MTD packet using the baseline scheme, where $E[P_R^{BL}]$ is the amount of power needed to successfully reserve access and is obtained from (4.34), substituting $P_{m,U}$ with $P_{m,B}$, which denotes the fixed power employed by the MTDs to communicate with B.

Finally, the latency reduction gain $\mathcal{G}_D$ is defined as,

$$\mathcal{G}_D \triangleq \frac{E_{BL}[D]}{E_N[D]} \tag{4.43}$$

where $E_{BL}[D] = E[D_A]$, since the baseline scheme assumes the direct communication between MTDs and B, and $E_N[D]$ is obtained from (4.37).

4.6 Numerical Results

In this section, we examine the results obtained related to the key performance metrics derived in the previous section. Our analysis has been validated by performing computer simulations with enough repetitions to ensure numerical stability. Table 4.2 shows all the simulation parameters considered. To evaluate the benefits of our ap-

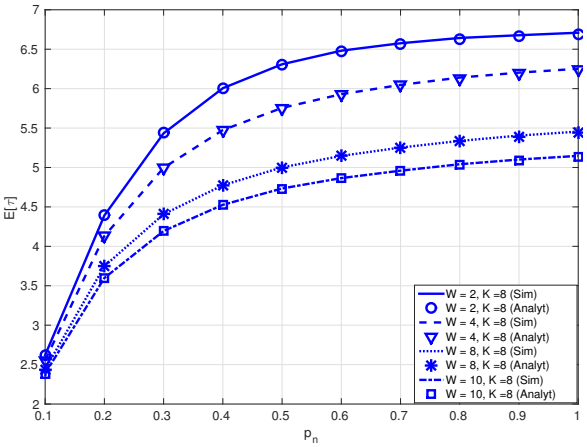
Parameter	Value	Parameter	Value
x_D	10 m	Γ_m	-3 dB
x_U	200 m	T	1 ms
$\bar{h}$	1	T_s	0.2 ms
$P_{m,U}$	-20 dBm	B_W	1 MHz
σ^2	-97 dBm	Q	2
α	3	D_m	64 bits
$P_{m,B}$	18 dBm	P_O	0.01

Table 4.2: Simulation Parameters

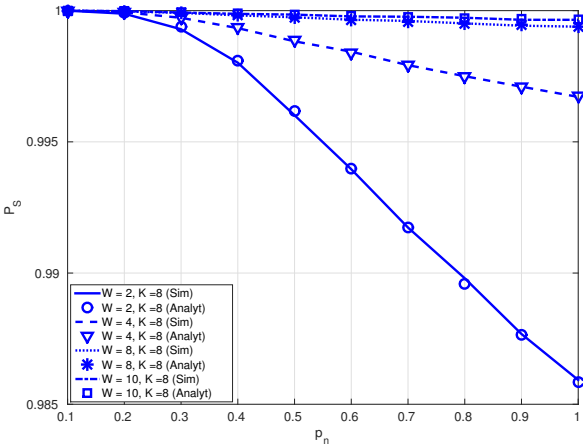
proach, we then compare the trunking-based protocol with the baseline scheme, where we adopt an FSA policy as described in Sec. 4.4.3.

4.6.1 Performance for different backoff parameters

We first evaluate the protocol performance for $R = 20$ and $M = 30$. Although the impact of the FWBO algorithm on the system performance can be investigated by changing both K and W, we chose to vary only the latter, since increasing K leads to higher delay and energy consumption, while higher values of W increase the delay, without demanding high energy. For this reason, we set the number of retransmissions K to 8, which represents a reasonable parameter for the considered traffic profile. Fig. 4-7a shows the average throughput in terms of the number of MTD packets aggregated per frame. We notice that a small window size W increases the throughput, as a larger number of packets can be potentially delivered to U during a frame, while higher values of W leads to a longer waiting time and to a smaller amount of delivered packets by each MTD. Fig. 4-7b shows the packet success probability as the overall load increases and for



(a)



(b)

Figure 4-7: Throughput (a) and success probability (b) for different K and W ($R = 20$, $M = 30$).

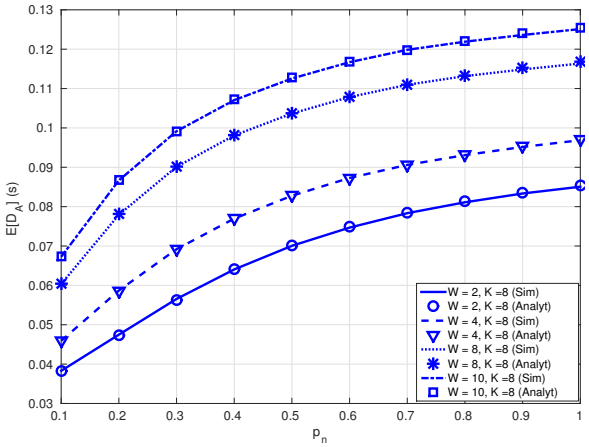


Figure 4-8: Average access delay for different K and W ($R = 20$, $M = 30$).

the same values of W. For $W = 2$, we can observe that P_S rapidly decreases in the medium-to-high access load region, i.e., $p_n > 0.4$. In other words, by increasing the load, the probability of selecting a single mini-slot decreases as the backlogged machines will attempt to retransmit more frequently due to the high level of congestion. This results in a higher number of packets dropped, whereas the effect of a high access load is almost negligible when $W \geq 8$. The average access delay defined as the average time needed to successfully transmit a packet is shown in Fig. 4-8. As expected, the larger W, the higher $E[D_A]$, since the increased window size forces all the retransmitting machines to remain backlogged for a longer time, until the next retransmission attempt.

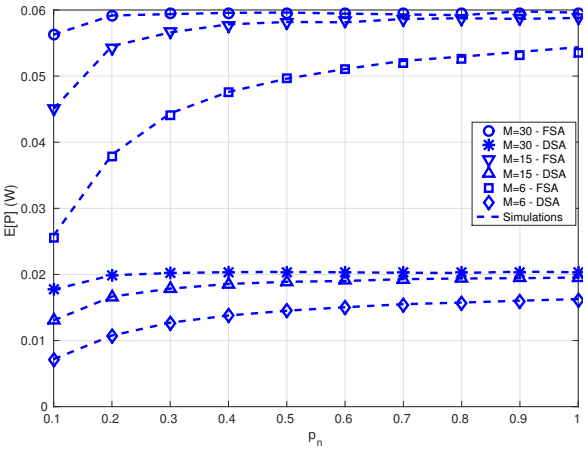
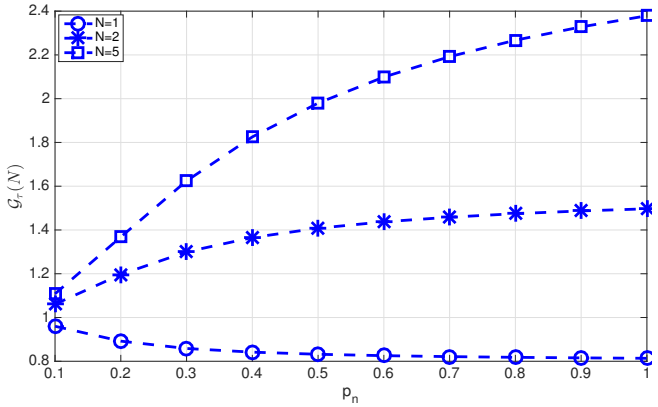


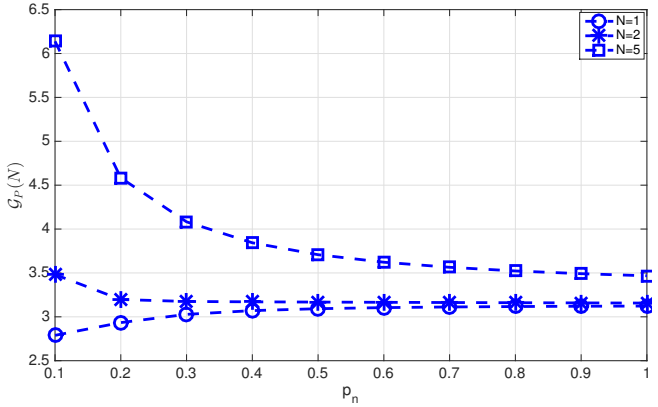
Figure 4-9: Average power spent to transmit an MTD packet for FSA and DSA ($R = 9, Q = 3, R_u = 0$).

4.6.2 FSA-DSA policy comparison

Fig. 4-9 shows the average power spent to transmit an MTD packet in the case of FSA and DSA policies. To evaluate the amount of power required to convey the machine packets, we set $R_u = 0$ such that CU packets coming from upper layers do not affect the performance computation. As expected, the DSA policy leads to a lower power consumption, since the number of trunking slots used can be dynamically increased depending on the number of successfully transmitting machines. It is also worth noticing that larger M requires less power per MTD packet, as more MTD packets can be simultaneously aggregated during a given frame.



(a)



(b)

Figure 4-10: $\mathcal{G}_\tau(N)$ (a) and $\mathcal{G}_P(N)$ (b) for different values of N ($R = 9, Q = 3, R_u = 0$).

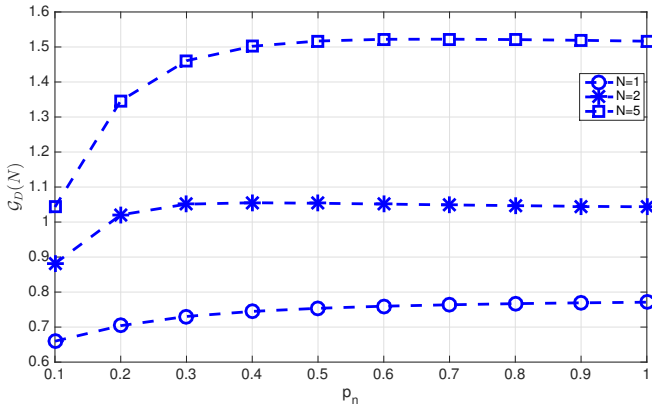


Figure 4-11: $\mathcal{G}_D(N)$ for different values of N ($R = 9, Q = 3, R_u = 0$).

4.6.3 Trunking gain evaluation

To evaluate the benefits of the trunking scheme, we consider a multi-cluster scenario, where the machines are equally distributed among non-overlapping clusters. We assume an FSA policy with $Q = 3$ and $N = 1, 2, 5$. Accordingly, the frame length L is set to 15, while M and R are chosen equal to 30 and 9, respectively.

The throughput gain is shown in Fig. 4-10a. We first note that high values of $\mathcal{G}_\tau(N)$ can be achieved by increasing the number of CUs N. As expected, for $N > 1$, the trunking approach can help to support a larger number of MTDs in comparison with the baseline scheme, while the single-cluster configuration ($N = 1$) leads to a gradual gain reduction due to the lower number of slots available in the aggregation subframe. On the other hand, the gain remains below 2 if the set of machines is simply split in half ($N = 2$). We also point out that for a low-to-medium access load, the gain increases more quickly due to the smaller number of potentially active MTDs per cluster, whereas this increase becomes slower when the system

approaches the saturated conditions, i.e., $p_n \approx 1$.

Fig. 4-10b shows the gain in terms of energy efficiency for the transmission of a single MTD packet. The highest values of $\mathcal{G}_P(N)$ can be reached by using all the available trunk links ($N = 5$), especially in low access load conditions, i.e., $p_n \leq 0.4$, where the power consumption is limited by the smaller cluster sizes. Interestingly, for $N = 1$, the gain increases and becomes approximately constant for $p_n > 0.3$, where the higher number of active MTDs leads to a remarkable power reduction, as the packet aggregation turns out to be more efficient.

Finally, we examine the gain in latency reduction $\mathcal{G}_D(N)$, which is illustrated in Fig. 4-11. By considering a single trunk link ($N = 1$), the baseline scheme guarantees better performance due to the higher number of slots available for data transmission and to the absence of the trunking link, which yields to additional delay due to the outage. We note that $\mathcal{G}_D(N)$ is lower than 1 and a small increment can be obtained only in the high access load region, while higher delay is achieved for $\lambda < 0.5$. On the other hand, we obtain $\mathcal{G}_D(N) > 1$ for $N = 2$ and $p_n > 0.1$, as a smaller number of MTDs are served by each CU, leading to a lower probability of packet retransmission. Furthermore, for $N = 5$, the increase in the gain is more evident and the highest values are achieved in the medium-to-high load region, i.e., $p_n > 0.5$. It is also important to notice that the gain slightly decreases after reaching the maximum value because of the increased number of active MTDs. This also means that each trunking configuration is characterized by an optimal activation probability p_n, maximizing $\mathcal{G}_D(N)$.

4.7 Conclusions

In this chapter, we proposed a novel MAC protocol to accommodate the traffic generated by MTDs, mitigating the congestion caused by the simultaneous massive access on a cellular BS. Our solution exploits D2D communications establishing between multiple MTDs and

a CU, which can simultaneously convey the MTT and its own data to the attached BS. We evaluated our protocol by employing a Markov chain, first deriving the key performance metrics related to the links machine-user and then incorporating the effect of the trunking cellular uplink. Furthermore, we compared the trunking scheme with a traditional direct access, by varying the number of aggregator users, with the aim of assessing the trunking gain in terms of throughput, energy efficiency and latency reduction.

Chapter 5

A Mobility Driven Joint Clustering and Relay Selection for IEEE 802.11p/WAVE Vehicular Networks

5.1 Introduction

Next-generation vehicles are expected to be digitally connected and capable of continuously collecting information on the surrounding environment, such as road conditions or obstacle presence, which are of paramount importance for making the driving experience more secure and enjoyable. Over the past few years, several academic and industrial projects have focused on designing fully autonomous cars, able to facilitate the parking or to prevent accidents as well as to

determine the optimal route towards the destination [59], paving the way to the new paradigm of *self-driving* or *driverless* vehicle. In this context, Vehicular Ad Hoc NETworks (VANET) can foster the required vehicle interconnection, extending the widely adopted and inexpensive Wireless Local Area Network (WLAN) technology towards vehicular scenarios [60]. Nevertheless, vehicular networks pose several challenges regarding the communication architecture specific design as well as security and privacy aspects which are essentially the reason of the reduced VANET deployment [61]. One of the many goals of VANET is to support traffic safety and make the driving experience more safe and comfortable, through the set up of intelligent and cooperative environments [62]. According to this vision, VANET entities are usually classified in Road Side Units (RSUs), typically located on traffic lights, and On-Board Units (OBUs), installed on mobile vehicles to enhance networking capabilities.

5.1.1 Related Work and Proposed approach

To support the vehicle-to-vehicle communication (V2V), IEEE has recently approved a suite of 802.11p/1609 Wireless Access in Vehicular Environments (WAVE) protocols [63,64]. WAVE system architecture includes the whole set of specifications to allow communication in the vehicular environments. In particular, MAC and PHY are based on IEEE 802.11a with fixed operating frequencies in the DSRC (Dedicated Short Range Communication) band of 5.85-5.925 GHz. Moreover, the resource management scheme is channel alternating; there is a MAC logical synchronization service which periodically switches the interested nodes among channels, i.e., the Control CHannel (CCH) and the multiple Service CHannel (SCH) [63]. The coordination is usually based on coordinated universal time (UTC), with synchronization interval length of 100 ms. It is also worth pointing out that the standard groups the users into WAVE Basic Service Sets (WBSSs), even though both RSU and OBU can be in charge of resource management.

One of the crucial aspects of IEEE 802.11p is the reduced con-

trol protocol scalability with respect to infrastructed wireless networks, especially for high density scenarios (e.g., a highway during rush-hour, or typical urban mobility patterns). As a consequence, the end-to-end communications could be severely affected and commonly adopted routing strategies [65] could be ineffective. Other approaches, as Hierarchical Cluster Based Routing (HCB) [66] adopts a two-layer communication architecture, supporting short and long range transmissions, which are not feasible in WAVE standards. To cope with this impairment, a clustering approach could be advantageously applied, disseminating information among groups in a broadcast/multicast fashion. This aspect has been previously investigated in [67] with reference to sparse highway scenarios. In this work, this approach is extended towards high density traffic conditions, where a certain degree of correlation among vehicles emerges such that a group mobility pattern (convoy) is expected. Our clustering protocol manages both cluster set up and maintaining and communications among different clusters. As for the former aspect, we refined the reference metric to select a Cluster Head (CH), by taking into account the node connectivity in terms of neighbors reachable within an estimated coherence time interval. Communications between clusters traveling in opposite directions are then handled by opportunistically exploiting the inter-contact time [68] and by selecting a suitable relay node.

The rest of this chapter is organized as follows: in Sec. 5.2 we characterize the typical vehicular mobility patterns, while giving evidence of driving profiles correlation which motivates the introduction of the proposed protocol detailed in Sec. 5.3, supporting both intra and inter cluster communications. Simulation results are presented in Sec. 5.4, while conclusions are drawn in Sec. 5.5.

5.2 System Model

In order to motivate and analyze the performance of the proposed clustering algorithm, we focused on generic vehicular mobility model.

It deals with a typical European highway, with two carriageways, each composed of two traffic lanes. In particular, the right lane is occupied by slower nodes, while the left lane is the passing lane, i.e., for faster vehicles. The model relies on the car following model [69], especially used in civil engineering to simulate the traffic behavior for a single lane under several traffic conditions. In more detail, this model describes the movements of a vehicle traveling along a straight road, keeping a security distance from a leading node; every time a vehicle exceeds the security distance, it must reduce its current speed, according to a certain deceleration factor.

During the set-up phase, a certain number of nodes is deployed over a roadway section (e.g., 1 km) depending on the vehicle density. Moreover, each node is assigned an initial speed randomly chosen according to a uniform distribution in a range between the minimum and maximum allowed speed, depending on the considered lane. In order to maintain the total number of nodes deployed constant over time, we modeled the playground as a toroid; once a vehicle reaches the border, it reenters the playground through the opposite border. The model represents the time evolution of the distance between adjacent vehicles according to the following rule [69]:

$$D = l + \beta v + \gamma v^2, \tag{5.1}$$

where D is the minimum desired distance, l is the actual vehicle length [m], v is the vehicle speed [m/s], β is the driver reaction time [s], and γ is the reciprocal of twice the maximum average deceleration of a leading vehicle. As for the γ parameter, it highlights the braking performance between the leading and following vehicle and can be neglected, assuming almost equal braking for all the vehicles. Finally, it is possible to choose the percentage of vehicles traveling on the passing lane with respect to the total number of nodes deployed.

The above mentioned model has been implemented by resorting to finite difference approach adopting the following parameters to characterize each investigated scenario:

- Vehicle density on a single carriage δ_N.
- Motorway length L [m].
- Percentage of nodes located in the passing lane N_{left} .
- Simulation time T_{sim} (in number of TSs of 100 ms).
- The maximum and minimum speed v authorized according to the specific lane [km/h].

Fig. 5-1 shows the scenario considered for the simulation set-up. Specifically, we analyze the protocol performance in a L-meter long section of the two-carriage motorway with two lanes for each motion direction.

The proposed model allows to study the overall traffic behavior under typical conditions experimented on a two-carriage motorway. For instance, the overtaking phase started by too slow vehicles can often produce traffic congestions, where vehicles on both the lanes of a single carriage are forced to temporarily adapt the driving style, (i.e., speed, acceleration and breaking instants). As a result, the vehicle distribution over time can be described by using a *convoy mobility pattern*, where vehicles join or leave a leading convoy depending on current traffic conditions. The presence of groups motivates the clusterization: this is pointed out in Fig. 5-2, where the cumulative frequency of inter vehicles distance is shown, for both right and left lane. In particular, statistical results obtained show that approximately 95% and 85% are within a coverage radius of 100 m, depending on the corresponding lane.

5.3 Proposed Protocol

As mentioned earlier, IEEE 802.11p/WAVE standard architecture is meant to support wireless communications in high mobility vehicular environments. A typical WAVE system consists of fixed RSUs and mobile OBUs installed in vehicles. In this chapter we focus only on vehicle-to-vehicle (V2V) communications, where OBUs move along a

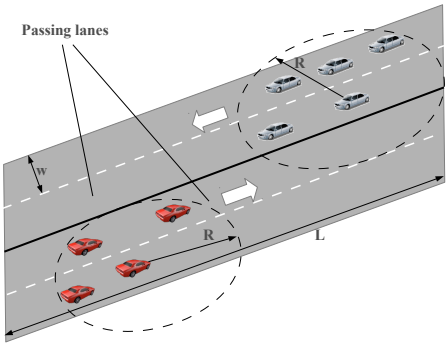


Figure 5-1: Illustration of the considered scenario.

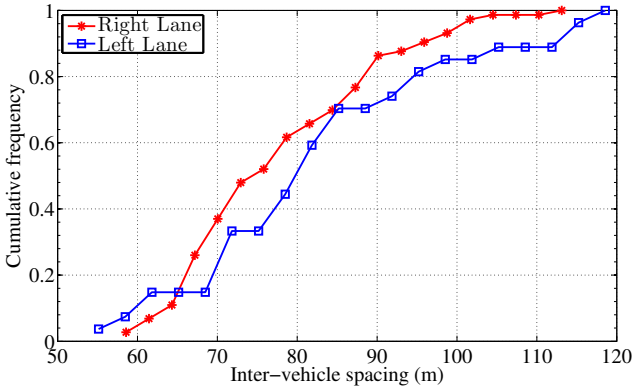


Figure 5-2: Cumulative frequency of inter vehicles distance for both right and left lanes.

fixed direction. We also assume that information about vehicle speed, position and driving direction is periodically updated and exchanged by means of Cooperative Awareness Messages (CAM) broadcast by each vehicle to all the neighbors with a frequency between 2 and 10 Hz, as foreseen in the WAVE system [70].

As explained in Sec. 5.2, the typical mobility pattern is comprised of convoys, thus motivating the clustering of vehicles to allow the gathering, fusion and dissemination of sensed data, and enhancing the context awareness. Differently from previous work [67], we rely on the vehicle information to select the cluster head (CH) and to join different cluster as seamlessly as possible. The protocol phases are further introduced and characterized.

5.3.1 Cluster Formation and Maintenance

This phase basically aims at splitting the network into 1-hop clusters of different sizes. Instead of considering the node degree, i.e., the number of 1-hop neighbors, the proposed protocol relies on the mobility characteristics and correlation properties among the nodes within the same radio coverage. By exploiting information provided by CAM packets, an estimated connectivity time is computed and used to select a proper CH node. For the sake of simplicity, we assume no interference among vehicles traveling along different carriages, i.e., only nodes being on the same carriage may take part in the clustering process. Furthermore, we consider the clustering election process completely periodic and triggered every T_u TSs.

Fig. 5-3 shows the reference system model considering only a single carriage with two lanes. Furthermore, the entire algorithm for the metric computation is shown in Algorithm 1.

To estimate the generic distance $\hat{d}_{ij}$ of the node j from the radio coverage border of node i, the latter needs to know current position and speed of the former. We assume the coverage radius $R = \hat{d}_{ij} + d_{ij}$, as the distance h is typically much smaller than the coverage radius

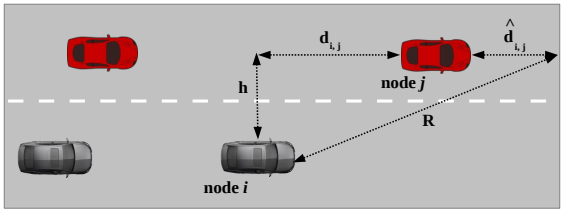


Figure 5-3: Illustration of the reference system model.

Algorithm 1 Metric computation for node i

1: **for** $j = 1$ to ν **do**
2: **if** $v_i < v_j$ **then**
3: Get $\delta_{v,ij}$
4: Compute $\hat{T}_{c,ij}$
5: **end if**
6: **end for**
7: Compute $\hat{T}_{c,i}$

R. Thus, $\hat{d}_{ij}$ can be obtained as follows:

$$\hat{d}_{ij} \doteq R - d_{ij}. \tag{5.2}$$

The speed difference between the i-th and j-th node is:

$$\delta_{v,ij} \doteq v_j - v_i, \tag{5.3}$$

where v_i and v_j are the speed of the i-th and j-th node, respectively. It is worth pointing out that if $\delta_{v,ij} < 0$, the connectivity time cannot be estimated, as node i would be faster than node j and could not be considered as reference by node i. This event implicitly introduces a threshold in the d_{ij} process, preventing a faster node from computing the metric and broadcasting it. Then, the estimated connectivity time between i-th and j-th node can be obtained as follows:

$$\hat{T}_{c,ij} = \frac{\hat{d}_{ij}}{\delta_{v,ij}}. \tag{5.4}$$

Finally, by averaging this metric with respect to the number of the neighbors involved in the process ν, we obtain the average estimated connectivity time $\hat{T}_{c,i}$:

$$\hat{T}_{c,i} = \frac{\sum_{j=1}^{\nu} \hat{T}_{c,ij}}{\nu}. \tag{5.5}$$

Algorithm 2 shows the overall CH election algorithm. First of all,

Algorithm 2 CH selection

1: Get $\hat{T}_c$ metric from all the neighbors
2: **if** Current node is not isolated **then**
3: Find neighbor with max $\hat{T}_c$
4: Send Broadcast 1-hop (selected CH ID)
5: **end if**

a generic node receives and collects the $T_{c,k}$ metric related to all its 1-hop neighbors, i.e., the possible CH candidates. Then, if this node is not isolated, namely it has at least a neighbor, the CH is selected by searching for the neighbor with the maximum T_c value. Finally, the CH identifier (ID) is broadcast to all the neighbors in order to inform the selected CH about the election result.

5.3.2 Inter-cluster Communication

In this Section we characterize a novel inter-cluster communication scheme which allows the low latency delivering of high priority messages among adjacent clusters located on different carriages. This proposal extends the intra-cluster communications framework simplifying the clusters management (i.e., CHs re-election) by selecting a couple of relays which are in charge of providing connectivity and, consequently, allowing CHs to further disseminate the recent information to all the members. Differently from our previous investigation [67], focused on sparse highway scenarios, we assume here a certain degree of correlation among vehicles such that a group mobility pattern is kept stable, as explained in Sec. 5.2. Thus, the merging of different clusters is avoided, while maximizing the information exchange among them without setting up any routing. In more detail, unlike urban traffic jam scenarios, in a typical high mobility traffic scenario, an end-to-end path between the source and destination cannot be reliably established. Moreover, the high speed difference experienced by nodes traveling in opposite directions can affect the inter-cluster connectivity and lead to overall performance degradation, especially in networks with high density of clusters. To face this problem, we designed a relaying technique able to establish robust connections between clusters. Instead of considering a flooding-based approach, a *dissemination* based-mechanism is assumed, being the latter a more scalable solution thanks to the limited number of broadcast transmissions [71]. Furthermore, our solution exploits the bidirectional mobility of vehicles exchanging information in a *carry-and-forward* fashion, according to the well known paradigm of the De-

lay Tolerant Networks (DTN) [72]. For this reason, in our approach we refer to a Custody Transfer Protocol (CTP) [73] as a mechanism in charge of dealing with the retransmission of previously buffered messages. Basically, a node holding the custodian of a message cannot delete it until a reliable transmission of the latter is completed, with the aim of increasing the reliability of the packet transmission.

The principle behind our proposal consists in cluster relay (CR) selection performed by CHs once a potential cluster reconfiguration occurs; in particular, the cluster member characterized by the lowest relative speed is chosen. Accordingly, this solution maximizes the connectivity among approaching clusters, as the slowest nodes guarantee a reduced disconnection probability. The overall procedure can be summarized in three different steps:

1. CH node selects a CR according to the above mentioned criterion;
2. CR node receives and holds cluster related information periodically sent by the CH;
3. Every time an inter-cluster communication opportunity exists, CRs belonging to different clusters establish a connection and exchange data by relaying on the CTP protocol.

5.4 Simulation Results

5.4.1 Scenario Characterization

Numerical simulations have been performed by using the mobility model described in Sec. 5.2. In particular, scenario A focuses on a single two-lane carriage and describes traffic congestion situations randomly generated by the presence of slow vehicles on both the lanes. Simulator parameters are shown in Table 5.1. We chose a simulator playground 10 km long, consisting of two parallel and straight lanes with width of 8 m each. Moreover, a total number of 250 vehicles are randomly deployed, as the node density is set to 25 nodes/km,

Parameters	Scenario A	Scenario B
T_s	100 ms	100 ms
L	10 km	10 km
w	8 m	16 m
Number of lanes	2	4
δ_N	25 nodes/km	15 nodes/km
$N_{left}(\%)$	30%	30%
T_{sim}	300 s	300 s
Min-max v (left line)	27.77 - 36.11 m/s	27.77 - 36.11 m/s
Min-max v (right line)	22.22 - 30.55 m/s	22.22 30.55 m/s
R	$25 - 250$ m	$25 - 250$ m
f_b	10 Hz	10 Hz
T_u	5 T_s	5 T_s

Table 5.1: Simulation parameters for scenario A and B.

with a percentage of 30% of vehicles located on the passing lane. As for vehicle speeds, the values are set in a random fashion according to minimum and maximum authorized speeds on each lane. Furthermore, we analyzed the system evolution over a time interval of 5 minutes, using one hundred repetitions of the simulations in order to achieve reliable results. On the other hand, scenario B describes a standard dual carriage motorway with two lanes for each motion direction. As shown in Table I, we also set the overall length and width to 10 km and to 8 m, respectively. In this case, the node density is equal to 15 nodes/km that corresponds to a total number of 150 nodes deployed on each single carriage[1] . Finally, remaining parameters are selected similarly to the previous scenario, with the percentage of nodes in the passing lane equal to 30%.

We considered a free-space radio channel with a fixed coverage

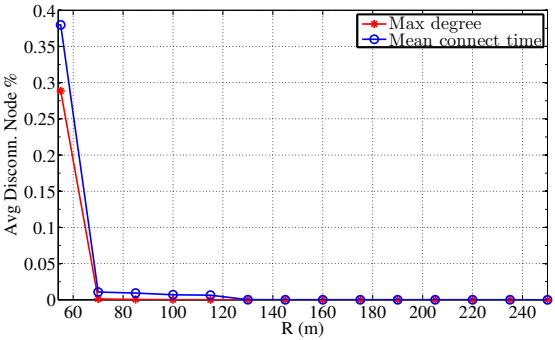
[1]The total number of nodes simulated is equal to 300.

radius R for all the nodes and we also performed simulations as R changes. Moreover, the beaconing rate (i.e., the frequency of CAM messages) f_b is fixed to 10 Hz, i.e., each vehicle sends a beacon once per WAVE Synchronization Interval T_s . Finally, we assumed that the clustering update interval T_u varies from 2 T_s to 10 T_s in order to analyze the algorithm performance as its reactivity changes.

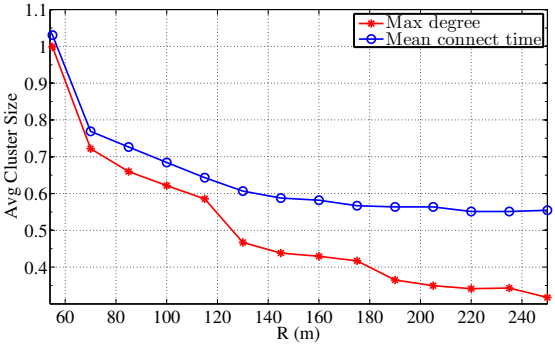
5.4.2 Numerical Results

In order to evaluate the performance of the proposed clustering scheme, we compared it with an algorithm based on the maximum one hop degree D, where nodes with the maximum number of neighbors are elected as CHs. First of all we focus on Scenario A. In Fig. 5-4a, the two approaches are compared in terms of percentage of disconnected nodes as a function of R, pointing out a substantial equivalence especially for medium-to-high R, where 99% of nodes are fully connected. Anyway, this result is achieved in a different way by the schemes. In particular, some differences are highlighted in Figs. 5-4b and 5-5a, since the mean connectivity time approach presents higher cluster size and lower percentage of CHs with respect to the maximum one hop degree scheme. Note that in the case of mean T_c, the average cluster size turns out to be constant for values grater than 160 m, while in the other case it decreases if R increases. Considering the cluster life-time as a function of R, one can observe how the proposed scheme outperforms the D-based approach. Fig. 5-5b shows the simulation results obtained. Better performance is achieved for $R < 160$ m, while the difference is almost negligible for higher values of R.

We also investigated the impact of the update interval T_u and coverage radius R on the inter-cluster communication. More in detail, Figs. 5-6a and 5-6b show the average number of occurrences of CH as relay as function of T_u and R, respectively. It is worth noticing that the former parameter does not affect the relay selection, as shown in Fig. 5-6b. Furthermore, the probability of CH being the best relay is greater than 50% for $R < 200$ m, meaning that considering CHs in the relay election procedure can be beneficial to enhance the overall

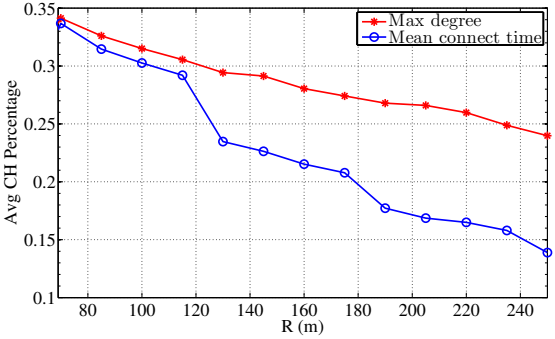


(a)

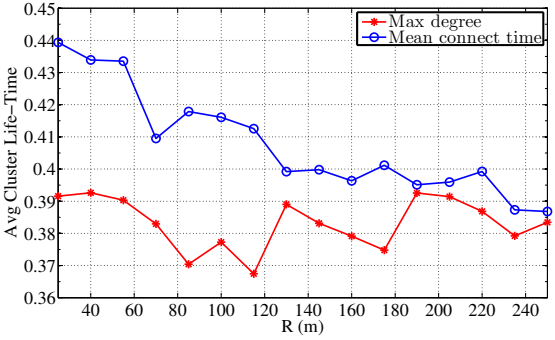


(b)

Figure 5-4: Average percentage of disconnected nodes (a) and average cluster size over average number of neighbors (b) as a function of R (scenario A).

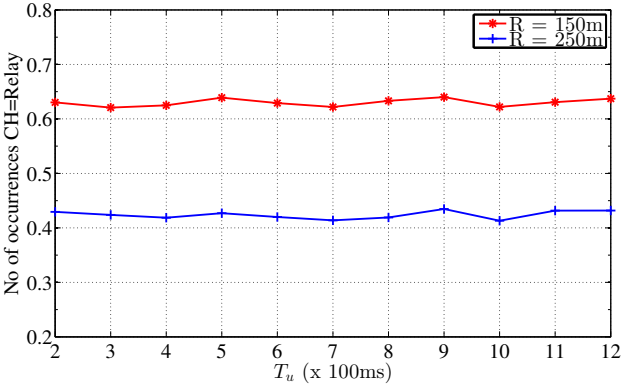


(a)

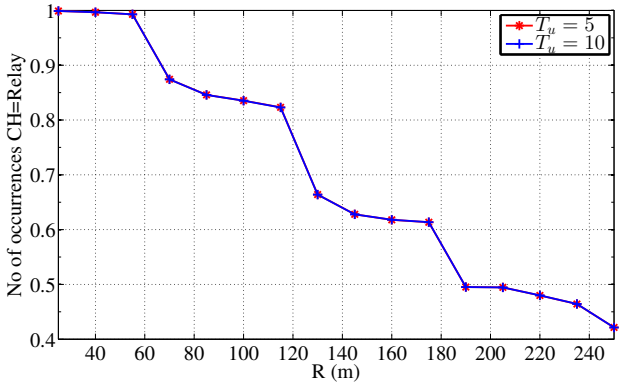


(b)

Figure 5-5: Percentage of CHs (a) and average cluster life-time normalized with respect to the simulation duration (b) as a function of R (scenario A).



(a)

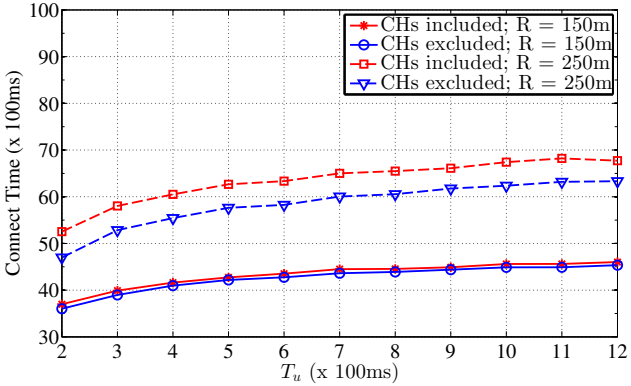


(b)

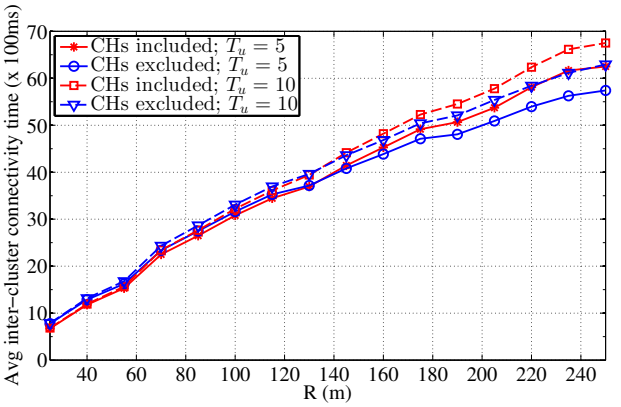
Figure 5-6: Average number of occurrences of CH as relay as function of T_u (a) and R (b) (scenario B).

performance.

Another key element of the inter-cluster communications is the



(a)



(b)

Figure 5-7: Average relay-to-relay connectivity time in *ms* as a function of T_u (a) and R (b) (scenario B).

average relay-to-relay connectivity time, i.e., the average time interval during that two relay nodes traveling with opposite directions can

exchange information. Figs. 5-7a and 5-7b show results obtained by varying T_u and R, respectively, including and excluding CH nodes from the relay selection. In the first case, the relay-to-relay connectivity time is slightly influenced by T_u, while in the other case increasing the coverage radius allows to achieve better performance.

5.5 Conclusions

The wide adoption of VANET paradigm through 802.11p/WAVE standard is presently limited, among other issues, by the networking capabilities which dramatically affect data handling and, thus, the safety of the driving experience. Focusing on V2V communications, we proposed a clustering approach for gathering and disseminating information among groups in a broadcast/multicast fashion. We extended our previous approach towards high density traffic conditions, where a certain degree of correlation among vehicles emerges such that a group mobility pattern can be modelled. In particular, we addressed the cluster set-up and maintaining phase, as well as the communication establishment among different clusters. As for the former aspect, we derived a metric which takes into account the node connectivity within an estimated coherence time interval, leading the CH selection. On the other hand, the latter aspect is addressed by opportunistically exploiting the inter-contact time among clusters traveling in opposite directions with the aim of selecting a relay node. After characterizing a typical vehicular mobility pattern, which motivates the proposal protocol, the proposed approach has been carefully described along all its phases for both intra and inter cluster communications. The protocol validation has been focused on a practical scenario adopting IEEE 802.11p/WAVE communication standard, highlighting good performance in terms of connection probability and overall complexity.

Chapter 6

Conclusions and Future Research Directions

6.1 General Conclusions

In this thesis, we proposed different solutions to enable M2M communications in diverse application scenarios, such as remote monitoring and intelligent transportation systems. Along with existing approaches extensively adopted in literature, including clustering, packet aggregation and trunking, we leveraged the D2D communication mode introduced in recent releases of LTE-A, which allows devices to directly exchange data under the network supervision. This in turn guarantees numerous benefits in terms of energy efficiency, throughput and latency, and provides a high degree of security and interference mitigation due to the licensed spectrum.

In chapter 2, we introduced a novel multi-hop networking scheme for devices employing D2D connectivity, overlaying an LTE-A system where H2H and M2M can coexist without increasing interference or affecting the overall network performance. We designed a centralized protocol to assist devices in establishing a D2D link, providing

support for proximity discovery and device pairing. Furthermore, an efficient resource allocation mechanism assigns radio resources to each D2D pair based on the connectivity graph representing D2D links available within the same cell, preventing potential bottlenecks on the links more involved in the forwarding of the M2M traffic.

Chapter 3 described an alternative approach to mitigate the RAN overload originated from the massive number of devices requesting access within the same cell. We developed a multiple access scheme combining D2D and cellular connections, where the M2M traffic is delivered to the BS through a CU located in the proximity of a certain number of MTDs employing low-power direct connectivity. In other words, the CU acts as a relay and is responsible of aggregating machine packets with its own traffic and send the aggregated data through the cellular uplink, thus following the paradigm of *trunked links*. We then compare the system performance with a traditional direct access scheme, illustrating the benefits of the trunking approach in terms of throughput and energy efficiency.

In chapter 4, we further enhanced our trunking-based protocol by incorporating a backoff algorithm aiming at reducing the probability of collisions among MTDs simultaneously requesting access to the CU. To support the analysis of the performance metrics, such as throughput, probability of successful packet delivery and delay, we also elaborated an accurate Markov chain modeling the temporal evolution of an access request made by an idle device. In addition, we evaluated the performance in a multi-cluster configuration, where MTDs are assigned to different clusters served by multiple CUs, which can forward the traffic on the correspondent trunked uplink connection without generating inter-cluster interference. This solution allows to define the trunking gains, i.e., gains in performance metrics obtained by employing the multi-cluster configuration with respect to a traditional direct access scheme, pointing out the advantages provided as the number of available relay users increases.

In chapter 5, M2M communications were investigated in the context of VANETs, where connected vehicles exchange crucial infor-

mation to support numerous functions, such as preventing accidents or avoiding traffic congestion. Assuming all vehicles equipped with IEEE 802.11p/WAVE radio interfaces, we devised a clustering protocol based on the correlation degree of mobile nodes, such that vehicles moving with similar speeds and directions are grouped and connected to a single cluster-head vehicle, which is also responsible of establishing inter-cluster connections. We finally illustrated the benefits of this approach in terms of the connection probability by developing a simulator modeling a realistic vehicular environment, taking into account potential mobility variations between approaching vehicles.

6.2 Directions for Future Research

The open challenges to be addressed in the future are numerous. First of all, the performance of the overlay D2D networking scheme needs to be investigated by adopting alternative routing algorithms for ad-hoc networking, such as OLSR or AODV, evaluating the time the network needs to converge and the amount of signalization overhead produced. Furthermore, power consumption and node mobility must be considered in the resource allocation algorithm, since devices more involved in the packet forwarding will consume power more quickly, thus leading to outage and consequent network reconfiguration. The proximity discovery has also to be enhanced by developing a specific beaconing procedure between two nearby devices, as the overhearing mechanism could be difficult to apply.

In the trunking protocol, an interesting problem is represented by the lack of complete channel state information at the CU. Analyzing the protocol in scenarios with channel uncertainties may help to understand the limits of the trunked link and find alternative power control policies to support the transmission on the time-varying channel. Moreover, the effect of different traffic patterns and limited buffer size on the CU must be investigated. The amount of data collected may lead to buffer overflow situations, which can affect the transmission of user data as well as increase the packet delay experienced by

MTDs.

For vehicular communications, our model can be improved considering interference generated by vehicles traveling with opposite directions or simulating high-dense traffic conditions as in the case of urban scenarios, which can perplex the clustering formation and reduce the network availability. Furthermore, comparisons with existing clustering approaches need to be made in order to evaluate the benefits provided by the proposed protocol. Another open issue is how to reduce the amount of isolated vehicles. This could be addressed by installing another radio interface, such as LTE, capable of cooperating with IEEE 802.11p/WAVE and of keeping all the vehicles synchronized.

Bibliography

[1] G. Wu, S. Talwar, K. Johnsson, N. Himayat, and K. Johnson, "M2M: From mobile to embedded internet," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 36–43, 2011.

[2] A. Zanella, M. Zorzi, A. dos Santos, P. Popovski, N. Pratas, C. Stefanovic, A. Dekorsy, C. Bockelmann, B. Busropan, and T. Norp, "M2M massive wireless access: Challenges, research issues, and ways forward," in *Globecom Workshops (GC Wkshps), 2013 IEEE*, Dec 2013, pp. 151–156.

[3] C. Kahn and H. Viswanathan, "Connectionless access for mobile cellular networks," *Communications Magazine, IEEE*, vol. 53, no. 9, pp. 26–31, September 2015.

[4] J. Wu, T. Zhang, Z. Zeng, and H. Chen, "Study on discontinuous reception modeling for M2M traffic in LTE-A networks," in *Communication Technology (ICCT), 2013 15th IEEE International Conference on*, Nov 2013, pp. 584–588.

[5] S. Razavi and D. Yuan, "Reducing signaling overhead by overlapping tracking area list in LTE," in *Wireless and Mobile Networking Conference (WMNC), 2014 7th IFIP*, May 2014, pp. 1–7.

[6] A. Ali, W. Hamouda, and M. Uysal, "Next generation M2M cellular networks: challenges and practical considerations," *Com-*

munications Magazine, IEEE, vol. 53, no. 9, pp. 18–24, September 2015.

[7] F. Ghavimi and H.-H. Chen, "M2M Communications in 3GPP LTE/LTE-A Networks: Architectures, Service Requirements, Challenges, and Applications," *Communications Surveys Tutorials, IEEE*, vol. 17, no. 2, pp. 525–549, Secondquarter 2015.

[8] T. Adame, A. Bel, B. Bellalta, J. Barcelo, and M. Oliver, "IEEE 802.11AH: the WiFi approach for M2M communications," *Wireless Communications, IEEE*, vol. 21, no. 6, pp. 144–152, December 2014.

[9] X. Xiong, K. Zheng, R. Xu, W. Xiang, and P. Chatzimisios, "Low power wide area machine-to-machine networks: key techniques and prototype," *Communications Magazine, IEEE*, vol. 53, no. 9, pp. 64–71, September 2015.

[10] A. Lo, Y. Law, and M. Jacobsson, "A cellular-centric service architecture for machine-to-machine (M2M) communications," *Wireless Communications, IEEE*, vol. 20, no. 5, pp. 143–151, October 2013.

[11] 3GPP, "TR36.888 - Study on provision of low-cost Machine-Type Communications (MTC) User Equipments (UEs) based on LTE," 3GPP, Tech. Rep., 2012.

[12] S. Andreev, O. Galinina, and Y. Koucheryavy, "Energy-Efficient Client Relay Scheme for Machine-to-Machine Communication," in *GLOBECOM, 2011 IEEE*, Dec 2011, pp. 1–5.

[13] H. kwan Lee, D. M. Kim, Y. Hwang, S. M. Yu, and S.-L. Kim, "Feasibility of cognitive machine-to-machine communication using cellular bands," *Wireless Communications, IEEE*, vol. 20, no. 2, pp. 97–103, April 2013.

[14] H. Shariatmadari, R. Ratasuk, S. Iraji, A. Laya, T. Taleb, R. Ja?ntti, and A. Ghosh, "Machine-type communications: current status and future perspectives toward 5G systems," *Com-*

munications Magazine, IEEE, vol. 53, no. 9, pp. 10–17, September 2015.

[15] V. B. Mišić, J. Mišić, X. Lin, and D. Nerandzic, "Capillary machine-to-machine communications: the road ahead," in *Ad-hoc, Mobile, and Wireless Networks*. Springer, 2012, pp. 413–423.

[16] D. Astely, E. Dahlman, G. Fodor, S. Parkvall, and J. Sachs, "LTE release 12 and beyond," *Communications Magazine, IEEE*, vol. 51, no. 7, pp. 154–160, 2013.

[17] D. Feng, L. Lu, Y. Yuan-Wu, G. Li, G. Feng, and S. Li, "Device-to-Device Communications Underlaying Cellular Networks," *Communications, IEEE Transactions on*, vol. 61, no. 8, pp. 3541–3551, 2013.

[18] C.-H. Yu, K. Doppler, C. Ribeiro, and O. Tirkkonen, "Resource Sharing Optimization for Device-to-Device Communication Underlaying Cellular Networks," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 8, pp. 2752–2763, 2011.

[19] Y.-D. Lin and Y.-C. Hsu, "Multihop cellular: a new architecture for wireless communications," in *INFOCOM 2000. Nineteenth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE*, vol. 3, Mar 2000, pp. 1273–1282 vol.3.

[20] A. Laya, L. Alonso, and J. Alonso-Zarate, "Is the Random Access Channel of LTE and LTE-A Suitable for M2M Communications? A Survey of Alternatives," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–13, 2013.

[21] S.-Y. Lien, K.-C. Chen, and Y. Lin, "Toward ubiquitous massive accesses in 3GPP machine-to-machine communications," *Communications Magazine, IEEE*, vol. 49, no. 4, pp. 66–74, 2011.

[22] K. Zheng, F. Hu, W. Wang, W. Xiang, and M. Dohler, "Radio resource allocation in LTE-advanced cellular networks

with M2M communications," *Communications Magazine, IEEE*, vol. 50, no. 7, pp. 184–192, 2012.

[23] W. Qiu, P. Tarasak, and H. Minn, "Orthogonal Multicarrier Division Multiple Access for Multipoint-to-Multipoint Networks," *Communications, IEEE Transactions on*, vol. 61, no. 9, pp. 3841–3853, 2013.

[24] G. Madueo, C. Stefanovic, and P. Popovski, "Reliable Reporting for Massive M2M Communications With Periodic Resource Pooling," *Wireless Communications Letters, IEEE*, vol. 3, no. 4, pp. 429–432, Aug 2014.

[25] S. Xiang, T. Peng, Z. Liu, and W. Wang, "A distance-dependent mode selection algorithm in heterogeneous D2D and IMT-Advanced network," in *Globecom Workshops (GC Wkshps), 2012 IEEE*, 2012, pp. 416–420.

[26] K. Doppler, C.-H. Yu, C. Ribeiro, and P. Janis, "Mode Selection for Device-To-Device Communication Underlaying an LTE-Advanced Network," in *Wireless Communications and Networking Conference (WCNC), 2010 IEEE*, 2010, pp. 1–6.

[27] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, "Design aspects of network assisted device-to-device communications," *Communications Magazine, IEEE*, vol. 50, no. 3, pp. 170–177, 2012.

[28] T. Winter, "RPL: IPv6 Routing Protocol for Low-Power and Lossy networks," 2012.

[29] V. Mhatre and C. Rosenberg, "Design guidelines for wireless sensor networks: communication, clustering and aggregation," *Ad Hoc Networks*, vol. 2, no. 1, pp. 45 – 63, 2004.

[30] T. Kwon and J. Cioffi, "Random Deployment of Data Collectors for Serving Randomly-Located Sensors," *Wireless Communications, IEEE Transactions on*, vol. 12, no. 6, pp. 2556–2565, June 2013.

[31] K.-C. Chen and S.-Y. Lien, "Machine-to-machine communications: Technologies and challenges," *Ad Hoc Networks*, vol. 18, no. 0, pp. 3 – 23, 2014.

[32] V. Pourahmadi, S. Fashandi, A. Saleh, and A. Khandani, "Relay Placement in Wireless Networks: A Study of the Underlying Tradeoffs," *Wireless Communications, IEEE Transactions on*, vol. 10, no. 5, pp. 1383–1388, May 2011.

[33] A. Chrapkowski and G. Grube, "Mobile trunked radio system design and simulation," in *Vehicular Technology Conference, 1991. Gateway to the Future Technology in Motion., 41st IEEE*, May 1991, pp. 245–250.

[34] A. Asadi, Q. Wang, and V. Mancuso, "A Survey on Device-to-Device Communication in Cellular Networks," *Communications Surveys Tutorials, IEEE*, vol. PP, no. 99, pp. 1–1, 2014.

[35] J. Wieselthier, A. Ephremides, and L. Michaels, "An exact analysis and performance evaluation of framed ALOHA with capture," *Communications, IEEE Transactions on*, vol. 37, no. 2, pp. 125–137, Feb 1989.

[36] H. Vogt, "Efficient object identification with passive RFID tags," in *Pervasive Computing*. Springer, 2002, pp. 98–113.

[37] M.-S. Alouini and A. Goldsmith, "Capacity of Rayleigh fading channels under different adaptive transmission and diversity-combining techniques," *Vehicular Technology, IEEE Transactions on*, vol. 48, no. 4, pp. 1165–1181, Jul 1999.

[38] G. Rigazzi, N. K. Pratas, P. Popovski, and R. Fantacci, "Aggregation and trunking of M2M traffic via D2D connections," in *IEEE ICC 2015 - Mobile and Wireless Networking Symposium (ICC'15 (03) MWN)*, London, United Kingdom, Jun. 2015.

[39] F. Vazquez-Gallego, J. Alonso-Zarate, A. M. Mandalari, O. Briante, A. Molinaro, and G. Ruggeri, "Performance Evaluation of Reservation Frame Slotted-ALOHA for Data Collection M2M

Networks," in *European Wireless 2014; 20th European Wireless Conference; Proceedings of*, May 2014, pp. 1–6.

[40] "MTC simulation results with specific solutions," *3GPP TSG RAN WG2, R2-104662*, August 2010.

[41] 3GPP, "TR37.869 - Study on enhancements to machine-type communications (MTC) and other mobile data applications; radio access network (RAN) aspects," 3GPP, Tech. Rep., 2013.

[42] "3GPP TR 23.887, Study on Machine-Type Communications (MTC) and other mobile data applications communications enhancements," September 2014.

[43] H. Wu, C. Zhu, R. La, X. Liu, and Y. Zhang, "FASA: Accelerated S-ALOHA Using Access History for Event-Driven M2M Communications," *Networking, IEEE/ACM Transactions on*, vol. 21, no. 6, pp. 1904–1917, Dec 2013.

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

[45] "3GPP TR 22.888, Study on enhancements for Machine-Type Communications (MTC) (Release 12)," March 2013.

[46] H.-L. Fu, P. Lin, H. Yue, G.-M. Huang, and C.-P. Lee, "Group Mobility Management for Large-Scale Machine-to-Machine Mobile Networking," *Vehicular Technology, IEEE Transactions on*, vol. 63, no. 3, pp. 1296–1305, March 2014.

[47] A. Aijaz and H. Aghvami, "PRMA based Cognitive Machine-to-Machine Communication in Smart Grid Networks," *Vehicular Technology, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2014.

[48] K.-H. Huang, S.-A. Lin, S.-Y. Ho, Y.-C. Lin, I.-W. Lu, J.-C. Huang, and C.-T. Chou, "Analysis of a prioritized medium

access control for 2-hop machine-to-machine (M2M) communication networks," in *Communications Workshops (ICC), 2014 IEEE International Conference on*, June 2014, pp. 435–440.

[49] D. Niyato, P. Wang, and D. I. Kim, "Performance Modeling and Analysis of Heterogeneous Machine Type Communications," *Wireless Communications, IEEE Transactions on*, vol. 13, no. 5, pp. 2836–2849, May 2014.

[50] S.-C. Lin, L. Gu, and K.-C. Chen, "Providing statistical QoS guarantees in large cognitive machine-to-machine networks," in *Globecom Workshops (GC Wkshps), 2012 IEEE*, Dec 2012, pp. 1700–1705.

[51] J. Liu, Y. Kawamoto, H. Nishiyama, N. Kato, and N. Kadowaki, "Device-to-device communications achieve efficient load balancing in LTE-advanced networks," *Wireless Communications, IEEE*, vol. 21, no. 2, pp. 57–65, April 2014.

[52] S. Andreev, O. Galinina, A. Pyattaev, K. Johnsson, and Y. Koucheryavy, "Analyzing Assisted Offloading of Cellular User Sessions onto D2D Links in Unlicensed Bands," *Selected Areas in Communications, IEEE Journal on*, vol. 33, no. 1, pp. 67–80, Jan 2015.

[53] G. Fodor, D. Della Penda, M. Belleschi, M. Johansson, and A. Abrardo, "A comparative study of power control approaches for device-to-device communications," in *Communications (ICC), 2013 IEEE International Conference on*, June 2013, pp. 6008–6013.

[54] T.-H. Chuang, M.-H. Tsai, and C.-Y. Chuang, "Group-Based Uplink Scheduling for Machine-Type Communications in LTE-Advanced Networks," in *Advanced Information Networking and Applications Workshops (WAINA), 2015 IEEE 29th International Conference on*, March 2015, pp. 652–657.

[55] D. Malak, H. S. Dhillon, and J. G. Andrews, "Optimizing Data Aggregation for Uplink Machine-to-Machine Communication Networks," *arXiv preprint arXiv:1505.00810*, 2015.

[56] N. Pratas and P. Popovski, "Underlay of low-rate machine-type D2D links on downlink cellular links," in *Communications Workshops (ICC), 2014 IEEE International Conference on*, June 2014, pp. 423–428.

[57] ——, "Zero-Outage Cellular Downlink With Fixed-Rate D2D Underlay," *Wireless Communications, IEEE Transactions on*, vol. 14, no. 7, pp. 3533–3543, July 2015.

[58] J. B. Seo and V. C. M. Leung, "Design and analysis of backoff algorithms for random access channels in UMTS-LTE and IEEE 802.16 systems," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 8, pp. 3975–3989, 2011.

[59] "Google Self-Driving Car Project," http://www.google.com/selfdrivingcar/.

[60] H. Hartenstein and K. P. Laberteaux, "A tutorial survey on vehicular ad hoc networks," *Communications Magazine, IEEE*, vol. 46, no. 6, pp. 164–171, 2008.

[61] Y. Toor, P. Muhlethaler, and A. Laouiti, "Vehicle Ad Hoc networks: applications and related technical issues," *Communications Surveys Tutorials, IEEE*, vol. 10, no. 3, pp. 74–88, Third 2008.

[62] F. Chiti and R. Fantacci, "Intelligent and cooperative environments via internet of sensors: Perspectives and case study analysis," in *Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International*. IEEE, 2013, pp. 19–23.

[63] R. Uzcategui and G. Acosta-Marum, "WAVE: a tutorial," *Communications Magazine, IEEE*, vol. 47, no. 5, pp. 126–133, 2009.

[64] Č. Stefanović, D. Vukobratović, F. Chiti, L. Niccolai, V. Crnojević, and R. Fantacci, "Urban infrastructure-to-vehicle traffic data dissemination using UEP rateless codes," *Selected Areas in Communications, IEEE Journal on*, vol. 29, no. 1, pp. 94–102, 2011.

[65] F. Li and Y. Wang, "Routing in vehicular ad hoc networks: A survey," *Vehicular Technology Magazine, IEEE*, vol. 2, no. 2, pp. 12–22, 2007.

[66] Y. Xia, C. K. Yeo, and B. S. Lee, "Hierarchical cluster based routing for highly mobile heterogeneous MANET," in *Network and Service Security, 2009. N2S'09. International Conference on.* IEEE, 2009, pp. 1–6.

[67] F. Chiti, R. Fantacci, and R. Mastandrea, "A low complexity clustering approach enabling context awareness in sparse VANETs," in *Global Communications Conference (GLOBECOM), 2012 IEEE.* IEEE, 2012, pp. 61–66.

[68] H. Zhu, S. Chang, M. Li, K. Naik, and S. Shen, "Exploiting temporal dependency for opportunistic forwarding in urban vehicular networks," in *INFOCOM, 2011 Proceedings IEEE.* IEEE, 2011, pp. 2192–2200.

[69] N. Wisitpongphan, F. Bai, P. Mudalige, V. Sadekar, and O. Tonguz, "Routing in sparse vehicular ad hoc wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 25, no. 8, pp. 1538–1556, 2007.

[70] T. Mangel and H. Hartenstein, "5.9 GHz IEEE 802.11 p inter-vehicle communication: Non-Line-of-Sight reception under competition," in *Vehicular Networking Conference (VNC), 2011 IEEE.* IEEE, 2011, pp. 155–162.

[71] T. Nadeem, P. Shankar, and L. Iftode, "A comparative study of data dissemination models for VANETs," in *Mobile and Ubiquitous Systems-Workshops, 2006. 3rd Annual International Conference on.* IEEE, 2006, pp. 1–10.

[72] S. Olariu and M. C. Weigle, *Vehicular networks: from theory to practice*. Crc Press, 2009.

[73] T. De Cola and M. Marchese, "Joint use of custody transfer and erasure codes in DTN space networks: Benefits and shortcomings," in *Global Telecommunications Conference (GLOBECOM 2010), 2010 IEEE*. IEEE, 2010, pp. 1–5.