Spatiotemporal Analysis of Wireless Internet of Things Networks

Vom Promotionsausschuss der
Technischen Universität Hamburg

zur Erlangung des akademischen Grades

Doktor-Ingenieur (Dr.-Ing.)

genehmigte Dissertation

von

Mustafa Muhamed Talat Ali Muhamed Emara

aus

Geddah, Saudi-Arabien

2021

Vorsitzender des Prüfungsausschusses:
Prof. Dr.-Ing. R. Grigat

1. Gutachter:
Prof. Dr.-Ing. Gerhard Bauch

2. Gutachter:
Prof. Marco Di Renzo

3. Gutachter:
Prof. Hesham ElSawy

Tag der mündlichen Prüfung:
26.10.2021

Mustafa Emara
DOI: https://doi.org/10.15480/882.3918

Summary

The wireless future industry will be dominated by a prevalent wireless integration of smart phones, wearables, sensors, tablets, drones, and other objects into a massive integrated system. A plethora of diverse services within every vertical segment is rapidly emerging within the context of the massive Internet of Things (IoT) wireless networks. Over the past decades, the features and functionalities of wireless networks have become more complex, which calls for a new way of thinking via designing the wireless networks from a combined communication and computation perspective. This thesis discusses the need to rigorously study the spatiotemporal dynamics of large scale IoT networks for diverse requirements, deployments, and use cases for both communication and computation pillars. For the communication pillar, analytical models are presented to model wireless networks, while considering random spatial deployment and dynamic temporal traffic models. Use cases capturing prioritized multi-stream, time and event-triggered traffic, and task offloading are considered in this thesis. By virtue of the spatiotemporal analysis, novel spatiotemporal metrics as well as the Pareto frontiers that represent stability region for network operation are derived and discussed.

For the computation pillar, because of the massive number of running services, the presence of demanding computations within the network is inevitable. One solution is to let such computations be executed at a remote data center. However, such approach is not only inefficient due to bandwidth constraints, but also hinders the performance of time-sensitive and location-aware applications due to the imposed network delay. Consequently, we investigate throughout this thesis the advents of Multi-access Edge Computing (MEC)-assisted networks, focusing on the latency and task execution efficiencies in task-offloading use cases. A novel computation-based cell association criterion is proposed to exploit both the communication and the computation resources within a heterogeneous network. Additionally, when it comes to safety-critical use cases within the automotive vertical, it is shown that in contrast to conventional remote cloud-based cellular architecture, the deployment of MEC infrastructure can substantially prune the end-to-end latency as well as the experienced information freshness. Furthermore, joint consideration of contention-based communications for task offloading, parallel computing, and occupation of failure-prone MEC computation resources are inspected. Finally, the availability and reliability of wireless links running diverse services are quantified, with the aim to reveal the incurred performance trade-offs between spatial and temporal resource provisioning.

In summary, this thesis provides a unified insight on modeling, designing, and assessing future wireless IoT networks, while considering different communication and computation key enablers. Such a joined communication and computation perspective is essential to meet the envisaged requirements of the future applications and services in diverse market segments.

Acknowledgments

Completing this thesis, a product of several years of hard work, learning, laughter, and hardships, I feel deeply indebted to many people who have greatly inspired and supported me throughout this journey. I would like to thank my supervisor Prof. Bauch for allowing me to be in his group and for providing valuable feedback throughout the past years. I would like to thank wholeheartedly Prof. Hesham ElSawy who is more than a collaborator for me. He is a friend, older brother, and a mentor. His dedication, work ethics, and motivating attitude is one of the main reasons I was able to successfully complete this thesis. I am lucky to have worked and collaborated with him and I wish our friendship will strengthen with the years. I would also like to thank Prof. Marco Di Renzo for agreeing to be on my PhD committee, his research impact has inspired me greatly when I was starting my research work.

I would like to express my sincere gratitude to my colleagues at Intel, especially Michael, Kilian, Markus, Leo, Ingolf, Miltos, Honglei, and Dario for the support throughout my four and half years at Intel. I was lucky to have Michael as my manager for more than four years, he was always there for me providing guidance on the personal and professional levels. His managerial style allowed me to have the required space to pursue different research problems with complete freedom. Markus has also hepled and supported me a lot during my last year at Intel. Sharing the office with Kilian as well as most of this journey with him has been a crazy and an interesting ride, just like one of our skiing trips. I hope our friendship will remain strong throughout the years. I was grateful to be collaborating with Miltos over my years at Intel. I am also quite thankful for my colleagues at Qualcomm, especially Marco, Michael, Lu Gao, and Ebraam for their support and effort on helping me complete the final milestone of this journey.

I owe hugely to my dear family, Talaat, Wafaa, Yasmeen, Yousra, and Ahmed. Their permanent love and confidence in me have encouraged me to go ahead in my study and career. I hope I can express my love and appreciation to you in the future. Your support and sacrifices have made it possible for me to successfully finish this chapter. I will always be grateful and indebted for all what you have done for me. I would like also to thank my dear friends and brothers as well for so many years, Hazem, Tarek, Karim, Hossam, Omar, Ehab, Alaa, and others, it feels always like home when we are together.

Finally, to Nada, my fiancé and soon wife-to-be, thanks for being there with me throughout the challenging phase of this journey. Your support and kindness made it easier to overcome the final lap of this long race. I am grateful to be sharing this with you and I look forward to our future together.

Table of contents

Abbreviations

AoI	age of information
ACK	acknowledgment
AMF	Access and mobility management function
ASF	Authentication server function
BS	base station
BH	back-haul
CCDF	complementary cumulative distribution function
C-V2X	cellular-V2X
CAM	cooperative awareness message
CTMC	continuous time Markov Chain
DTMC	discrete time Markov chain
EA	equal allocation
ETSI	European telecommunication standards institute
eMBB	enhanced mobile broadband
5G	fifth generation
FCFS	first come first serve
FDMA	frequency division multiple access
Geo	geometric
IoT	internet of things
I/O	input/output
KPI	key performance indicator
LT	Laplace transform
LTE	long term evolution
MAM	matrix analytic method
MAC	medium access control
MEC	multi-access edge computing

MGF moment generating function

mMTC massive machine type communications

NACK non-acknowledgment

NSSF Network slice selection function

NRF Network repository function

NEF Network exposure function

NR new radio

PPP Poisson point processes

PDF probability density function

PMF probability mass function

PA priority agnostic

PAoI peak age of information

PH phase

PCF Policy control function

QoS quality of service

QCI QoS class identifier

QBD quasi-birth-death

RAN radio access network

RSRP reference signal received power

SINR signal to interference noise ratio

SIR signal to interference ratio

SMF Session management function

3GPP third generation partnership project

TSP transmission success probability

TDMA time division multiple access

URLLC ultra reliable low latency communication

UDM Unified data management

UPF User plane function

VRU vulnerable road user

V2X vehicle-to-everything

VM virtual machine

WA weighted allocation

WLAN wireless local area network

Nomenclature

$\mathbf{s}$	Absorption vector of a PH type distribution
$\tilde{\mathbf{v}}_i$	Absorption vector of the i-th priority queue
P_a	Active probability of a device
1_m	All ones vector of size m
$\mathcal{I}_m$	All ones matrix matrix of size $m \times m$
0_m	All zeros vector of size m
$\Delta(t)$	Age of information at the t-th time stamp
$\mathcal{A}_{\text{dep}}$	Area of deployment
λ_a	Average task/packet arrival rate at an IoT device
$\mathcal{Y}$	Average ratio of packet arrival to packet service
κ	Average number of devices per BS per channel
μ_{ser}	Average packet's service rate at the BS
C_{BH}	Backhaul system capacity
$\mathcal{B}$	Bandwidth available at a given BS
$\mathcal{B}_{k,i}$	Bandwidth allocated to the k-th device when served by an i-th tier BS
$\Omega_{\mathbf{i},\mathbf{j}}(\theta,\xi,\Phi)$	Binary spatial availability function between the locations i and j: $1_{\{\mathbb{P}\{\text{SIR}_{\mathbf{i},\mathbf{j}} \geq \theta\} \geq \xi\}}$
$\mathbf{B}_{1,\mathbf{i}}, \mathbf{C}_{\mathbf{i}}, \mathbf{B}_{2,\mathbf{i}}$	Boundary sub-stochastic matrices for the i-th priority queue
$F_X(\cdot)$	CDF of the random variable X
h	Channel power gain of the intended signal
g	Channel power gain of the interfering signal
$h_{i,j}$	Channel power gain between the i-th transmitter and j-th receiver
$\mathcal{H}_{k,M}$	Closest cluster of M vehicles of the k-th VRU
$\bar{(\cdot)}$	Complement operator (i.e., $\bar{a} = 1 - a$)
$\mathcal{C}$	Computational power of a MEC host in cycles/second
φ	Convergence tolerance parameter
$\mathcal{A}_t$	Computational resources availability
d	Computation degradation factor within a MEC host
d_i	Departure probability of the i-th priority queue
$\tilde{d}_n$	Discretized departure probability of the n-th QoS class
T	Duty cycle for time-triggered traffic generation
$\mathcal{T}^k_{\text{E2E, C}}$	End-to-end latency of the k-th device's for the conventional deployment
$\mathcal{T}^k_{\text{E2E, MEC}}$	End-to-end latency of the k-th device's for the MEC deployment
$\mathcal{E}$	Extended packet delay budget

$\|\|\cdot\|\|$	Euclidean distance operation
$r_{i,j}$	Euclidean distance between i and j; $i, j \in \mathbb{R}^2$
$\mathcal{T}^{\text{exc}}_{k,i}$	Execution latency of the k-th device's packet at the i-th BS
$\mathbb{E}_X\{\cdot\}$	Expectation of the random variable X
$\mathcal{F}$	Failure rate of a VM within a MEC host
$\mathcal{Y}_{k,i}$	Fraction of computational resources allocated to the k-th device at an i-th tier BS
${}_2F_1(a,b,u;z)$	Gaussian hyper-geometric function that is defined as ${}_2F_1(a,b,u;z) = \sum_{k=0}^{\infty} \frac{(a)_k (b)_k z^k}{(u)_k k!}$
G_i	Generation time of the i-th packet at its source
α	Geometric random variable for packet generation event
$\odot$	Hadamard's product
$\mathcal{I}$	Inter-arrival time between two consecutive packets
$\mathbf{I}_m$	Identity matrix matrix of size $m \times m$
$1_{\{\cdot\}}$	Indicator function
ρ	Initialization vector of a given PH type distribution
$\mathcal{J}_0$	Joint probability of no packets in all N priority queues
$\mathcal{J}_i$	Joint probability of having no packets with priority higher than i and at least a single i-th priority packet
$L(\cdot)$	Lebesgue measure
$\mathbf{R}$	Matrix analytic method matrix
$\mathbf{M}$, $\mathbf{m}$	Matrix and vector representation
$\mathbb{M}_X\{\cdot\}$	Moment generating function of the random variable X
$M_{b,E}$	Moments of the transmission success probability under time triggered traffic for $b = 1,2$
$M_{b,T}$	Moments of the transmission success probability under time triggered traffic for $b = 1,2$
σ^2	Noise power
x_F	Number of failed VMs within a MEC host
x_I	Number of idle VMs within a MEC host
x_O	Number of occupied VMs within a MEC host
$\mathcal{K}$	Number of tiers in a heterogeneous network
C	Number of uplink channels available at the BS
$\mathcal{V}_{\text{loc}}$	Number of VMs in a device
$\mathcal{V}_{\text{MEC}}$	Number of VMs in a MEC host
$\mathcal{N}_{\text{VRU}}$	Number of deployed VRUs within the network
Q_i	Number of packets in the i-th priority queue
f_k	Number of processing operations per input bit for k-th device task
N	Number of priority classes
$\tilde{N}$	Number of QoS classes
$\mathcal{O}$	Offloading success probability
l_k	Packet/ task size in bits of the k-th device
Δ_{p}	Peak Age of information
η	Path-loss exponent, where $\eta > 2$
ε	Path-loss inversion compensation factor

$f_X(\cdot)$	PDF of the random variable X
Θ_E	Percentile of active devices for event-triggered traffic
Θ_T	Percentile of active devices for time-triggered traffic
Θ_τ	Percentile of active devices with $\nu \neq \tau$
ξ	Percentile of devices that achieves an $\mathsf{SIR} \geq \theta$
β	PH type distribution initialization vector
$\mathbf{S}$	PH type distribution transient matrix
Φ	PPP modeling base stations locations in $\mathbb{R}^2$
Ψ	PPP modeling devices locations in $\mathbb{R}^2$
$\mathbf{P}_i$	Probability transition matrix for the i-th priority/class
$\mathcal{T}_{k,i}^{\text{radio}}$	Radio latency between the k-th device and i-th BS
$\mathcal{G}^{\text{c}}$	Ratio of computational power of two consecutive tiers $\mathcal{G}_i^{MECTx} = \frac{\mathcal{C}_i}{\mathcal{C}_{i+1}} > 1$
$\mathcal{G}^{\text{r}}$	Ratio of transmission power of two consecutive tiers $\mathcal{G}_i^{\text{Tx}} = \frac{P_i}{P_{i+1}} > 1$
$\mathbb{R}^2$	Real coordinate space of dimension 2
$\mathbb{E}_X^!\{\cdot\}$	Reduced Palm Expectation of random variable X
μ_r	Relative computational rate ($\mu_r = \mu_{\text{MEC}}/\mu_{\text{loc}}$)
$\mathcal{R}$	Repair rate of a VM within a MEC host
$\mathbf{S}$	Sub-stochastic matrix of a PH type distribution
$\mathbf{S}_i$	Stochastic non-boundary transient matrix of the i-th priority queue
$\mathbf{S}_{0,i}$	Stochastic boundary transient matrix of the i-th priority queue
$\mathcal{A}_s$	Spatial availability
$\mathcal{S}_{\text{s}}$	Set of stable QoS classes under time-triggered traffic
$\mathcal{U}_{\text{u}}$	Set of unstable QoS classes under time-triggered traffic
$\mathcal{H}_{M,k}$	Set of closest M vehicles to the k-th VRU
q_i	Size of the i-th priority queue
λ	Spatial intensity of base stations
μ	Spatial intensity of devices
μ_{veh}	Spatial intensity of vehicles
τ	State probability vector for a given CTMC
$\mathbf{x}_i$	Steady state probability for the i-th priority/ class queue
$\mathcal{S}_i$	Set containing the number of i-th priority packets
$\mathcal{V}_i$	Set containing the tuples of number of packets with priority higher than i
$\mathcal{Y}_{\text{v}}$	State space for the proposed vacation-based priority queues
$\mathcal{Q}$	State space for a given CTMC
$\mathbf{A}_0$	Sub-stochastic matrix that captures backward transitions within a QBD
$\mathbf{A}_1$	Sub-stochastic matrix that captures same-level transitions within a QBD
$\mathbf{A}_2$	Sub-stochastic matrix that captures forward transitions within a QBD
θ	Successful decoding threshold
$\mathcal{A}_t$	Temporal availability
μ_{loc}	Task execution rate at a local device
μ_o	Task execution rate of a single VM within a MEC host

μ_{MEC}	Task execution rate at a given VM within a MEC host
T_s	Task instruction transmission time
$\mathcal{C}_t$	Task execution capacity
$\mathcal{R}_t$	Task execution retainability
$I_\xi(a,b)$	The regularized incomplete beta function $I_\xi(a,b) = \int_0^\xi t^{a-1}(1-t)^{b-1}\mathrm{d}t$
$\gamma(a,b)$	The lower incomplete gamma function $\gamma(a,b) = \int_0^b t^{a-1}e^{-t}\mathrm{d}t$
$[\mathbf{A}]_{i,j}$	The i-th row and j-th element of the matrix $\mathbf{A}$
ν	Time-offset for time-triggered traffic
$\mathcal{R}_t$	Temporal reliability
$\mathcal{A}_i$	Transmission availability of the i-th priority queue
D_i	Transmission delay of the i-th priority queue
P	Transmission power of a transmitter
$\mathbf{Q}$	Transition rate matrix of a given CTMC
P_{s}	Transmission success probability
$P_{\text{s},i}$	Transmission success probability of the i-th priority queue
$\mathcal{T}_i$	Transmission access probability of the i-th priority queue
ρ	Uplink power control threshold
$\mathbf{V}_i$	Vacation visit matrix of the i-th priority queue
χ_i	Vacation probability of the i-th priority queue
$\mathbf{v}_i$	Vacation initialization vector of the i-th priority queue
$\mathcal{W}$	Waiting time of a packet in its queue

Chapter 1

Introduction

1.1 Background and Motivation

The evolution of mobile networks is characterized by a growing traffic demand (currently dominated by video content [1]), a paradigm shift in the consumed services, where content sharing and social behavior are redefining network utilization, and a massive number of devices [2]. The roll-out of fifth generation (5G) systems will witness a dramatic increase of device-to-device connections [3] due to the progressive increase of internet of things (IoT) traffic and services, that will be dominated by several new vertical business segments (e.g., automotive and mobility, factories of the future, health-care, media, and entertainment) [4]. As a result, the need for efficient use of network resources is continuously increasing. However, current mobile systems have been planned and deployed so far with the mere aim of enhancing radio coverage and capacity [5]. Hence, future networks will need to effectively support heterogeneous services, variable in both space and in time.

The massive deployment of IoT devices such as smart-phones, tablets and wearables, along with advanced wireless network capabilities, has led to prominent research in the field of wireless communication to address the resulting challenges [6]. Despite steady improvement in the capabilities of the hardware components (e.g., computing units, battery and memory), the majority of IoT devices are still not capable of fully supporting the requirements of the emerging computation-intensive and delay sensitive applications [7]. Additionally, the realization of the network objectives of reliable communication, low latency, and efficient computation, significantly relies on efficient network design, analysis, and optimization, where a joint communication-computation perspective should be considered. Throughout this thesis, we aim to address this joint point of view by first, modeling the envisaged large scale IoT networks and second, by augmenting the computing capabilities of devices by allowing them to use remote cloud servers to address their computation demanding tasks.

1.1.1 IoT and Massive Connectivity

The IoT paradigm is paving the way to ensure seamless connectivity, efficient networking, and continuous monitoring within different market segments [8]. Emerging segments entail, among other examples, smart cities, connected vehicles, industrial IoT, health-care, and smart homes, which are all tied with the IoT technology advancement [4]. According to a recent analysis by Cisco, the number of mobile devices

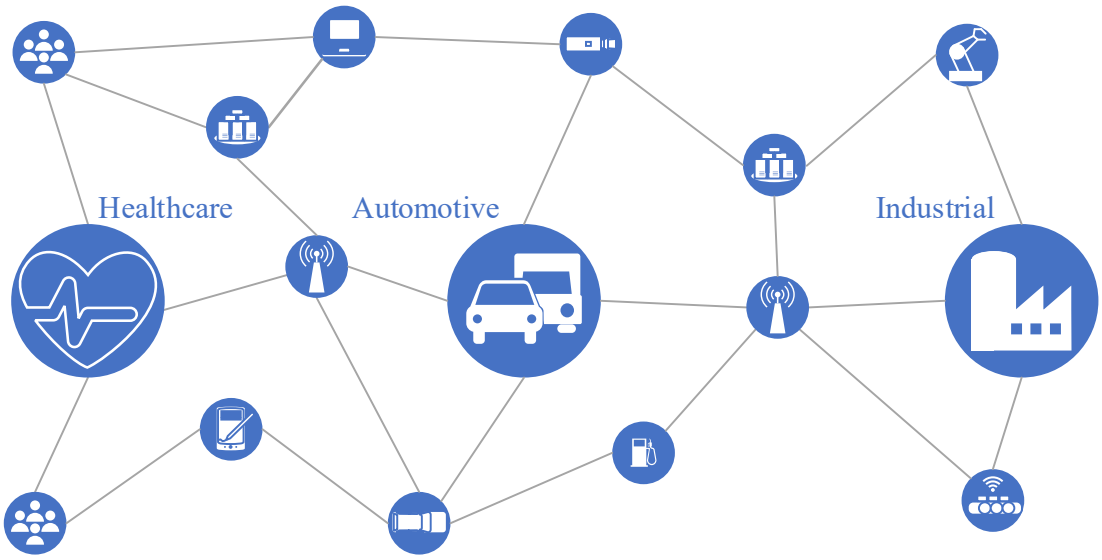


Figure 1.1: Envisaged IoT penetration in different vertical segments.

will grow to 13.1 billion by 2023 at a compound annual growth rate of 8% between 2018 and 2023 [9]. Moreover, it is predicted in the same study that 5G systems will support more than 10% of global mobile connections. Such growth is fueled by the aforementioned emerging markets. Such markets generate services that have characteristics which may be bandwidth-hungry (surveillance, video conferencing, traffic monitoring), latency-critical (industrial IoT, autonomous vehicles, human-machine interaction), and may cause spatial or temporal activity spikes (e.g., sporting events).

In Figure 1.1, an envisioned ecosystem is shown, in which a versatile and massive collection of silos, devices and network components is interconnected. In the health-care silo, applications related to health, patient monitoring and identification or collection of medical data are expected to coexist. Health-care applications are generally characterized by low mobility, medium data rates and high service availability and reliability. In the automotive silo, connected and autonomous vehicles are envisioned to form, along with smart stations and road side units, a connected network, which is characterized by high mobility, low latency and jitter and high reliability. Finally, the industrial silo encompasses applications that are designed to optimize the plant, logistics and supply chain management, which are characterized by stringent delay and reliability constraints. In addition, the deployment of edge servers within the ecosystem is expected, in order to extend the computation resources within the network to the edge. Thus, lower network congestion, improved resource optimization, and enhanced user experience can be realized. By leveraging the radio access network (RAN), edge deployment will improve significantly the experienced latency, yielding efficient resources utilization, thus, accommodating a larger number of services [10]. Owing to the large scale IoT deployment, as depicted in Figure 1.1, one can deduce that current 5G and beyond systems, will be poised to induce a significant surge in demand for networking infrastructures, computation resources and data, in order to accommodate the anticipated requirements of the heterogeneous running applications [11].

1.1.2 The Communication and Computation Pillars

For many years, cellular networks have been evolving to offer global coverage, quality of service (QoS) support, security, low cost of deployment, scalability, mobility and roaming support. The third generation partnership project (3GPP) standards development organization covers the radio access and the core transport network, and defines the interfaces for non-3GPP networks [12]. Focusing on 5G-new radio (NR), it is globally considered the first cellular technology designed to provide efficient coexistence of diverse 5G services with highly heterogeneous requirements [13]. In particular, new service classes have been introduced targeting to address services spanning enhanced mobile broadband (eMBB), massive machine type communications (mMTC) and ultra reliable low latency communication (URLLC) [14]. The eMBB service addresses the bandwidth-hungry, data-intensive and human-centric use cases. The demand within this use case is continuously increasing due to the emerge of new applications with more stringent requirements (e.g., virtual, augmented and mixed reality). On the other hand, URLLC considers latency-sensitive services that necessitate extremely high service reliability and availability, as found in industrial automation, autonomous vehicles and machine-centric applications. Finally, mMTC entails providing spectral and energy efficient connectivity to a massive number of low cost and low energy IoT devices with sporadic traffic.

In parallel, the wireless local area network (WLAN) technology (i.e., IEEE 802.11) has been since many years and continues to be a first choice for fixed and local area broadband wireless access for home and enterprise networks, and being an alternative technology in terms of data throughput performance, cost, and interoperability support [15]. Investigating both, WLAN and 3GPP, the evolution of radio access technologies got driven primarily by peak data throughput following timely the Ethernet road-map with an offset of many years as shown in Figure 1.2. Meanwhile, the advancement of the communication peak data throughput by WLAN and 3GPP along with the advancement in the computing performance, quantified via floating point operations per second, delivers some motivating insights regarding the relationship between the communication and computation growth. It is evident that computing, supported by reliable communication, is the cornerstone for realizing the promised gains. It can be also articulated that such a computation-communication ecosystem will be hindered by the communication systems capabilities to meet the imposed demand. Thus, one of the main challenges of next generation systems is to offer sufficient understanding, novel solutions, and architectural designs that balance efficiently computation performance with the communication and access capabilities.

A result of the staggering number of running applications and services is the presence of demanding computations within the network. One solution is to let such computations be executed at a remote data center. However, such approach is not only inefficient due to bandwidth constraints but also hinders the performance of time-sensitive and location-aware applications due to the imposed network delay for offloading data to the cloud, and computation dependencies between data generated by nearby sensors. A natural alternative would to bring the computation resources closer to the devices, so the network delays can be alleviated and the additional computation resources can be efficiently utilized. In this context, the challenges of large scale IoT deployment in future wireless systems that will be addressed throughout this thesis are summarized as follows.

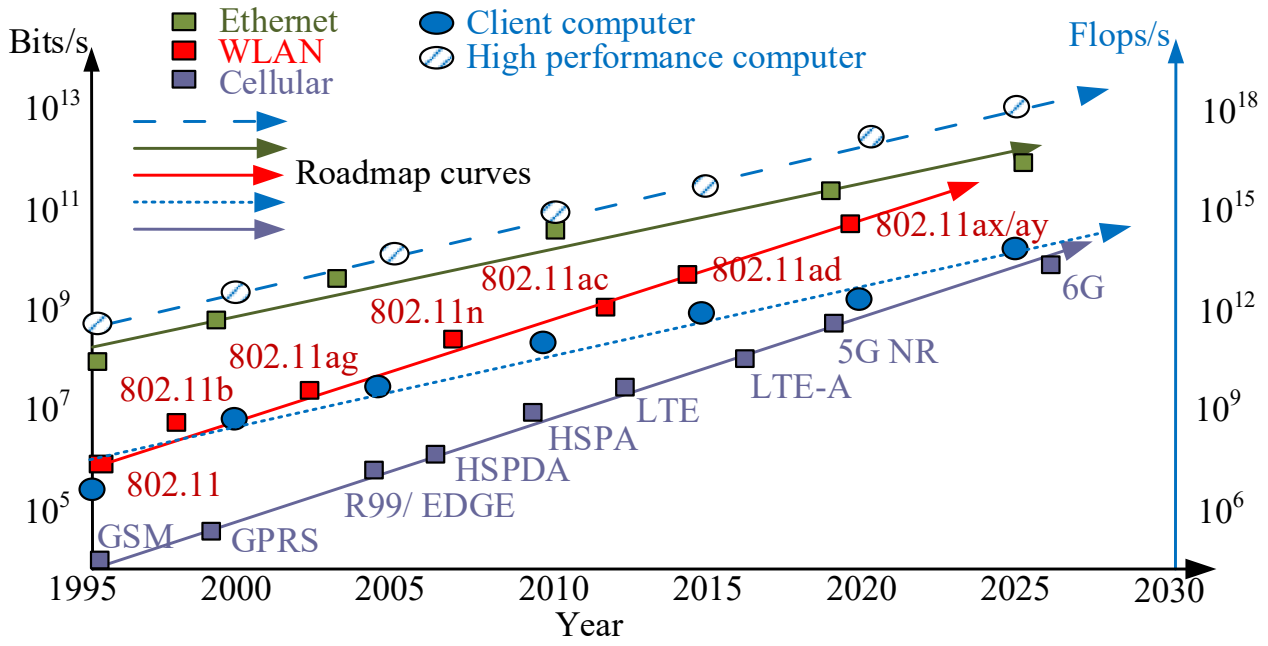


Figure 1.2: Computing trends, cellular, WLAN and Ethernet technology road-maps [16].

- Massive connectivity: based on the anticipated number of coexisting services, massive number of IoT devices will be trying to utilize the scarce available computation and communication resources. Due to the sharing nature of the wireless channel, understanding the effect of such large-scale access within the deployment area is paramount to meet the targeted requirements.
- Services heterogeneity: resulting from the wide range of running applications and services impose the necessity that each service be properly differentiated and addressed. This results in the need to support diverse characteristics in terms of mobility, latency, power efficiency, traffic, number of devices and computation requirements.
- Demanding computation tasks: massive devices coexist with varying profile activity, where some of them are idle, others are connected, while running voice/ video/ data services, whereas others aim to access the network in order to delegate processing tasks that cannot be locally addressed in a timely fashion.
- Spatiotemporal randomness: exists based on the anticipated highly dynamic network topology. Temporal and spatial randomness are inevitable in the dynamic geographical deployment, traffic burstiness, and tasks arrival and departure.

1.2 Scope and Organization

In order to address the identified challenges, this thesis aims at addressing the communication and computation pillars from different perspectives. We utilize analytical tools from stochastic geometry, queueing theory, and reliability theorem to model, analyze and investigate large scale IoT networks. *For the communication pillar*, our objective is to develop general, flexible, and rigorous frameworks for different traffic models under a spatiotemporal viewpoint. Such frameworks are used for analyzing the

system behavior with respect to the different system design parameters. Consequently, many design insights and trade-offs for the considered networks are obtained. *For the computation pillar*, we propose architectural modifications to enable efficient and faster task execution via task offloading to the multi-access edge computing (MEC) hosts, co-located at the base stations (BSs). In addition, we showcase MEC latency reduction gains through a deployed vehicular network. Moreover, novel key performance indicators (KPIs) related to dependable network functionality are considered in order to provide a novel service-centered perspective of the network. To this end, the contributions and organization of the thesis can be summarized as follows:

Chapter 2

This chapter provides the foundations of the analytical tools and key enablers that are utilized throughout the thesis. First, the spatial, temporal and spatiotemporal modeling frameworks and their detailed building blocks are discussed. The spatial and temporal models encompass stochastic geometry and time Markovian models, respectively. Afterwards, the different KPIs that are employed within the thesis are presented and explained in order to provide a holistic overview on the performance assessment conducted in later chapters. Finally, an introduction to MEC technology and its architectural components within cellular networks is presented to lay out the foundations of the computation pillar proposed within this thesis.

Chapter 3

This chapter develops a novel priority-aware spatiotemporal framework to characterize large scale IoT uplink networks with prioritized multi-stream traffic. A systematic and tractable scheme to model the prioritized traffic is introduced. Such a scheme alleviates the curse of dimensionality resulting from the state of the art schemes. Additionally, dedicated and shared channel priority-aware access strategies are presented, and bench-marked against a priority-agnostic scheme. The impact of prioritized uplink transmission on the performance of different priorities is highlighted, in terms of transmission probabilities and delay. Additional performance metrics as average number of packets, peak age of information (PAoI), delay distribution, and Pareto frontiers for different parameters are presented, which give insights on stable operation of uplink IoT networks with prioritized traffic.

Chapter 4

Moving to time and event-triggered data traffic models, this chapter provides a spatiotemporal framework that captures the information freshness within large scale IoT uplink networks. In contrast to the typical user analysis that is conducted in Chapter 3, this chapter builds upon the idea of the meta distribution of the transmission success probability and derives key expressions for the location-dependent performance of devices under the two mentioned traffic variants. Numerical evaluations are conducted to validate the proposed mathematical framework and assess the effect of traffic load on the realized information freshness. The results unveil a counter-intuitive superiority of the event-triggered traffic over the time-triggered one in terms of information freshness, which is due to the underlying temporal interference

correlations. Insights regarding the network stability frontiers and key design recommendations are presented and discussed.

Chapter 5

Moving to the computation pillar, this chapter advocates the advents of MEC deployment to reduce the experienced latency within heterogeneous and vehicular networks. First, a novel computation-based cell association criterion is proposed to exploit both the communication and the computation resources within a heterogeneous network. It is shown that, for a range of disparities between radio and MEC capabilities between tiers, the proposed computation association criterion provided gains in terms of the experienced one-way latency, as compared to the conventional association criterion. Additionally, when it comes to safety-critical use cases within a vehicular network, we showcase that in contrast to conventional, remote cloud-based cellular architecture, the deployment of MEC infrastructure can substantially prune the end-to-end communication latency and the experienced information freshness.

Chapter 6

Focusing on successful task execution, in this chapter, novel definitions of dependability attributes for communication and computation services are provided. First, joint consideration of contention-based communications for task offloading and parallel computing as well as the occupation of failure-prone MEC computation resources are inspected. The influence of various system parameters on dependability metrics such as (i) computation resources availability, (ii) task execution retainability, and (iii) task execution capacity are investigated. Second, the availability and reliability of wireless links, running diverse services, are quantified via a novel spatiotemporal framework, with the aim to reveal the incurred performance trade-offs between spatial and temporal resources provisioning.

Chapter 7

Finally, this chapter summarizes the thesis. A summary of the different proposed spatiotemporal frameworks and their key results are discussed. The impact of MEC deployment on different performance metrics is reviewed and different task execution dependability insights are recapitalized. Finally, future research directions are pointed out.

Chapter 2

Wireless Networks Modeling

Modeling and analysis of wireless networks is a comprehensive topic that entails an enormous number of building blocks, especially for the envisaged use cases that combine the communication and computation aspects of future network. To yield the overview relevant to the thesis scope, we focus throughout this chapter on key concepts and tools that are employed within the coming technical chapters. First, concepts addressing spatial, temporal and spatiotemporal models of cellular networks are presented in Section 2.1. Afterwards, the different KPIs adopted throughout this thesis are presented and explained in Section 2.2. Finally, MEC technology, its architecture and relation to the IoT technology advancement is discussed in Section 2.3.

2.1 Preliminaries for Cellular Networks Modeling

Throughout this section, some mathematical preliminaries and tools that are utilized throughout the thesis will be presented. The spatial and temporal models are discussed in Subsections 2.1.1 and 2.1.2, respectively, whereas the fused spatiotemporal perspective is provided in Subsection 2.1.3.

2.1.1 Spatial Modeling

To address the exploding and diverse heterogeneous service requirements, network parameters (e.g., physical and medium access control (MAC)) should be carefully modeled and optimized through a collective cross-layer framework in order to showcase meaningful insights. The modeling phase aims at obtaining mathematical expressions that characterize the network behavior. The inputs for the modeling expressions are the network parameters (i.e., network geometry, MAC protocol, propagation environment, etc.) and the outputs are the different KPIs of interest. Afterwards, performance analysis can be conducted, in which the system response to different network parameters is analyzed, to understand the system behavior, performance trends, trade-offs, and design insights. Analytical frameworks that characterize the performance of large scale networks provide a trade-off between model practicality in mimicking real networks and complexity. Such models in the literature were facilitated by resorting to simplifications such as the Wyner model [17], which considered only one or two interfering cells. Other models such as [18] aggregated the network-wide interference into a single random variable that is then empirically fit to some distribution. Equidistant interference was considered in [19], which is far from

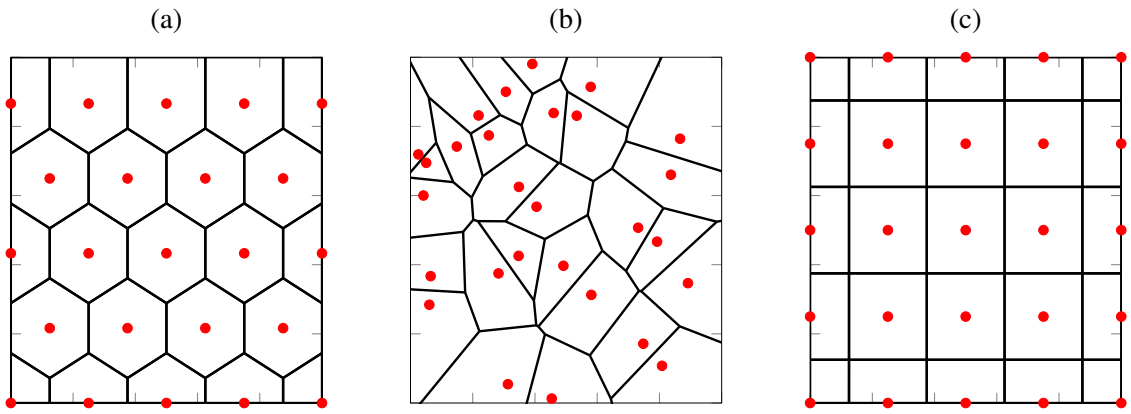


Figure 2.1: Cellular deployment variants: (a) hexagonal (b) PPP (c) square. BSs and their coverage regions are depicted via red circles and black lines, respectively.

actual network deployments. Another widely accepted model is the grid-based model, which is suited for large-scale infrastructure-based wireless networks [20]. In the grid-based model, the BSs are deployed on a hexagonal or square lattice as shown in Figure 2.1(a) and (c), respectively. Nevertheless, real network planning and deployment are far from being uniform as highlighted in [21]. Such disparity from real deployments increases in the case of heterogeneous BSs deployment, where each tier is characterized by different radio capabilities.

An alternative to the aforementioned simplified models is simulations, which aim at exhaustively complex simulations in order to average out the many sources of randomness, such as fading distributions, noise, and BSs and device locations. These simulations can be extremely time consuming and generally require continuous human maintenance and development. Although system-level simulations will continue to be indispensable for cellular network analysis and design, the need for a complementary analytical approach for the purposes of bench-marking and comparison has long been called for. In this context, stochastic geometry is a powerful tool that has been utilized in the last decade to capture the random network's topology and to provide tractable yet accurate expression of different KPIs [22, 21]. Specifically, stochastic geometry study the spatial average performance, over large enough spatial realizations of a network, whose nodes (i..e, BSs, devices or both) are deployed following a predetermined distribution [23]. Spatially-averaged performance implies that each spatial deployment realization is weighted by its probability of occurrence [24].

To this regards, point processes are employed to mimic the spatial distribution of the BSs and devices [25]. The Poisson point processes (PPP) is regarded as the commonly adopted point process when modeling wireless network due to its mathematical tractability [26]. Due to its spatial randomness, alternative point processes are adopted to model repulsion between the points, such as the Matérn hard core point process [25] and the Poisson cluster point process. Nevertheless, such non-PPP based models are not as tractable as the PPP-based models. An example of a network deployment based on a PPP is shown in Figure 2.1(b). The PPP-based deployment are advantageous compared to grid-based models when it comes to deriving mathematical expressions for complex network scenarios. Such expressions are cumbersome to realize when dealing with the location-dependent grid-based models.

In what follows, we briefly highlight some of the key mathematical properties of the PPP. Let $\Lambda = \{\mathbf{x}_i;\, i \in \mathbb{N}\}$ be a PPP, which is a countably-finite collection of points in the d-dimensional Euclidean

space $\mathbb{R}^d$, where $\mathbf{x}_i \in \mathbb{R}^d$ represents the coordinates of the i-th point. Given a generic set of points, denoted by $\mathcal{A} \in \mathbb{R}^2$, the number of points $\Xi(\mathcal{A}) = |\Lambda \cap \mathcal{A}| \sim poisson(\lambda)$, with the following probability mass function (PMF)

$$\mathbb{P}\{\Xi(\mathcal{A}) = k\} = \frac{(\lambda L(\mathcal{A}))^k}{k!} \mathrm{e}^{-\lambda L(\mathcal{A})}, \tag{2.1}$$

where λ is the intensity of the PPP and $L(.)$ is the Lebesgue measure [22] [1]. An additional major property of the PPP is that for another set of points $\mathcal{B} \in \mathbb{R}^2$, such that $\mathcal{A} \cap \mathcal{B} = \emptyset$, the number of points in each set, i.e., $\Xi(\mathcal{A})$ and $\Xi(\mathcal{B})$, are independent. Throughout this thesis, we will consider only homogeneous PPP, which is a class of PPP that is only characterized via its intensity measure λ [24]. In the following, we present some of the main statistical properties of a homogeneous PPP that are utilized in this thesis. For a more detailed study on this subject, the reader is kindly referred to [23–25].

- **Campbell's theorem**: converts an expectation of a random sum over a PPP to an integral. This enables the computation of the aggregate interference in a network. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a measurable function, then Campbell's theorem states that

$$\mathbb{E}\left\{\sum_{x_i \in \Lambda} f(x_i)\right\} = \int_{\mathbb{R}^d} \lambda f(x) \mathrm{d}x. \tag{2.2}$$

- **Probability generating functional**: converts an expectation of a random product over a PPP to an integral. Such a property enables the Laplace transformation of the aggregate interference. Let $f : \mathbb{R}^d \to [0,1]$ be a real-valued function, then the probability generating functional states that

$$\mathbb{E}\left\{\prod_{x_i \in \Lambda} f(x_i)\right\} = \exp\left(-\lambda \int_{\mathbb{R}^d} (1 - f(x))\, \mathrm{d}x\right). \tag{2.3}$$

- **Slivnyak's theorem**: states that for a PPP Λ, because of the independence between all of the points, conditioning on a point $\mathbf{x}_i$ does not change the distribution of Λ. In other words, a PPP observed from any generic location remains the same irrespective of having a point on that location. In practice, Slivnyak's theorem enables the analysis of cellular networks, as the case of downlink where it allows the treatment of interference as coming from a PPP despite removing the serving BS from that PPP. Such an equivalence is expressed mathematically as $\mathbb{P}^!\{a\} = \mathbb{P}\{a\}$, where $\mathbb{P}^!\{a\}$ is the reduced Palm probability of event a [25].

- **Independent thinning**: the thinned point process Λ_{thin} obtained from Λ by randomly and independently removing some points with probability p is a PPP with intensity $p\lambda$. Independent thinning can be applied to capture the portion of active devices within the network as will be shown in the coming sections.

- **Displacement**: let $\mathbf{F}_{x_i}$ be a random translation on the PPP Λ such that its distribution depends on $\mathbf{x}_i$, then the resulting displaced point process Λ_{dis} is a PPP such that $\Lambda_{\text{dis}} = \{\mathbf{x}_i \in \Lambda : \mathbf{x}_i + \mathbf{F}_{\mathbf{x}_i}\}$. The displacement property is adopted in modeling mobility and uplink interference in wireless network.

[1] The Lebesgue measure for $d = 2$ represents the area of a given set.

To this end, network deployment can be modeled depending on the system model that is of interest. However, an underlying limiting aspect of the stochastic geometry based models is the *full buffer* assumption, which assumes that the transmitter has always backlogged packets to be transmitted [22, 21, 27, 23]. In reality, a device oscillates between idle and active states, depending on the underlying traffic model and the experienced radio conditions. Thus, conventional models based on stochastic geometry are oblivious to the temporal traffic evolution and the underlying queueing dynamics at each device. Before delving into the coupled temporal fluctuation at the nodes with their large scale spatial characterization, we briefly present some aspects from queueing theory that will be utilized throughout this thesis.

2.1.2 Temporal Modeling

Queueing models are utilized to account for the temporal dynamics of packets (i.e., arrival, waiting and service) and their contention at a communication node [28]. The design of a queueing model may become very complex depending on its features, which may or may not be common in other types of models. Throughout this thesis we consider a single node queueing model, which is characterized by the arrival process (A), the service process (B), the number of hosts (C) in parallel, the service discipline (D) and the queue size (E). In this regard, using the famous Kendall's notations, a single node queue can be represented as *A/B/C/D/E*. The arrival process of a queueing model describes the distribution of the inter-arrival times of packets (or users), as well as how many packets (or users) arrive simultaneously [29]. For continuous systems, the most common assumed arrival process is the Poisson process, in which the inter-arrival times for individual devices or class, are independent and identically distributed as exponential random variables with a shared average rate [30]. This distributional assumption is often prized for its memory-less property, among other features, which allow for tractable analysis [28]. For discrete systems, the phase (PH) type have been considered as a fundamental corner stone in stochastic modeling as it allows numerical tractability of some difficult problems and in addition several distributions encountered in queueing seem to resemble the PH distribution [31]. In particular, every PH type distribution can be represented by the tuple $(\rho, \mathbf{S})$. In this regard, let us consider an $m+1$ absorbing discrete time Markov chain (DTMC) with a state space $\mathcal{J} = \{0, 1, 2, \cdots, m\}$ and let state 0 be the absorbing state. The vector $\rho = [\rho_1 \ \rho_2 \ \cdots \ \rho_m]$ is the initialization vector, such that $\rho 1_m = 1$. In particular, ρ_i is the probability that the system starts from a transient state i, $1 \leq i \leq m$, and $S_{i,j}$ is the probability that the system transitions from the i-th transient state to the j-th transient state. In this regard, $\mathbf{S}$ is an m-dimensional sub-stochastic transient matrix constructed as follows[2]

$$\mathbf{S} = \begin{bmatrix} S_{1,1} & S_{1,2} & \cdots & S_{1,m} \\ S_{2,1} & S_{2,2} & \cdots & S_{2,m} \\ \vdots & \vdots & \cdots & \vdots \\ S_{m,1} & S_{m,2} & \cdots & S_{m,m} \end{bmatrix}. \tag{2.4}$$

[2] A matrix in which the elements is each row sum to at most one is denoted as sub-stochastic matrix.

Moreover, a PH type distribution is defined as an absorbing Markov chain [29], which is defined mathematically as

$$\mathbf{T} = \begin{bmatrix} 1 & 0 \\ \mathbf{s} & \mathbf{S} \end{bmatrix}, \tag{2.5}$$

where $\mathbf{s} \in \mathbb{R}^{m\times 1}$ represents the absorption probability from a given transient state and is given by $\mathbf{s} = 1_m - \mathbf{S}1_m$. Before delving into how to represent different traffic models via the PH type distribution, let $\alpha \in (0,1]$ be a geometric random variable that models the packet arrival probability, i.e. $\mathbb{P}\{\text{packet arrival}\} = \alpha$ and $\mathbb{P}\{\text{no packet arrival}\} = \bar{\alpha}$. Examples of different traffic distributions utilizing the PH type distribution are as follows:

- **Negative binomial ET traffic:** $\rho = [1\ 0\ \cdots\ 0]$ and

$$\mathbf{S} = \begin{bmatrix} \bar{\alpha} & \alpha & & & \\ & \bar{\alpha} & \alpha & & \\ & & \ddots & \ddots & \\ & & & \bar{\alpha} & \end{bmatrix}. \tag{2.6}$$

- **Mixed geometric ET traffic:** $\rho = [\rho_1\ \rho_2\ \cdots\ \rho_m]$ and

$$\mathbf{S} = \begin{bmatrix} \bar{\alpha}_1 & & & \\ & \bar{\alpha}_2 & & \\ & & \ddots & \\ & & & \bar{\alpha}_m \end{bmatrix}, \tag{2.7}$$

  where α_i and $\rho_i \in [0,1]$; $\forall i = \{1,2,\cdots,m\}$.

- **Bernoulli ET traffic:** $\rho = 1$ and $\mathbf{S} = \bar{\alpha}$.

Regarding the service process of the DTMC, we consider that the packet departures from the queues are based on a signal to interference noise ratio (SINR) capture model. That is, a packet departs from the queue if the achieved SINR exceeds a certain threshold. Considering the network-wide interaction between the devices in uplink communications, first come first serve based interactive queues have been well investigated in the literature for the collision model [32, 33], which ignore the mutual interference between the devices. Collision models assume everything to be static and deterministic, and hence, the collision event cannot be resolved. However, things are more challenging and involving in the SINR capture model. Accordingly, we adopt wireless-based queue abstraction model, in which every wireless link between two nodes can be abstracted with a queue with a given departure probability. In each of the technical chapters, we will build upon the presented model according to the system model at hand. In addition, it will be clearly explained how the SINR model implicitly considers the network-wide aggregate interference along with the devices temporal dynamics which are governed by their traffic. Apart from the arrival and service processes, single-host model with finite and infinite queue sizes are adopted in this thesis. In addition, a first come first serve (FCFS) queueing discipline is adopted, where packets/ users within a given queue are addressed based on their relative temporal arrival time-stamp. In

what follows we will briefly summarize time Markov chains, which are powerful tools to characterize the temporal dynamics of a queue.

Time Markov Chains

Discrete and continuous Markov chains are utilized throughout this thesis to track the temporal dynamics of different system parameters (e.g., packets arrivals, number of users, task offloading instructions, etc.). Without loss of generality, we will briefly present some of the key concepts of the DTMC, whereas for continuous time Markov Chain (CTMC), similar definitions can be deduced. Let $X_0, X_1, \cdots, X_n; n \in \mathbb{N}$ be a discrete time stochastic process with countable state space $\mathcal{D} = \{i_1, i_2, \cdots, i_n\}$. If $\mathbb{P}\{X_{j+1} = i_{j+1} | X_j = i_j, X_{j-1} = i_{j-1}, \cdots, X_0 = i_0\} = \mathbb{P}\{X_{j+1} = i_{j+1} | X_j = i_j\}$ holds for any j and $\mathcal{D}$, then X_j is said to be a DTMC. Furthermore, if $\mathbb{P}\{X_{j+m+1} = i | X_{j+m} = v\} = \mathbb{P}\{X_{j+1} = i | X_j = v\}$, $\forall (i,v) \in \mathcal{D}$, $\forall (j,m) \geq 0$, then the DTMC is said to be time-homogeneous or stationary [29]. Let $p_{v,i}(j) = \mathbb{P}\{X_{j+1} = i | X_j = v\}$ denotes the transition probability from state v at time j to state i at time $j+1$.[3] All the probability transitions between the different states in $\mathcal{D}$ are collected in the probability transition matrix $\mathbf{P} \in \mathbb{R}^{|\mathcal{D}| \times |\mathcal{D}|}$, in which $\sum_{v \in \mathcal{D}} p_{v,i} = 1$, which implies that each row of $\mathbf{P}$ sums to one. One of the most encountered DTMC in discrete time queues is the quasi-birth-death (QBD) [29]. For the queues with infinite queue size, the transition matrix $\mathbf{P}$ of an QBD is given in the block partitioned form as follows

$$\mathbf{P} = \begin{bmatrix} \mathbf{B} & \mathbf{C} & & & \\ \mathbf{E} & \mathbf{A}_1 & \mathbf{A}_0 & & \\ & \mathbf{A}_2 & \mathbf{A}_1 & \mathbf{A}_0 & \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{2.8}$$

An important feature of the QBD is that it only transitions a maximum of one level up or down. In addition, it is skip-free to the left and to the right. Assuming that the matrices $\mathbf{A}_k$, $k = 0,1,2 \in \mathbb{R}^{n \times n}$ and matrix $\mathbf{B} \in \mathbb{R}^{m \times m}$, then, $\mathbf{C} \in \mathbb{R}^{m \times n}$ and $\mathbf{E} \in \mathbb{R}^{n \times m}$. The matrix analytic method (MAM) is a powerful mathematical tool, utilized to analyze DTMCs [31, 29]. The key matrices that form the ingredients of the MAM are the $\mathbf{R}$ and $\mathbf{G}$ matrices, which are the minimal non-negative solutions to the following equations [31]

$$\mathbf{R} = \mathbf{A}_0 + \mathbf{R}\mathbf{A}_1 + \mathbf{R}^2\mathbf{A}_2, \tag{2.9}$$

$$\mathbf{G} = \mathbf{A}_2 + \mathbf{A}_1\mathbf{G} + \mathbf{A}_0\mathbf{G}^2. \tag{2.10}$$

The computations of $\mathbf{R}$, $\mathbf{G}$ as well as the steady state probability of the DTMC is dependent on the block matrices of $\mathbf{P}$ as will be explained in more details in the next chapters. Finally, we would like to emphasize that throughout this thesis, different queueing models will be introduced and studied, we postpone the treatment of each specific model to its respective chapters in order to provide a self-contained analysis within each technical chapter.

[3] Throughout this thesis, we consider cases where the transition between states is independent of the time, i.e. $p_{v,i}(j) = p_{v,i}$, $\forall j$.

2.1.3 Spatiotemporal Modeling

As mentioned in the previous subsection, there has been considerable effort toward characterizing and understanding the performance of wireless links in large-scale networks by using tools from stochastic geometry [24, 23, 25]. Typical system-level expressions for a variety of network statistics, e.g., coverage, throughput, or delay can be captured, by capturing the spatial and physical layer attributes [26]. This intrinsic elegance has marketed stochastic geometry as a disruptive tool for performance evaluation among various wireless systems. However, an underlying limiting aspect of the stochastic geometry based models is the full buffer assumption, which assumes that the transmitter has always backlogged packets to be transmitted. Thus, conventional models based on stochastic geometry are oblivious to the temporal traffic evolution and the underlying queueing dynamics at each device. Moreover, tools from queueing theory are suited for analyzing the per-node temporal dynamics, which provide little insight on the network-wide interactions within the network. To account for the temporal domain, recent efforts have integrated queueing theory with stochastic geometry, offering a full spatiotemporal characterization of the large-scale networks [34–42]. This *spatiotemporal* network perspective triggered a plethora of challenging research directions that addresses the modeling, analyzing and optimizing the network from spatial and temporal perspectives.

Throughout this thesis, the different proposed frameworks entail the macroscopic and microscopic scales of large scale wireless-networks as highlighted in Figure 2.2. The *microscopic scale* is per device level that addresses the temporal dynamics at each device, whereas the *macroscopic scale* represents a holistic view of the network that captures the mutual interaction (i.e., mutual interference and contention among the resources) among the devices (i.e, queues). Hence, the microscopic and macroscopic scales can be regarded as, respectively, a zoom in that shows the behavior of each device and a zoom out that shows the behavior of the entire network. This integrated analysis of the network can be considered as extension to the well established interactive queues problem in the literature [32, 33], in which the collision model, which ignores the mutual interference between the devices, has been widely adopted. However, adopting such an integrated view enables a more holistic understanding of the network's behavior as well as the effect of the different system parameters. For the sake of organized presentation, the detailed discussion of the related spatiotemporal underlying models is left to the technical chapters. That is, for each chapter, the specific system model, related state of the art works and key novelty are discussed there.

2.2 Performance metrics

Through this subsection we will review the main KPIs that are adopted throughout this thesis. The choice of of a specific KPI in the technical chapters is motivated by its relevance to the investigated use case and its effect on different system design insights.

2.2.1 Link Quality

The quality of a given link can be assessed via its achieved SINR (or signal to interference ratio (SIR) in case of interference-limited networks). The SINR distribution characterizes also the aggregate

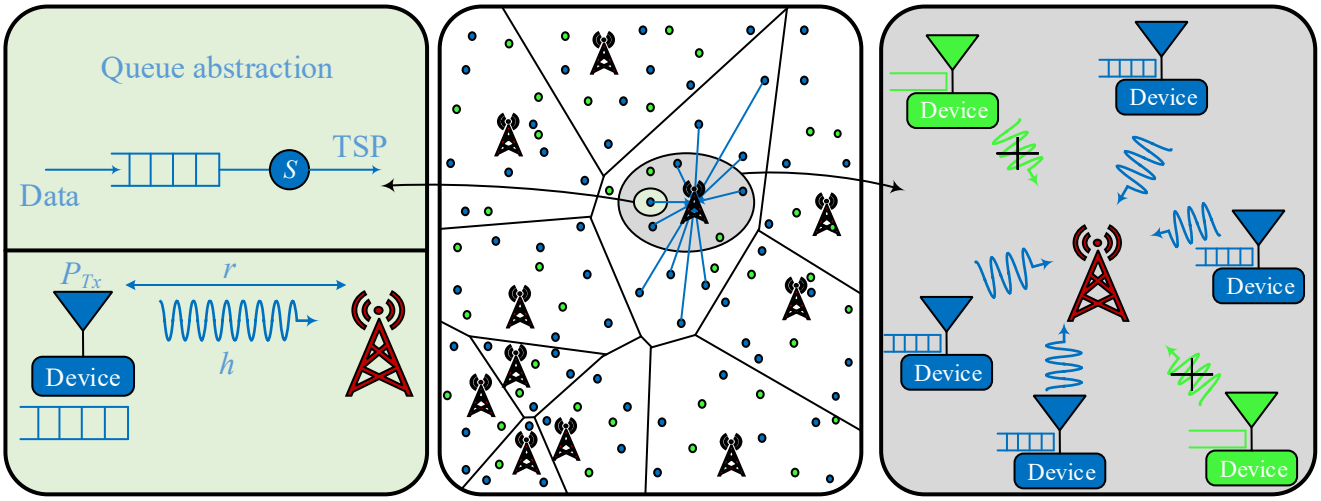


Figure 2.2: Spatiotemporal model with microscopic (macroscopic) network scale highlighted on the left (right). Blue (green) represent active (idle) devices.

interference within the network. Mathematically, the SINR at a generic i-th receiver communicating with a generic j-th transmitter can be characterized as

$$\mathsf{SINR}_{i,j} = \frac{P_j h_{i,j} r_{i,j}^{\eta}}{\sum_{k \in \mathcal{Z}_j} a_k P_k h_{i,k} r_{i,k}^{\eta} + \sigma^2}, \tag{2.11}$$

where P_j is the transmission power of the j-th transmitter, $h_{i,j}$ and ${r_{i,j}}^{\eta}$ are the channel gain, that are Rayleigh distributed with unity gain and the Euclidean distance between the i-th receiver and its j-th transmitter, with the path-loss exponent $\eta > 2$, $\mathcal{Z}_j$ is the set of transmitters utilizing the same channel as the j-th transmitter, a_k equals one if the k-th transmitter is active, and zero otherwise, and σ^2 is the noise power. For the sake of completiness, (2.11) represents the SINR of a generic link direction, nevertheless, in this thesis we consider mainly uplink transmissions. As a result, P_j represent the transmission power of the j-th device to its serving i-th BS. Morevoer, due to the randomness encompassed in (2.11), resulting from the spatial locations, fading, activity profiles, etc., interest is focused on the transmission success probability, denoted by P_{s}, which represents the complementary cumulative distribution function (CCDF) of the SINR. The transmission success probability entails spatial and temporal averaging and is mathematically expressed as

$$P_{\mathrm{s}} = \mathbb{P}\{\mathsf{SINR} > \theta\}, \tag{2.12}$$

where θ is the decoding threshold. The expression in (2.12) can be thought of equivalently as:

- the probability that a randomly chosen device can achieve $\mathsf{SINR} > \theta$,
- the average fraction of devices within the network who at any time achieve $\mathsf{SINR} > \theta$,
- the average network's area fraction that is in coverage at any time.

Moreover, a packet transmitted in a given time slot is considered successfully decoded, if the accompanying instantaneous SINR is greater than θ. For the case of unsuccessful decoding, a negative

acknowledgment is sent via a dedicated channel and the transmitter attempts a re-transmission in the following time slot. We would like to emphasize that throughout this thesis, the latency incurred by acknowledgment messages is not considered, as it is negligible compared to data transmission latency. Additionally, a quasi-deterministic channel model is adopted, such that the experienced channel gain (i.e., $h_{i,j}$) by the i-th transmistter is constant over the packet transmission period (i.e., slot). On the other hand, in case of retransmissions, a new randomized channel gain is observed by that transmitter.

2.2.2 Pareto Frontiers

The temporal fluctuation of traffic introduces a significant source of randomness to the network, which is captured, tracked, and analyzed, as mentioned earlier, using tools from queueing theory [28]. This randomness raises important questions that cannot be answered via the full-buffer analysis. Some of these questions are i) given that packets arrive following a given traffic model at a device, how long on average a packet takes to be successfully transmitted; ii) how such random packet arrival in different devices affect their interactions in terms of mutual interference; iii) will the devices be able to deliver all of the generated packets or will they run into instability, and iv) what are the network parameters that guarantee that the devices will be able to deliver all generated packets.

As mentioned earlier, prior stand alone stochastic geometry or stand alone queueing theory models cannot address these questions. This is attributed to the fact that devices are randomly deployed, with deterministic or stochastic traffic, and random interaction in both space and time. Hence, a combined stochastic geometry and queueing theory model, denoted as spatiotemporal model, should be used to model such network and answer the aforementioned questions. Considering the spatiotemporal network's performance, the Pareto frontiers define regions where the queues employed at the transmitters are guaranteed to be operating within a stable region. Thus, operating beyond the Pareto frontiers, will yield the network (of queues) unstable. Throughout the Chapters 3 and 4, we study, characterize and analyze the Pareto frontiers for large scale uplink IoT networks.

2.2.3 Latency

Achieving low latency is an important target for many of the computation and communication applications [43]. Moreover, MEC promised latency reductions are facilitated by the computation of such intensive tasks at the MEC host. As a result, assessing the different latency components of the end-to-end experienced latency, can help identify latency bottlenecks and provide insights on how to alleviate them if possible. Throughout this thesis, the following latency components will be investigated:

- Radio latency: represents the latency incurred to transmit a packet in either downlink or uplink. Parameters such as mutual interference, transmission power and radio resources, among others, affect this metric.

- Queueing latency: results from the waiting time of a packet within its queue till it is successfully received at its destination. Latencies resulting from packet re-transmissions are implicitly included in this metric.

- Network latency: results from the different network components such as back-haul, transport and core network. This metric will be utilized to analyze the MEC latency reduction gains.
- Execution latency: considers the time required to process a given task either locally or at the MEC host. Parameters such as task size, processing power, and number of required cycles to process an input bit will be considered in subsequent chapters.

2.2.4 Information Freshness

Different from the experienced latency and inter-delivery time, the timeliness and retainability of continuous updates of nodes within a system are overarching requirements to meet the different data transfer, monitoring, timing, and scaling challenges [44, 45]. This implies continuous information update about the real-time states between a given source and its targeted destination. As presented in Chapter 1, this is essential for IoT and its underlying architecture, that include among others, ubiquitous sensors and autonomous actuators [46]. Regarding the computational aspect, timely execution of tasks is essential to ensure timely result delivery and an enhanced quality of experience [47].

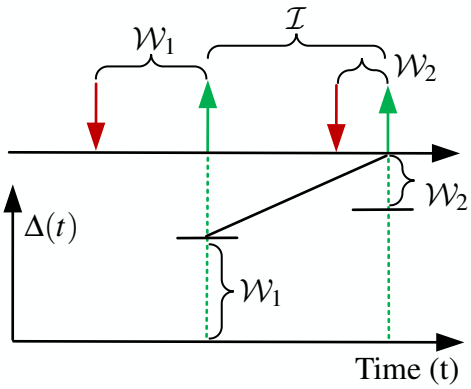
To characterize the freshness of information at the receiver, we adopt the age of information (AoI) metric, that was first introduced in [48]. The AoI accumulates the transmission delay in addition to the time elapsed between successive system updates [49]. To account for the information freshness, the AoI increases linearly when there are no packets in the system as shown for a single source-destination pair in Figure 2.3. Compared to traditional time metrics (e.g. packet waiting time which is denoted by w_1 and w_2 for two consecutive packets in the aforementioned figure), AoI captures the timeliness of updates in a way those traditional metrics do not [50, 51]. Assume that the i-th packet is generated at time G_i, then $\Delta_i(t+1)$ is computed recursively as

$$\Delta_i(t+1) = \begin{cases} \Delta_i(t)+1, & \text{transmission failure,} \\ t - G_i + 1, & \text{otherwise} \end{cases} \tag{2.13}$$

For an M/M/1 FCFS queuing model, in which update packets are generated at the source following a Poisson process with average rate λ_a and service times are independent and identically distributed exponentials with average service rate μ_{ser}, the average AoI can be evaluated as reported in [48] as

$$\Delta = \frac{1}{\mu_{\text{ser}}}\left(1 + \frac{\lambda_a}{\mu_{\text{ser}}} + \frac{\mu_{\text{ser}}^2}{\lambda_a^2(1-\frac{\mu_{\text{ser}}}{\lambda_a})}\right). \tag{2.14}$$

In Figure 2.3, we compute the average waiting time $\mathbb{E}\{\mathcal{W}\}$, average inter-arrival time $\mathbb{E}\{\mathcal{I}\}$ and average AoI $\mathbb{E}\{\Delta\}$ for $\mu_{\text{ser}} = 1$ and three different values of λ_a. It is observed that for relative extreme values of λ_a (i.e., 0,01 and 0.99), information is outdated (i.e., high AoI values). In such cases, focusing on the delay solely is not sufficient, as it neglects that impact of the inter-delivery time. It can also be observed from the reported values in Figure 2.3 that a desirable low AoI is realized when packets with low waiting times are delivered regularly. Thus, AoI adds a new perspective to the system view that is overcoming the shortcomings of the typical delay metrics. Moreover, through this thesis, we consider the PAoI, which is an alternative and more tractable metric compared to the AoI. It is defined as the value of age resulted



λ_a	$\mathbb{E}\{\mathcal{W}\}$	$\mathbb{E}\{\mathcal{I}\}$	$\mathbb{E}\{\Delta(t)\}$
0.01	1.01	100	101
0.52	2.08	1.92	3.48
0.99	100	1.01	100.02

Figure 2.3: Age of information illustration and M/M/1 example.

immediately prior to receiving the i-th update [52]. Mathematically, the time averaged PAoI is computed as

$$\Delta_{\mathrm{p}} = \mathbb{E}\{\mathcal{W}\} + \mathbb{E}\{\mathcal{I}\}. \tag{2.15}$$

The increased focus on the PAoI stems from the guaranteed system performance insights it unveils. In addition, the minimization of the PAoI may be required for time critical applications [53]. The PAoI will be investigated for event-triggered traffic, prioritized traffic and vehicular use cases in the coming chapters.

2.2.5 Dependability Attributes

A fundamental question for future computation and communication systems is how to characterize and quantify the ability to operate fault-free. In both recent 3GPP specifications and academic research works, URLLC use cases have been studied with similar questions in mind. Generally, metrics such as packet error ratio, latency and jitter, have been well understood in order to assess a given system performance. These metrics, though fundamentally meaningful from the radio communication perspective, need to be looked collectively with the computational service demands. Consequently, applying dependability parameters describe not only the proportion of fault-free functioning of the communication system, but also its readiness to function and the ability to preserve and restore the desired function. From a dependability point of view, temporal and spatial availability of a service and reliability of its operation, help us understand a new perspective of the system's functionality. A glimpse of the different dependability attributes is illustrated in Figure 2.4, where adopting such measures is inevitable for 5G and beyond systems [54]. Understanding and optimizing such attributes facilitate addressing the plethora of challenges and their stringent requirements, which is needed to ensure service operation with virtually no failures during the operation time [55]. The need to bridge the gap between traditional radio-link and service-level key performance indicators is imminent, via providing an insight on the system components from a dependable perspective.

To this end, tools from reliability theory will be utilized throughout the thesis to assess the computation and communication paradigm via novel defined KPIs. Reliability theory involves the development of mathematical methods in order to evaluate the reliability, maintainability, availability, and safety of

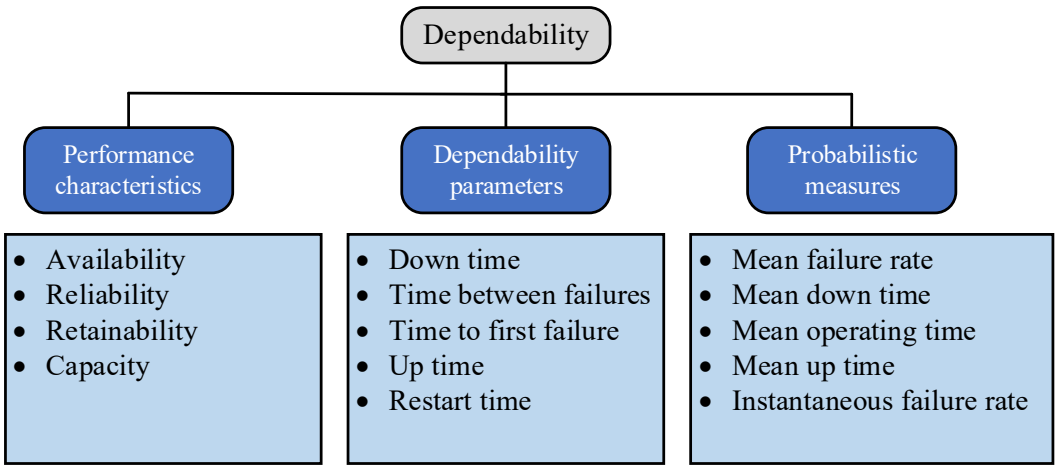


Figure 2.4: Assessment of different dependability attributes for dependable systems.

technical components, equipment, and systems [56]. We utilize such concepts to study efficient service availability, its reliability and computational-execution related KPIs.

2.3 Multi-access Edge Computing

To realize efficient computing, caching, and data analytic resources at the network edge, MEC is introduced by the European telecommunication standards institute (ETSI) industry specification group as a mean of extending intelligence to the edge of the network along with higher processing and storage capabilities [10]. MEC enables the implementation of MEC applications as software-only entities that run on top of a virtualization infrastructure, which is located close to the network edge [57]. MEC deployment will introduce computing capabilities at the edge of the network and will provide an open environment targeting low packet delays due to close proximity to end users [58]. As a result, this will minimize network congestion and improve resource optimization, user experience, and the overall network performance.

From a standardization point of view, 3GPP indicates the interoperability of MEC deployment in the 5G network as presented in [12]. Seamless integration of MEC into 5G is illustrated in Fig 2.5. The presented architecture comprises two parts: the 5G service-based architecture on the left and an MEC reference architecture on the right. The network functions defined in the 5G architecture and their roles are briefly summarized as follows [59, 10]:

- Access and mobility management function (AMF): handles mobility and access procedures (e.g., connection and mobility management, termination of the RAN control plane, integrity protection and access authentication/ authorization).
- Session management function (SMF): performs session management-related functionalities, such as , session establishment, charging and support for roaming, and downlink data notification.
- Network slice selection function (NSSF): assists in the selection of suitable network slice instances for users and the allocation of necessary AMFs.

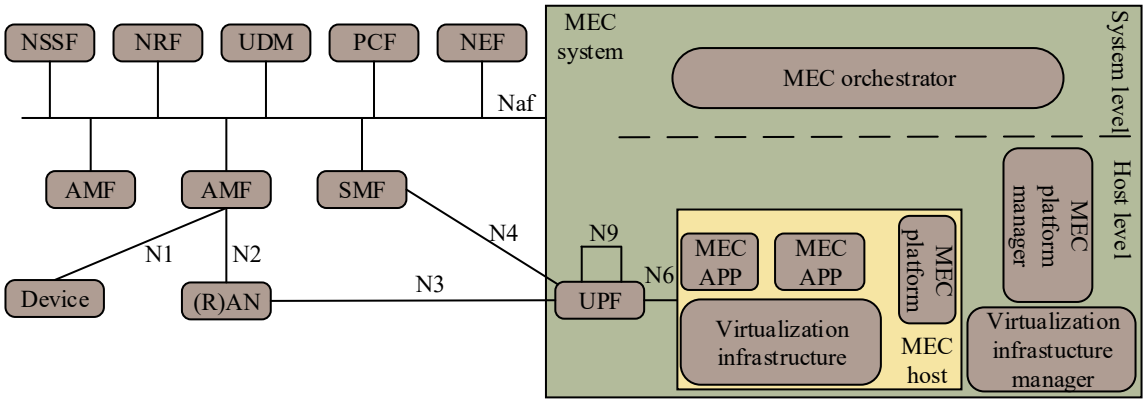


Figure 2.5: Integration of 5G service-based architecture and a generic MEC system [59].

- Network repository function (NRF): entails the discovery of network functions and their supported services.
- Unified data management (UDM): handles user subscription and identification services.
- Policy control function (PCF): unifies the network policies and provides policy rules to control plane functions (e.g., traffic steering).
- Network exposure function (NEF): acts as a service-aware border gateway for providing secure communication with the services supported by the network functions.
- Authentication server function (ASF): performs authentication procedures.
- User plane function (UPF): provides functionalities to facilitate user plane operations, e.g., packet routing and forwarding, data buffering, and allocation of IP address.

Although the aforementioned integrated architecture clarifies how the logical system entities can interconnect and interoperate, it does not specify where the edge cloud is physically located. Deployment options can vary from central to remote [60]. In practice, the logical architecture defined in the standards does not aim to answer the frequently asked question: *where is the edge?* So far, only qualitative studies are present in the literature, providing generic guidelines for decision makers [10, 58]. In fact, when it comes to physical deployment options, the MEC stakeholders (i.e., network operators, cloud providers, and infrastructure owners) need to clarify how this logical mapping is translated into a practical deployment blueprint. However, quantitative analysis that demonstrates the performance gains resulting from MEC system deployment is still to be conceived in different use cases. As a first attempt to provide meaningful assessment to MEC gains, we consider throughout this thesis that, a MEC host, that addresses the different services within the network, is physically co-located with each BS. Before delving into the considered technical use cases, it is first worth highlighting the inter-connection between MEC and IoT network deployments.

Enabled by the IoT paradigm, MEC has opened many new frontiers for network operators, services and content providers to deploy versatile, heterogeneous and continuous services on IoT devices [3]. In this sense, MEC and IoT are viewed as complementary technologies, in which MEC empowers

computational-limited IoT devices with significant additional computational capabilities, to execute computationally demanding tasks, via task offloading [7]. On the other hand, IoT provides MEC with a plethora of devices that can utilize its promised gains, ranging from sensors and actuators to smart vehicles and industrial automation [61]. As presented in [3], the IoT-MEC collaboration offers i) less traffic passing through the network's infrastructure, ii) latency reduction for applications and services, and iii) scalability gains in terms of offered services. Among such gains, latency reduction gains, introduced by MEC due the reduced physical and virtual communication distance, stands out as a key enabler for many new market segments [61].

Chapter 3

Prioritized Multi-stream Traffic: Spatially Interacting Vacation Queues

This chapter develops a novel priority-aware spatiotemporal mathematical model to characterize massive IoT networks with uplink prioritized multi-stream traffic. In such networks, heterogeneous traffic is envisaged, where packets generated at each device should be differentiated and served according to their priority. Stochastic geometry is utilized to account for the macroscopic network wide mutual interference between the coexisting devices. DTMCs are employed to track the microscopic evolution of packets within each priority queue. To provide a systematic and tractable model, we decompose the prioritized queueing model at each device to a single-queue system with server vacation. To this end, the IoT network with prioritized multi-stream traffic is modeled as spatially interacting vacation queues. Dedicated and shared channel priority-aware access strategies are presented. A priority-agnostic scheme is used as a benchmark to highlight the impact of prioritized uplink transmission on the performance of different priorities in terms of transmission probabilities and delay. Additional performance metrics as average number of packets, PAoI, delay distribution, and Pareto frontiers for different parameters are presented, which give insights on stable operation of uplink IoT networks with prioritized multi-stream traffic.

This chapter is organized as follows. The background, literature review and our contributions are highlighted in Section 3.1. Section 3.2 provides the system model and the underlying physical and MAC assumptions. The proposed queueing model along with the microscopic intra-device interactions among the priority queues are presented in Section 3.3. Section 3.4 shows the macroscopic inter-device queueing interactions in terms of mutual interference. Simulation results are presented in Section 3.5. Finally, Section 3.6 summarizes the work and draws some conclusions.

3.1 Introduction

Traffic prioritization schemes in IoT is inevitable due to the IoT heterogeneous traffic such as regular traffic (e.g., reports or updates), query responses (e.g., diagnostics), special measurements, control packets, warnings, and alarms [62]. In addition, system alarms or failures need to be addressed almost immediately. Thus, heterogeneous multi-stream traffic is envisaged, where each traffic stream needs to be

differentiated and addressed according to its priority. Such traffic discrepancies impose new challenges on how to properly model the network. The necessity to meet the targeted QoS becomes more prominent with prioritized multi-stream traffic in mixed-criticality systems. For cellular systems, the concept of QoS class identifier (QCI) was first adopted in long term evolution (LTE) systems to characterize different services and to ensure that resources are allocated appropriately [63]. Each stream (i.e., data bearer) has a corresponding QCI, which indicates the service type, priority, and packet transmission requirements. Industrial automation is another sector that relies on prioritized traffic, where guaranteed performance regarding successful packet delivery and latency is an imminent KPI [64]. In particular, the IEEE 802.1 Qbv amendment, among its many features, introduces eight different priority classes that are assigned to an incoming traffic stream which define the service requirements of each stream [65].

In addition to traffic prioritization within the network, massive number of deployed IoT devices is foreseen as highlighted in the previous chapters [6]. Due to the shared characteristic of the wireless channel, mutual interference between the IoT devices is imminent. In this context, a key enabler of large scale IoT devices is the low cost of deployment, which is realized via distributed and uncoordinated devices. Due to its decentralized nature, grant-free access is adopted in uplink cellular transmissions, where the scheduling complexities imposed by the scheduling grants from the BSs are alleviated [66]. To this end, proper understanding and modeling of the prioritized traffic within the massive number of devices is required to i) characterize the performance; ii) understand the impact of different network parameters; iii) highlight common trends in the network's performance, and iv) provide design insights.

Queues with prioritized traffic have attracted wide attention in the queueing theory literature where different metrics (e.g., waiting time distribution and average queue length) are characterized [29, 67]. The incorporation of vacations to facilitate the analysis of priority queues is proposed in [68, 67, 69–72]. Nonetheless, the previously mentioned works consider only the interactions within a single queue and disregard the network-wide interaction between the devices prioritized multi-stream traffic. Focusing on the work addressing the spatiotemporal view, the work in [34] characterizes the delay outage and downlink SIR for a heterogeneous cellular network under random, FCFS and round-robin scheduling schemes. The authors in [35] present a spatiotemporal characterization for grant-free uplink transmissions in IoT network, where the performance of power-ramping and back-off transmission strategies are investigated. The work in [35] is extended and compared to scheduled (i.e., grant-based) uplink transmissions in [36] and it is shown that the network performance is highly dependent on the devices densities and traffic load. Analysis for small cell deployment is presented in [38], where the authors show the traffic load effect on the transmission success probability. For an ad-hoc network, [39] presents a fine-grained spatiotemporal characterization for location-dependent QoS classes in IoT networks.

Considering prioritized traffic under a spatiotemporal perspective, [40] studies the delay and throughput in a cognitive radio setup, in which a network of secondary users share the channel with a single primary user. Secondary users are allowed to access the channel with a probability that depends on the primary user's queue length. However, their proposed framework only considers two priority classes. Recently, a framework to characterize an N-class prioritized devices is proposed in [41], where users randomly share the available channel. However, the model in [41] is for prioritized devices, not traffic streams, and is only applicable to ad-hoc networks. In summary, none of the aforementioned works consider prioritized multi-stream traffic in uplink IoT networks. In addition, we are not aware of any

work in the literature that characterizes the spatiotemporal performance, stability frontiers, and delay under different channel allocation strategies.

When compared to the results presented in the aforementioned works, we provide an analytical framework that entails spatial macroscopic and microscopic scales of uplink large scale IoT networks with prioritized traffic. The analysis relies on the joint utilization of stochastic geometry and queueing theory. The spatial macroscopic scale denotes the network-wide interactions arising between the devices in terms of the packet departure probabilities, due to mutual interference between the simultaneously active devices. Tools from stochastic geometry are employed to characterize the network-wide aggregate interference. On the other hand, the spatial microscopic scale, investigated via tools from queueing theory, represents the priority queues temporal dynamics and their interactions. To track the priority class being served at a given time stamp, a two-dimensional geometric (Geo)/PH/1 DTMC is employed for each device. In summary, the main contributions of this chapter are summarized as:

- Develop a novel and tractable spatiotemporal framework, based on stochastic geometry and queueing theory, that jointly accounts for prioritized multi-stream traffic traffic in uplink large scale IoT networks;
- employ a two dimensional Geo/PH/1 DTMC at every IoT device to account for the temporal evolution of queues in response to the prioritized multi-stream traffic arrivals and departures;
- integrate the developed DTMCs within a stochastic geometry framework to account for interference-based intrinsic inter-dependency between the macroscopic- and microscopic-scales;
- compare the dedicated and shared allocation strategies with respect to various KPIs;
- present the Pareto frontiers that characterize the stability regions for different parameters.

3.2 System Model

3.2.1 Spatial & Physical Layer Parameters

This chapter studies a cellular uplink network, where the BSs and IoT devices are spatially deployed in $\mathbb{R}^2$ according to two independent homogeneous PPPs, denoted by Φ and Ψ with intensities λ and μ, respectively. Single antennas are employed at all devices and BSs. Grant-free access is assumed, where the devices attempt their transmissions on a randomly selected channel without a scheduling grant from their serving BS. In addition, single connectivity is considered where each device is served by its nearest BS. To alleviate congestion, a set of C channels are utilized by the network and a priority-aware access strategy is adopted by the devices to access the available channels. This corresponds to the Zadoff-Chu codes utilized in LTE and 5G system for the random access channels to request scheduling grants[66].[1] In this chapter, we analyze three channel allocation strategies for priority-aware packet transmission, namely, i) dedicated strategy for each priority class with equal channel allocation; ii) dedicated strategy

[1] For mathematical tractability, we consider only orthogonal channels (i.e., Zadoff-Chu codes stemming from the same root).

for each priority class with weighted channel allocation, and iii) shared strategy for all priority classes. For the dedicated strategy, each priority stream has an exclusive set of channels that can only be accessed by the devices to transmit their corresponding priority packets. For the shared strategy, all the channels can be accessed by all devices irrespective of the transmitted packet's priority.

An unbounded path-loss propagation model is adopted such that the signal power attenuates at the rate $r^{-\eta}$, where r is the distance and $\eta > 2$ is the path-loss exponent. Small-scale fading is assumed to be multi-path Rayleigh fading, where the signal of interest and interference channel power gains h and g, respectively, are exponentially distributed with unit power gain. All channel gains are assumed to be spatially and temporally independent and identical distributed. Full path-loss channel-inversion power control is adopted, which implies that all devices adjust their transmit powers such that the received uplink average powers at their serving BS is equal to a predetermined value ρ [73]. Moreover, a dense deployment of BSs is assumed, ensuring that every device is able to invert its path-loss almost surely. A packet generated at a given device is successfully decoded at its serving BS if the received SINR is larger than a predefined threshold θ.[2] Let d_i and $P_{\mathrm{s},i}$ denote the departure probability and the transmission success probability of an i-th priority packet, given a transmission attempt, respectively. Mathematically, both metrics can be evaluated as follows

$$d_i = \mathcal{T}_i P_{\mathrm{s},i}, \tag{3.1}$$

$$P_{\mathrm{s},i} = \mathbb{P}\{\mathsf{SINR}_i > \theta\}, \tag{3.2}$$

where SINR_i and $\mathcal{T}_i$ are the SINR and the transmission access probability of the i-th priority queue, respectively. It is worth noting that $P_{\mathrm{s},i}$ incorporates the inter-dependency between the macroscopic and microscopic scales of the network.

3.2.2 Temporal & MAC Layer Parameters

The proposed framework considers a synchronized, time slotted, and priority-aware system, in which packets of different priorities are generated at the devices. A prioritized multi-stream traffic model is considered such that packets are generated at each priority class independently of other classes. Hence, for a system with N priority classes, batch arrivals up to N packets can occur in every time slot. Independent geometric inter-arrival times are assumed between packets belonging to each priority class with parameters $\alpha_i \in [0,1]$, $i \in \{1,2,\cdots,N\}$. Through this chapter, traffic parameterized with lower indices has higher priority. Generally, we consider that each device has N-priority finite queues, each of size q_i, that accumulate generated packets according to their priorities. The devices employ a priority-aware transmission strategy that prioritizes the transmission of high priority over lower priority packets. Furthermore, spatially-uniform distributed traffic is considered, whereas the case of location-dependent traffic can be extended by adopting different point processes (e.g., Poisson cluster point process) [22].

Furthermore, it is assumed that arrival and departure of packets only occurs at the start of a time slot. If a high priority packet arrives at its respective queue while a lower priority queue is being addressed,

[2]Throughout this thesis, queueing activities at the BS are not considered and left for future works. Nevertheless, analysis developed in this chapter for the prioritized multi-stream traffic can be extended to study networks with queues at both the IoT devices and their serving BSs.

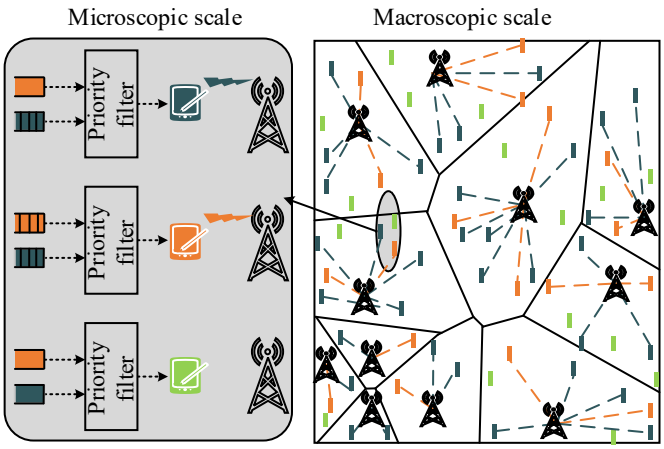


Figure 3.1: A snapshot realization of the network with two priority classes. Dark teal, orange and green rectangles represent devices with high priority packet, low priority packet and no packets in their queues, respectively. The Voronoi cells of the BSs are denoted by the solid black lines while the dashed lines denote the active transmissions between devices and their serving BSs.

service is interrupted and switched to the higher priority queue. The interrupted service is resumed after the high priority queue is empty. Thus, an inter-class preemptive discipline is considered along with an FCFS discipline within each priority queue. In addition, BSs have no knowledge regarding the status of the devices queues. For the dedicated channel allocation strategies, the device randomly and uniformly selects one of the channels dedicated for the addressed packet priority. For the shared strategy, the device randomly and uniformly selects one of the complete set of channels regardless of the packet priority. In both cases the channel selection process is repeated in each transmission attempt.

Pictorially, a snapshot realization of the network for two priority classes is shown in Figure 4.1. The right-hand side of the figure highlights a macroscopic network view and the left-hand side emphasizes the microscopic scale of three links. Due to the adopted preemptive priority discipline, imposed by the priority filter block, packets existing at high priority queues are prioritized for service (i.e., transmission) over packets existing in lower priority queues. If no high priority packets exist, the backlogged lower priority packets are served. In the case of having empty queues, no transmission is attempted and the device does not contribute to the network interference. It is worth noting that the time scale of channel fading, packet generation and transmission is much smaller than that of the spatial dynamics.Each spatial network realization for the adopted PPPs remains static over sufficiently large number of time slots, while channel fading, queue states, and device activities change from one time slot to another.

3.3 Temporal Microscopic Analysis

Throughout this section, a novel technique to model the prioritized multi-stream traffic is presented. In order to mathematically describe the different priority queues, a conventional way of characterizing the system is based on the following state space [29, Chapter 9]. Let $\mathcal{Y} = \{(z_{1,n}, z_{2,n}, \cdots, z_{N,n}) | z_{i,n} \in \{0, 1, \cdots, q_i\}\}$ and $i \in \{1, 2, \cdots, N\}, n = 1, 2, \cdots$, where $z_{i,n}$ denotes the number of i-th priority packets at the n-th time slot. Although tractable for the case of $N = 2$, the depicted state space becomes disproportionately complex for larger values of N [29]. As a result, we seek to introduce a scalable and tractable model for

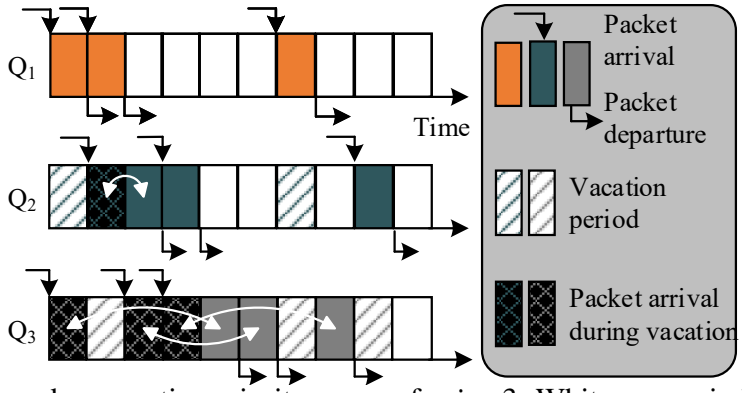


Figure 3.2: Vacation-based preemptive priority queues for $i = 3$. White curves indicates how low priority packets await service till higher priority queues are addressed.

a general number of priority classes based on vacation queues. For ease of mathematical exposition, the time-index n will be dropped hereafter.

Priority queues can be modeled using vacation queues, where low priority queues are forced into a vacation period to allow the high priority queues service [67, 69, 71, 72, 68]. In other words, the prioritized multi-stream traffic is decomposed into a single queue with vacations, where the server becomes alternatively available and unavailable for a given priority class. In our model the server represents the wireless link over which a packet is transmitted. A sever vacation means that the IoT device is utilizing the current uplink time slot to transmit a high priority packet and no lower priority packet can be transmitted within this time slot. The unavailability of the server, denoted as vacation, is due to serving higher priority packets.

An illustrative example for the vacation-based modeling of priority queues is shown in Figure 3.2. Due to its priority, the first priority queue is agnostic to the lower priority queues dynamics. On the other hand, the second priority queue will be in vacation till the first priority queue is empty. Similarly, the third priority queue will be in vacation till the two higher queues are empty. Conceptually, a given queue will go strictly to vacation if a packet resides in any of the higher priority queues. For ease of demonstration, Figure 3.2, assumes a hypothetical flawless server (i.e., $P_{\mathrm{s},i} = 1; \forall i = \{1,2,3\}$), thus, ignoring the events of packet transmission failures due to poor wireless channel conditions or high aggregate network-wide interference from mutually active devices.

In that sense, one can consider that the vacation period of the i-th priority queue is the summation of the busy periods of the higher queues. In this context, the i-th priority queue's vacation period can be modeled via an PH type distribution, which tracks the server's status whether it is serving the intended (i.e., i-th) priority queue or in vacation serving higher priority queues. By virtue of preemptive prioritization, there is no need to track any of the lower priority queues when analyzing the i-th priority class. Accordingly, the state space for the proposed vacation-based model $\mathcal{Y}_{\mathrm{v}} = \big\{\big(\mathcal{S}_i, \mathcal{V}_i\big) | i \in \{1,2,\cdots,N\},\big\}$, where $\mathcal{S}_i \in \{0,1,\cdots,q_i\}$ represents the number of packets at the i-th priority queue and $\mathcal{V}_i = \{(v_1, v_2, \cdots, v_{i-1}) | v_j \in \{0,1,\cdots,k_j\}\ \&\ \exists v_j > 0\}$ captures the vacation states of the server in terms of the number of packets in the higher priority queues. It is worth mentioning that, due to service preemption, any combination of non-empty higher priority queues is considered as a service vacation event for the lower priority queues. Utilizing such categorization of states, Figure 3.3 presents a two-dimensional

Geo/PH/1 Markov chain that is employed at each IoT device to track the packet's temporal evolution. The horizontal transitions represent the states of the server, denoted as phases, whether in vacation serving higher priority packets or serving the intended i-th priority queue. The vertical transitions represent the number of the packets in the i-th priority queue, denoted as levels. By virtue of the vacation-based categorization in $\mathcal{Y}_{\text{v}}$, Figure 3.3 represents the transitions between serving the third priority class ($\mathcal{S}_3$ is captured via the left hand states) and being in vacation serving higher priority classes ($\mathcal{V}_3$ is captured via the right hand state and its internal components).

In details, the PH type distribution of the server's vacation is represented via an absorbing Markov chain. When serving higher priority packets, the server will be looping in the transient states of the PH type distribution. Absorbing Markov chains are mathematically described via an initialization vector and a transient matrix. In our case, the initialization vector and transient matrix are denoted as $\mathbf{v}_i$ and $\mathbf{V}_i$, respectively. The initialization vector $\mathbf{v}_i$ captures all the possible initial states for vacations with their corresponding probabilities. That is, any combination of batch arrivals, that include higher priority packets, represents a legitimate initial state for the server vacation. For $i = 3$, all legitimate initial vacation states are illustrated in Figure 3.3. The sub-stochastic transient matrix $\mathbf{V}_i$ tracks the server's vacation through tracking the temporal evolution of packets in the higher priority queues. Adopting this vacation-based model allows a systematic and tractable approach to model a network with generic N priorities.

For simplicity and ease of understanding lets consider only the first three priority classes, shown in Figure 3.3, where it is represented the possible states (i.e., in terms of number of packets) of the third priority queue on the left hand side (depicted by $\mathcal{S}_3$). On the right hand side, we plot the vacation states (i.e., in terms of all combinations of the number of packets for the higher priority queues). Such vacations are represented by the states $\mathcal{V}_3 = \{(1,0),(0,1),(1,1),(1,2),(2,2),\cdots\}$, since the server has to serve first priority queues' packets and then proceed to the second priority queue, before serving the third priority queue packets. Now consider the server is serving a third priority packet. Such service will be interrupted if a higher priority packet arrives. It is important to note that there exist several higher priority packet arrival events due to the considered multi-stream traffic. That is, the higher priority packet arrival may be for the first queue only, the second second queue only, or both priority queues simultaneously. Such possible states for a start of vacation period are shaded in Figure 3 and labeled as *"Initial vacation possible states"*. Moreover, the phases that capture all the interactions between the first and second priority queues (i.e., busy period of both first and second priority queues) are embedded within the matrix $\mathbf{V}_3$. The vector representing the successful serving of the first and second priority queues is captured via $\tilde{\mathbf{v}}_3$, whereas $\mathbf{v}_3$ represents the vacation initialization vector. The probability of a packet arrival (first priority packet, second priority packet or both) is captured by χ_3. In other words, χ_3 represents the probability that the third priority queue starts a vacation. Now consider the third priority queue's perspective, what is important is just to know the busy period of the aggregate higher priority queues, which is how long the server will be in vacation before it starts serving its packets (i.e., third priority packets). Thus, the right hand side just represents the vacation phases, but upon zooming in, one observes all the possible phases that represent such vacation.

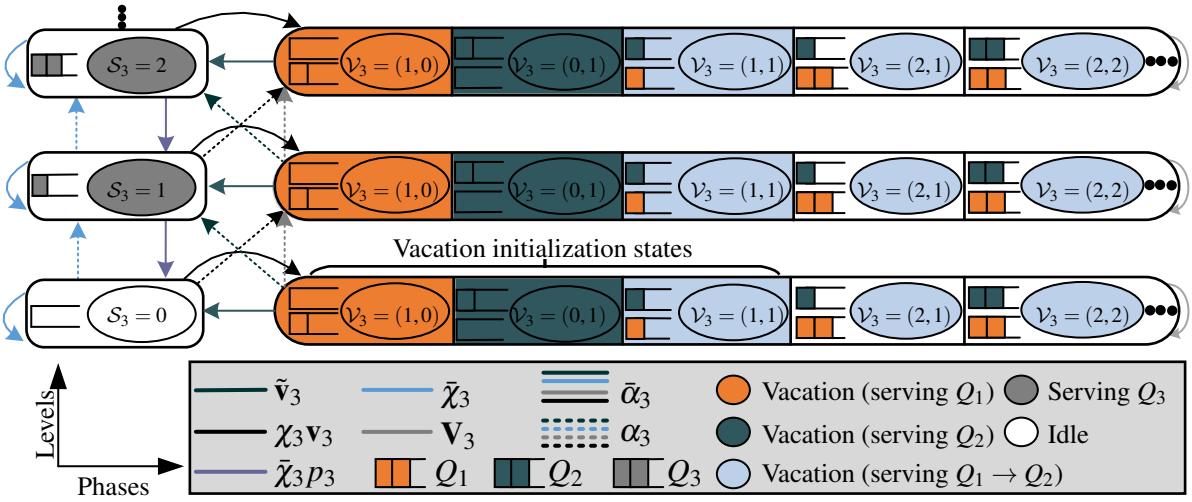


Figure 3.3: Two-dimensional DTMC modeling the vacation-based priority queues for $i = 3$. States for the first, second and third priority classes are depicted by red, green and blue circles, respectively. Solid (dashed) lines are all multiplied by $\bar{\alpha}_3$ (α_3).

3.3.1 Vacation-based Priority Queues Analysis

Let $m_i = \prod_{m=1}^{i-1}(k_m + 1)$ denotes the number of transient states in the PH type distribution of the i-th priority queue. For mathematical convenience, we utilize a two level PH type distribution. In the higher level, absorption denotes packet departure from the i-th priority queue. At the lower level, absorption implies that the server comes back from vacation and is serving the i-th priority packet. Such hierarchy facilitates the construction of the system transition matrix. The utilized higher level PH type distribution is denoted by the initialization vector and transient matrix tuple $(\boldsymbol{\beta}_i, \mathbf{S}_i)$, where $\boldsymbol{\beta}_i \in \mathbb{R}^{1\times m_i}$ and $\mathbf{S}_i \in \mathbb{R}^{m_i \times m_i}$. In details, $\mathbf{S}_i$ is the sub-stochastic transient matrix that incorporates all the transition probabilities (including whether the server is in vacation or not) until packet departure [29]. Starting from any state, the temporal evolution until a single packet departures is captured via the following absorbing Markov chain

$$\mathbf{T}_i = \begin{bmatrix} 1 & 0 \\ \mathbf{s}_i & \mathbf{S}_i \end{bmatrix}, \tag{3.3}$$

where $\mathbf{s}_i \in \mathbb{R}^{m_i \times 1}$ is the probability of being absorbed from a given transient phase and is given by $\mathbf{s}_i = 1_{m_i} - \mathbf{S}_i 1_{m_i}$. It is worth noting that $\mathbf{s}_i$ only have a non-zero element in the location corresponding to the serving state of the server, since a packet only departs while the server is not in a vacation. Exploiting the mentioned PH type distribution, a scalable formulation that captures the queueing dynamics can be given in the form of a QBD process [74]. In particular, the probability transition matrix $\mathbf{P}_i$ of the i-th priority queue is

$$\mathbf{P}_i = \begin{bmatrix} \mathbf{B}_{1,i} & \mathbf{C}_i & & & & \\ \mathbf{A}_{2,i} & \mathbf{A}_{1,i} & \mathbf{A}_{0,i} & & & \\ & \mathbf{A}_{2,i} & \mathbf{A}_{1,i} & \mathbf{A}_{0,i} & & \\ & & \ddots & \ddots & \ddots & \\ & & & & \mathbf{A}_{2,i} & \mathbf{B}_{2,i} \end{bmatrix}, \tag{3.4}$$

where $\mathbf{B}_{1,i}, \mathbf{C}_i$ and $\mathbf{B}_{2,i} \in \mathbb{R}^{m_i \times m_i}$ are the boundary sub-stochastic matrices.[3] In addition, $\mathbf{A}_{0,i}, \mathbf{A}_{1,i}$ and $\mathbf{A}_{2,i} \in \mathbb{R}^{m_i \times m_i}$ represent the sub-stochastic matrices that capture the transition down a level, in the same level, and up a level within the QBD, respectively.

In details, $\mathbf{B}_{1,i} = \bar{\alpha}_i \mathbf{S}_{0,i}$ captures all transitions from and to the idle state, where $\mathbf{S}_{0,i}$ is the stochastic transient boundary matrix. Similarly, $\mathbf{C}_i = \alpha_i \mathbf{S}_{0,i}$ captures the transitions to level 1, that represents an increment of the i-th priority packets. The forward transitions sub-matrix $\mathbf{A}_{0,i} = \alpha_i \mathbf{S}_i$ represents the case where a new packet arrives and no packet departs (i.e., vacation state or serving state with transmission failure). The local transitions sub-matrix $\mathbf{A}_{1,i} = \alpha_i \mathbf{s}_i \beta_i + \bar{\alpha}_i \mathbf{S}_i$ represents no packet arrival while in transient state or a simultaneous arrival of one packet and a departure of another packet of the same priority. The backward transitions sub-matrix $\mathbf{A}_{2,i} = \bar{\alpha}_i \mathbf{s}_i \beta_i$ captures the case of a packet being dispatched, leading to a decrement of the i-th queue packets. Finally, the boundary sub-matrix $\mathbf{B}_{2,i} = \alpha_i \mathbf{s}_i \beta_i + \mathbf{S}_i$ captures the events when the i-th queue is full. Note that packets of the i-th priority that arrive in this state are lost due to queue overflow. Due to the embedded vacation model, the initialization vector is expressed as $\beta_i = [1 \;\; 0_{m_i-1}]$ has only 1 at the serving state and zeros otherwise.

In order to construct the QBD via (3.4), the stochastic transient matrices $\mathbf{S}_i$ and $\mathbf{S}_{0,i}$ need to be computed. We first present preliminary definitions that facilitate the construction of $\mathbf{S}_{0,i}$ and $\mathbf{S}_i$. Let χ_i denote the probability that server starts a vacation while serving the i-th priority queue. Due to the adopted preemptive priority discipline, a vacation starts upon the arrival of any of the higher priority packets. Exploiting the independence between the traffic streams, χ_i equals

$$\chi_i = 1 - \prod_{m=1}^{i-1} \bar{\alpha}_m. \tag{3.5}$$

Let $\mathbf{v}_i \in \mathbb{R}^{1 \times m_i - 1}$ denotes the vacation initialization vector, which have only non-zero values at the legitimate initial vacation states. The two level PH type distribution used to build the QBD in (3.4) is constructed through the following proposition.

Proposition 1. *The stochastic transient matrices of the $i+1$-th priority queue with transmission success probability $P_{s,i}$, for the boundary $\mathbf{S}_{0,i+1}$ and non-boundary $\mathbf{S}_{i+1}$ states, are evaluated as*

$$\mathbf{S}_{0,i+1} = \tilde{\mathbf{S}}_{i+1}, \;\; \mathbf{S}_{i+1} = \tilde{\mathbf{S}}_{i+1} \odot \mathcal{Q}([\mathcal{I}_{m_i}]_{1,1}, \bar{P}_{s,i}), \text{ such that } \tilde{\mathbf{S}}_{i+1} = \begin{bmatrix} \bar{\chi}_i & \chi_i \mathbf{v}_i \\ \tilde{\mathbf{v}}_i & \mathbf{V}_i \end{bmatrix},$$

where the operator $\mathcal{Q}([\mathbf{A}]_{i,j}, b)$ replaces the element in the i-th row and j-th column of $\mathbf{A}$ with the scalar b, $\mathbf{V}_i = \mathbf{D}_i \mathbf{P}_i \mathbf{D}_i^T$ is the vacation visit matrix, χ_i is given in (3.5), $\mathbf{v}_i$ is the vacation initialization vector and $\tilde{\mathbf{v}}_i = 1_{m_i-1} - \mathbf{V}_i 1_{m_i-1}$ is the absorption vector. In addition, $\mathbf{D}_{i-1}$ is the selection matrix and equals $\mathbf{D}_{i-1} = [0_{m_i-1} \;\; \mathbf{I}_{m_i-1}]$.

Proof. See Appendix A.1. ■

Based on Proposition 1, the vacation states are initialized through the vector $\chi_i \mathbf{v}_i$ (black arrows in Figure 3.3), while all the vacation phases are captured by $\mathbf{V}_i$ (golden arrows in Figure 3.3). Successful

[3] Those matrices capture the transitions between idle to idle, idle to serving i-th priority queue, idle to vacation (serving $1 \leq j < i$ priority queues) and their complementary directions.

transmission of higher priority packets (i.e., end of vacation) is captured by $\tilde{\mathbf{v}}_i$ (green arrows in Figure 3.3). At this point, the steady state distribution of each priority queue at each device can be evaluated. Let $\mathbf{A}_i = \mathbf{A}_{0,i} + \mathbf{A}_{1,i} + \mathbf{A}_{2,i}$ and let π_i represent the unique solution of $\pi_i \mathbf{A}_i = \pi_i$, with the normalization condition $\pi_i 1_{m_i} = 1$. Since finite queues are considered at the devices, one is interested to determine the critical arrival probability after which the probability of having full queues starts to dominate and the queues tend to be always non-empty [75]. Through the rest of the chapter, we use the term overflow (non-overflow) region to denote operating beyond (below) such a probability. Mathematically, for the DTMC in (3.4) to be in the non-overflow region, the following condition must be satisfied

$$\pi_i \mathbf{A}_{2,i} 1_{m_i} > \pi_i \mathbf{A}_{0,i} 1_{m_i}. \tag{3.6}$$

The condition in (3.6) ensures that the departure probability of packets is higher than the arrival probability of packets, which ensures a low overflow probability. Consequently, the overflow probability can be highly reduced by increasing the queue size. When (3.6) is not satisfied, this implies that the packet departures cannot cope with the packet arrivals.

Let $\mathbf{x}_i = \begin{bmatrix} \mathbf{x}_{i,0} & \mathbf{x}_{i,1} & \cdots & \mathbf{x}_{i,q_i} \end{bmatrix}$ be the steady state probability vector where $\mathbf{x}_{i,j}$ incorporates the joint probabilities of having j i-th priority packets and all possible combinations of number of packets with priority higher than i. In particular, let $\mathbb{P}\{n_1, n_2, n_3, \cdots, n_i\}$ denotes the joint probability of having n_1 packets at the first priority queue, n_2 packets at the second priority queue and so on until n_i packets at the i-th priority queue, $\mathbf{x}_{i,j}$ can be represented as

$$\mathbf{x}_{i,j} = \Big[\mathbb{P}\{(\underbrace{0, \cdots, 0}_{i-1}, j)\} \cdots \mathbb{P}\{(k_1, \cdots, 0, j)\} \cdots \mathbb{P}\{(k_1, \cdots, 1, j)\} \cdots \mathbb{P}\{(k_1, \cdots, k_{i-1}, j)\} \Big].$$

In addition, let the scalar $x_{i,j}$ represents the probability of having j packets in the i-th priority queue, which is evaluated as $x_{i,j} = \mathbf{x}_{i,j} 1_{m_i}$. By virtue of the adopted preemptive discipline and observing Figure 3.2, it is clear that the third priority queue is only granted service when all higher priority queues are empty. Thus, the transmission probability $\mathcal{T}_i$ can be computed as

$$\mathcal{T}_i = \sum_{z_i=0}^{k_i} \mathbb{P}\{(0, 0, \cdots, 0, z_i)\}, \tag{3.7}$$

whereas for the first priority queue $\gamma_1 = 1$. Let $r_i = m_i(q_i + 1)$ be the number of possible states for the i-th queue, then the steady state solution for a stable system is characterized as follows.

Lemma 1. *The steady state distribution for the i-th queue with state transition matrix* $\mathbf{P}_i$ *is*

$$\mathbf{x}_i = 1_{r_i} \big(\mathbf{P}_i - \mathbf{I}_{r_i} + \mathcal{I}_{r_i} \big)^{-1}. \tag{3.8}$$

Proof. Since we are considering finite DTMC based on (3.4), the steady state vector $\mathbf{x}_i$ satisfies

$$\mathbf{x}_i \mathbf{P}_i = \mathbf{x}_i, \; \mathbf{x}_i 1_{r_i} = 1, \tag{3.9}$$

which is in the form of $\mathbf{A}\mathbf{x} = \mathbf{b}$. Employing [76, Lemma 1], the lemma can be proved. ■

3.3.2 Matrix Analytic Method Solution

The mathematical complexity required for the inversion in (3.8) can be cumbersome, specially for large number of priority classes and large queue sizes q_i. Thus, a less-complex and mathematically tractable solution is sought. To this end, the matrix analytic method is a powerful mathematical tool which is most suited to Markov chains with QBD structure [74],[29]. Based on the state transition matrix defined in (3.4), the following lemma derives the steady state distribution for the i-th priority queue.

Lemma 2. *The steady state distribution based on the matrix analytic method for the i-th queue is*

$$\mathbf{x}_{i,j} = \begin{cases} \Upsilon_i \mathbf{A}_{2,i} \big(\mathbf{I}_{m_i} - \mathbf{B}_{1,i}\big)^{-1}, & j = 0, \\ \Upsilon_i, & j = 1, \\ \mathbf{x}_{i,1} \mathbf{R}_i^{i-1}, & j > 1, \end{cases} \tag{3.10}$$

where $\Upsilon_i = \mathbf{x}_{i,0}\mathbf{C}_i\big(\mathbf{I}_{m_i} - \mathbf{A}_{1,i} - \mathbf{R}_i\mathbf{A}_{2,i}\big)^{-1}$ *and* $\mathbf{R}_i = \mathbf{A}_{0,i}(\mathbf{I}_{m_i} - \mathbf{A}_{1,i} - \mathbf{A}_{0,i}1_{m_i}\beta_i)^{-1}$ *is the matrix analytic method. In addition, (3.10) must satisfy the normalization* $\mathbf{x}_{i,0}1_{m_i} + \Upsilon_i(\mathbf{I}_{m_i} - \mathbf{R}_i)^{-1}1_{m_i} = 1$.

Proof. Based on [74, 29], $\mathbf{R}_i$ is the minimal non-negative solution to the quadratic equation $\mathbf{R}_i = \mathbf{A}_{0,i} + \mathbf{A}_{1,i} + \mathbf{A}_{2,i}$. Let $\mathbf{x}_{i,0}$ and $\mathbf{x}_{i,1}$ be the solution to

$$\begin{bmatrix} \mathbf{x}_{i,0} & \mathbf{x}_{i,1} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{i,0} & \mathbf{x}_{i,1} \end{bmatrix} \begin{bmatrix} \mathbf{B}_{1,i} & \mathbf{C}_i \\ \mathbf{A}_{2,i} & \mathbf{A}_{1,i} + \mathbf{R}_i\mathbf{A}_{2,i} \end{bmatrix}. \tag{3.11}$$

The employed DTMC has an advantageous feature that can be exploited, since $\mathbf{A}_{2,i}$ is a rank one matrix, which simplifies $\mathbf{R}_i$ to $\mathbf{R}_i = \alpha_i\mathbf{S}_i(\mathbf{I}_{m_i} - \bar{\alpha}_i\mathbf{s}_i\beta_i - \bar{\alpha}_i\mathbf{S}_i - \alpha_i\mathbf{S}_i1_{m_i}\beta_i)$. Given that (3.6) is satisfied, $\mathbf{R}_i$ has a spectral radius less than one [29]. The solution to (3.11) is

$$\mathbf{x}_{i,0} = \alpha_i\mathbf{x}_{i,0}\mathbf{S}_0(\mathbf{I}_{m_i} - \alpha_i\mathbf{s}_i\beta_i - \bar{\alpha}_i\mathbf{S}_i - \mathbf{R}_i\bar{\alpha}_i\mathbf{s}_i\beta_i)^{-1}\bar{\alpha}_i\mathbf{s}_i\beta_i(\mathbf{I}_{m_i} - \bar{\alpha}_i\mathbf{S}_{0,i})^{-1}, \tag{3.12}$$

with the normalization $\mathbf{x}_{i,0}1_{m_i} + \alpha_i\mathbf{x}_{i,0}\mathbf{S}_0(\mathbf{I}_{m_i} - \alpha_i\mathbf{s}_i\beta_i - \bar{\alpha}_i\mathbf{S}_i - \mathbf{R}_i\bar{\alpha}_i\mathbf{s}_i\beta_i)^{-1}(\mathbf{I}_{m_i} - \mathbf{R}_i)^{-1}1_{m_i} = 1$. Finally, $\mathbf{x}_{i,1}$ is obtained through solving (3.11) and $\mathbf{x}_{i,j} = \mathbf{x}_{i,1}\mathbf{R}^{-1}$. Substituting the component stochastic matrices, the lemma can be reached. ■

3.3.3 Vacation Model Verification

As verification, the proposed vacation-based preemptive model is compared against the conventional method presented in [29, Chapter 9] for the case of $N = 2$. Assuming a hypothetical fixed service probability $P_{\mathrm{s},i}$, Figure 4 compares the conventional method with the proposed one. It is observed that the vacation-based model exactly characterizes the priority queues evolution while offering a computationally convenient, tractable, and scalable model for larger number of priority classes, whereas for higher values of N, the conventional method becomes highly complex.

It is clear that in order to compute the steady state distributions $\mathbf{x}_{i,j}$ of the i-th queue, one need to compute $P_{\mathrm{s},i}$. Such inter-dependency highlights the interaction between the microscopic and macroscopic

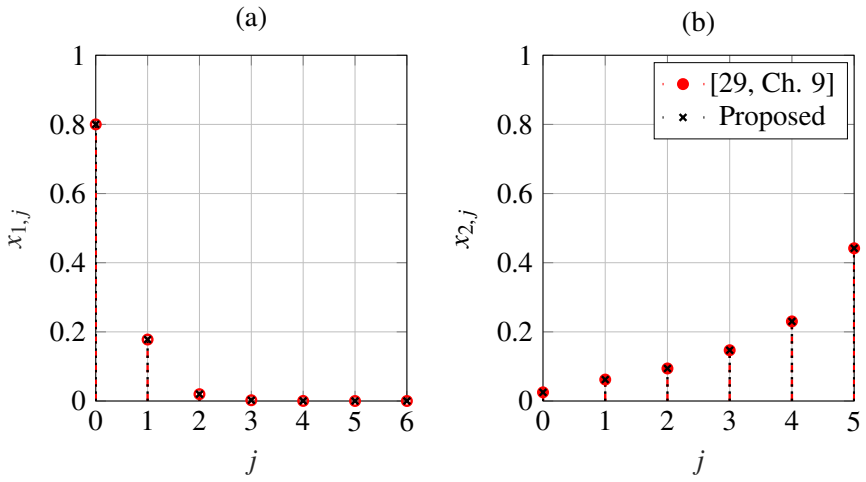


Figure 3.4: Steady state probabilities for (a) first (b) second priority class for $q_1 = 6, q_2 = 5, \alpha_1 = 0.1, \alpha_2 = 0.5$ & $P_{\mathrm{s},1} = P_{\mathrm{s},2} = 0.5$.

scales in the network. In what follows, we present the framework adopted to characterize $P_{\mathrm{s},i}$ based on stochastic geometry analysis.

3.4 Spatial Macroscopic Analysis

Based on (3.1), it is clear that $P_{\mathrm{s},i}$ is a function of the aggregate network mutual interference induced by the macroscopic interactions between the devices. This section utilizes stochastic geometry to delve into the network-wide interactions between devices and characterizes the transmission success probability defined in (3.1). Before proceeding further, we state two commonly used and core approximations that are utilized in this chapter for tractability and mathematical convenience.

Approximation 1. *(i) The spatial correlations between adjacent Voronoi cell areas are ignored. (ii) All devices in the network are assumed to perform (i.e., in terms of transmission success probability) as the typical device located at the origin.*[4]

Remark 1. *(i) Implies that all devices will have independent and identically distributed transmit powers to invert their path-loss to the serving BS. Such assumption is commonly used and verified in the literature [77, 73]. (ii) For static networks, the transmission success probability is location dependent, which is captured via the meta distribution [78] and can be incorporated to the spatiotemporal analysis as in [39, 38]. However, it is shown [35, 36, 79, 80] that such location dependence diminishes with path-loss inversion and random channel selection. (iii) The device becomes typical by spatial averaging. That is, the typical device's performance is obtained by averaging over different network realizations, fading parameters, and queue states [22]. Hence, no generality is lost in studying the statistics seen by the typical device.*

[4]Both approximations are validated in Section 5.3.5 against independent Monte Carlo simulations.

Exploiting Approximation 1(i) and 1(ii), the transmission success probability of an i-th priority packet transmitted from a typical device located at the origin can be further expressed as

$$P_{\mathrm{s},i} = \mathbb{P}\{\mathsf{SINR}_i > \theta\} = \mathbb{P}\Big\{\frac{\rho h_o}{I_i + \sigma^2} \geq \theta\Big\}, \tag{3.13}$$

where h_o is the channel gain between the device and its serving BS, σ^2 is the noise power, and I_i is the aggregate interference seen by an i-th priority packet, which is expressed as

$$I_i = \sum_{y_j \in \Psi \setminus y_o} 1_{\{a_{i,j}\}} P_j g_j ||y_j - z_o||^{-\eta}, \tag{3.14}$$

where y_j is the location of an interfering device (all active devices will be interfering except the typical device $\Psi \setminus y_o$), $a_{i,j}$ is the event that the device located at y_j is transmitting on the same channel as the typical device, P_j is its transmit power, g_j is the channel power gain between the interfering device and the serving BS, $||.||$ is the Euclidean norm, and z_o is the typical device's serving BS's location.

Remark 2. *It is worth noting that $P_{s,i}$ across different priority classes will only be different for the dedicated channel allocation, where the channel selection is dependent on the packet priority. Hence, a device sending an i-th priority packet may only experience interference form devices transmitting packets of the same priority. However, for the case of shared channel allocation, the transmission success probability is agnostic to packets priorities.*

Due to the assumed exponential distribution of h_o, the channel inversion power control and the definition of the Laplace transform (LT), (3.13) can be expressed as

$$P_{\mathrm{s},i} = \exp\Big\{-\frac{\sigma^2\theta}{\rho}\Big\}\mathcal{L}_{I_i}\Big(\frac{\theta}{\rho}\Big), \tag{3.15}$$

where $\mathcal{L}_{I_i}(.)$ is the LT of the aggregate interference I_i. One can observe from (3.15) the effect of fading, power control, and decoding threshold on the achieved transmission probabilities, which in return affects the queues temporal evolution. Thus, coupling the queues departure probabilities and the aggregate interference in the network. In the remaining of this section, we characterize the transmission success probability for three different channel allocation strategies.

3.4.1 Dedicated allocation

This scheme considers an orthogonal allocation among the active queues based on their priority. The interfering sources to an active transmission of the i-th priority queue can only be from the set of all active devices having packets to be transmitted in their i-th priority queue. The transmission success probability of an i-th priority packet under the dedicated allocation is derived as follows.

Theorem 1. *The transmission success probability $P_{s,i}$ of a packet belonging to the i-th priority class under the dedicated allocation strategy is given by*

$$P_{s,i} \approx \frac{\exp\left\{-\frac{\sigma^2\theta}{\rho} - \frac{2\theta\mathcal{J}_i\kappa_{i,m}}{(\eta-2)}{}_2F_1(1,1-2/\eta,2-2/\eta,-\theta)\right\}}{\left(1+\frac{\theta\mathcal{J}_i\kappa_{i,m}}{(1+\theta)c}\right)^c},$$
$$\overset{(\eta=4)}{=} \frac{\exp\left\{-\frac{\sigma^2\theta}{\rho} - \mathcal{J}_i\kappa_{i,m}\sqrt{\theta}\arctan\left(\sqrt{\theta}\right)\right\}}{\left(1+\frac{\theta\mathcal{J}_i\kappa_{i,m}}{(1+\theta)c}\right)^c}, \tag{3.16}$$

where $\mathcal{J}_i = \sum_{z_i=1}^{k_i}\mathbb{P}\{(0,0,\cdots,0,z_i)\}$ is the joint probability of having no packets with priority higher than i and at least a packet with priority i, $\kappa_{i,m} = \frac{\mu}{\lambda C_{i,m}}$ is the average number of devices per BS per channel, where $m \in \{EA, WA\}$ indicates equal-allocation or weighted-allocation dedication strategy. ${}_2F_1(\cdot)$ is the Gaussian hyper-geometric function that is defined as ${}_2F_1(a,b,u;z) = \sum_{k=0}^{\infty}\frac{(a)_k(b)_kz^k}{(u)_kk!}$ and $c = 3.575$. The approximation is due to Approximation 1(i) and the employed approximate probability density function (PDF) of the PPP Voronoi cell area in $\mathbb{R}^2$ as shown in (A.4).

Proof. See Appendix A.2. ■

The parameter $\kappa_i\mathcal{J}_i$ represents the portion of devices attempting a transmission of an *i*-th priority packet. Thus, interfering on the typical device that is attempting the transmission of its own *i*-th priority packet. Moreover, κ_i is affected by the number of channels assigned to each priority class. Through this chapter, we investigate two dedicated channel allocation strategies; namely, equal allocation (EA) and weighted allocation (WA). The former equally splits the total available channels among the existing priority classes, whereas the latter considers an allocation of channels that is dependent on that given priority class arrival probability. Mathematically, the number of allocated channels for the equal and weighted schemes are expressed as

$$C_{i,\mathrm{EA}} = \frac{C}{N},\ C_{i,\mathrm{WA}} = C\frac{\alpha_i}{\sum_{j=1}^{N}\alpha_j}. \tag{3.17}$$

3.4.2 Shared allocation

This strategy considers the case of inter-class channel multiplexing among all the active devices irrespective of the packet's priority that is to be transmitted. That is, all the active devices can mutually interfere regardless of the priority of the packets being transmitted. Hence, all the devices with non-empty queues are potential interferers to the typical device's packet. Recalling the preemptive-based mechanism, the probability of being a potential interferer is the complement of the joint probability that all the *N* priority queues are empty. In the following theorem, the transmission success probability of an *i*-th priority packet under the shared allocation is derived.

Theorem 2. *The transmission success probability $P_{s,i}$ of a packet belonging to the i-th priority class under the shared allocation strategy is given by*

$$P_{s,i} \approx \frac{\exp\left\{-\frac{\sigma^2\theta}{\rho} - \frac{2\theta\bar{\mathcal{J}}_0\kappa}{(\eta-2)}{}_2F_1(1,1-2/\eta,2-2/\eta,-\theta)\right\}}{\left(1+\frac{\theta\bar{\mathcal{J}}_0\kappa}{(1+\theta)c}\right)^c},$$
$$\overset{(\eta=4)}{=} \frac{\exp\left\{-\frac{\sigma^2\theta}{\rho} - \bar{\mathcal{J}}_0\kappa\sqrt{\theta}\arctan\left(\sqrt{\theta}\right)\right\}}{\left(1+\frac{\theta\bar{\mathcal{J}}_0\kappa}{(1+\theta)c}\right)^c}, \tag{3.18}$$

where $\mathcal{J}_0 = \mathbb{P}\{(0,0,\cdots,0,0)\}$ is the joint probability of having no packets in all the N priority queues, $\kappa = \frac{\mu}{\lambda C}$ is the average number of devices per BS per channel, and ${}_2F_1(\cdot)$ is the Gaussian hyper-geometric function and $c = 3.575$. The approximation is due to Approximation 1(i) and the employed approximate PDF of the PPP Voronoi cell area in $\mathbb{R}^2$ as shown in (A.4).

Proof. Since all the packets being transmitted experience the same aggregate interference under the shared allocation, $P_{s,i}$ of all the queues are identical. Furthermore, a device is attempting a transmission if it has any packets within its N priority queues. Thus, the portion of interfering devices within the network is $\mu\bar{\mathcal{J}}_0$, where $\mathcal{J}_0$ is the joint probability of having no packets in all the N priority queues. Finally, the theorem is realized following similar steps as Theorem 1. ■

In summary, the shared channel allocation strategy aims at allowing the devices to utilize all available channels. Thus, a given device will have a larger pool of channels to utilize for its transmission, while experiencing mutual interference from different priority transmission. On the other hand, the dedicated strategies provides a limited number of the channels for a given class, based on an allocation criteria, either equally or proportionally. This prohibits mutual interference from different priority transmission.

3.4.3 Iterative Solution

As discussed in Section 3.3, the idle probability of an i-th priority queue employed at a given IoT device governs the interference it causes within the network. In addition, the aggregate network interference affects the idle probability of each device. Thus, an inter-dependency exists between the devices activity and aggregate interference scales. Such inter-dependency can be solved iteratively as presented in Algorithm 1, which converges uniquely to a solution by virtue of the fixed point theorem [81]. Regarding the complexity, the dedicated allocation scheme is considered to be more complex compared to the shared one, as it requires an additional coordination step to compute the portion of channels available to each priority class. In order to conduct this, prior knowledge of the number of priority classes or the arrival probabilities α_i are required for the equal and weighted-allocation strategies, respectively. On the other hand, the shared allocation alleviates such step, as all the channels are available irrespective of the packet's priority class.

Algorithm 1 Iterative computation of $P_{\mathrm{s},i}$ and $\mathbf{x}_i$ for dedicated and shared channel allocation

procedure $((\alpha_1\ \alpha_2\ \cdots\ \alpha_N), \lambda, \mu, \eta, \theta, C, \varphi)$ ▷ φ is a convergence tolerance parameter
  initialize $\mathcal{J}_0$ and $\mathcal{J}_i \in [0,1]$
  while $||\mathbf{x}_i^k - \mathbf{x}_i^{k-1}|| \geq \varphi$ **do**
    Compute $P_{\mathrm{s},i}$ from (3.16)-dedicated or (3.18)-shared.
    Construct $\mathbf{S}_{0,i}$ and $\mathbf{S}_i$ from Proposition 1.
    if $\pi_i \mathbf{A}_{2,i} 1 > \pi_i \mathbf{A}_{0,i} 1$ **then** ▷ non-overflow (i.e., stability) condition
      Solve $\mathbf{x}_i$ based on Lemma 1 or Lemma 2.
      Compute $\mathcal{J}_0$ and $\mathcal{J}_i$ from $\mathbf{x}_i$.
      Compute $P_{\mathrm{s},i}$ from (3.16)-dedicated or (3.18)-shared.
    else
      Set $\mathcal{J}_0 = 0$ and calculate $\mathcal{J}_i$.
      Compute $P_{\mathrm{s},i}$ from (3.16)-dedicated or (3.18)-shared.
      Break.
    end if
    Increment k.
  end while
  return $P_{\mathrm{s},i}$ and $\mathbf{x}_i\ \forall i$.
end procedure

3.4.4 Performance Metrics

Based on the provided iterative framework, once can evaluate the steady state distribution of the N priority queues. To this end, a number of KPIs can be evaluated, which are insightful when designing and assessing massive prioritized multi-stream traffic IoT networks.

First, the departure probability is evaluated as $d_i = \mathcal{T}_i P_{\mathrm{s},i}$, where $\mathcal{T}_i$, defined in (3.7), is the probability that the sever is available to serve the i-th priority packet. Articulated differently, $\mathcal{T}_i$ is the probability that all higher priority queues are empty such that the device is able to send an i-th priority packet. Such transmission attempt succeeds with probability $P_{\mathrm{s},i}$ as given by (3.16) for the dedicated allocation and (3.18) for shared allocation. Let Q_i be the instantaneous number of packets at the i-th queue, then the average number of packets is

$$\mathbb{E}\{Q_i\} = \sum_{n=1}^{q_i} n\mathbb{P}\{Q_i = n\} = \sum_{n=1}^{q_i} n x_{i,n}. \tag{3.19}$$

For the i-th priority packet, its transmission will be postponed till all the packets belonging to higher classes are successfully served. Transmission availability for the i-th priority class in a generic device denotes the probability that the i-th priority queue is non-empty and that all higher priority queues are empty. Thus, transmission availability is evaluated as

$$\mathcal{A}_i = 1 - \sum_{j=1}^{i-1} \sum_{m_j=1}^{k_j} x_{j,m_j}. \tag{3.20}$$

A critical KPI in prioritized traffic is the information freshness, which is quantified via the age of information [82]. Specifically, we focus on the PAoI, which is defined as

$$\Delta_{\mathrm{p},i} = \mathbb{E}\{\mathcal{I}_i\} + \mathbb{E}\{\mathcal{W}_i\} + \mathbb{E}\{D_i\}, \tag{3.21}$$

where $\mathbb{E}[\mathcal{W}_i], \mathbb{E}[D_i]$ and $\mathbb{E}[\mathcal{I}_i]$ denote the average queueing delay, average transmission delay and inter-arrival delay, respectively. Based on the adopted geometric distribution for packets arrival, the average inter-arrival times simplifies to $\mathbb{E}[\mathcal{I}_i] = \frac{1}{\alpha_i}$. In addition, let W_i be the queueing delay (i.e., number of time slots spent in the queue before the service of the i-th priority queue starts) for a randomly selected packet, then the average queueing delay is given by

$$\mathbb{E}\{\mathcal{W}_i\} = \sum_{n=0}^{\infty} n\mathbb{P}\{\mathcal{W}_i = n\}, \tag{3.22}$$

where the temporal distribution of the delay (i.e., across different packets) can be obtained as $\mathbb{P}\{\mathcal{W}_i = 0\} = x_{i,0}$ and $\mathbb{P}\{\mathcal{W}_i = j\} = \sum_{k=1}^{j} \mathbf{x}_{i,k}\mathbf{G}_j^{(k)}\mathbf{1}$, where $\mathbf{G}_{i,j}^{(k)}$ represents the probability of having k packets in the i-th priority queue and being serviced in j time slots with

$$\mathbf{G}_{i,j}^{(k)} = \begin{cases} \mathbf{S}_i^{j-1}\mathbf{s}_i\beta_i & k = 1, \\ (\mathbf{s}_i\beta_i)^k & j = k, k \geq 1, \\ \mathbf{S}_i\mathbf{G}_{i,j-1}^{(k)} + \mathbf{s}\beta\mathbf{G}_{i,j-1}^{(k-1)} & k \geq j \geq 1. \end{cases} \tag{3.23}$$

Based on the considered PH type distribution for the vacation duration, let W_i be the number of time slots spent in the queue before the service starts for a randomly chosen packet. Averaging over all packets, the transmission delay can be computed as [29, Section 2.5.3]

$$\mathbb{E}\{D_i\} = \beta_i(\mathbf{I}_{m_i} - \mathbf{S}_i)^{-1}. \tag{3.24}$$

Finally, the PAoI is evaluated by plugging (3.22) and (3.24) into (3.21).

3.5 Simulation Results

Through this section various numerical results are presented that aim at (a) validating the proposed analytical model; (b) highlighting the influence of the different channel allocation strategies, and (c) showing priority-aware wireless-based system design insights.

3.5.1 Simulation Methodology

The developed simulation framework incorporates microscopic and macroscopic averaging, where the former addresses the steady state temporal statistics of the different queues employed at each device and the latter addresses the stochastic geometric network-wide performance. The simulation area is $10 \times 10\text{km}^2$ with a wrapped-around boundaries to ensure unbiased statistics imposed by the network boundary devices.

Unless otherwise stated, we consider the following physical layer parameters: $\kappa = 1$ devices/BS/channel, $C = 64$ channels, $\eta = 4$ and $\rho = \sigma^2 = -90$ dBm. For the MAC layer parameters, we consider three priority classes with $(\alpha_1, \alpha_2, \alpha_3) = (0.1, 0.25, 0.35)$, where all the queues have equal size (i.e., $q_1 = q_2 = q_3 = 8$). The proposed priority-aware transmission schemes are compared to a reference multi-stream priority agnostic (PA) FCFS queueing model. In such model, the transmission is granted on an FCFS basis, equally among the all existing N priority classes.

Synchronous time-slotted system is adopted and each microscopic simulation run is considered as a time slot where independent channel gains are instantiated and packets are generated probabilistically. The queue occupancy for each of the considered priority classes are tracked. For a transition from one time slot to another, packets are independently generated at every device for all queues based on the batch arrival process (i.e., α_i). Every device with a non-empty queue of the N queues tries to communicate its backlogged packets with its serving BS based on the employed preemptive priority-aware transmission strategy. For a device with non-empty i-th priority queue, a packet is dispatched from the i-th priority queue if and only if i) all higher priority queues are empty, and ii) the achieved uplink SINR_i on the selected channel is greater than θ. In order to ensure that the different queues at the devices are in steady state, simulation is first initiated with all queues at the devices as being idle and then it runs for a sufficiently high number of time slots until the steady-state is reached. Let $\hat{\mathbf{x}}^k = [\mathbf{x}^k_{1,0}, \mathbf{x}^k_{2,0}, \cdots, \mathbf{x}^k_{N,0}]$ denotes the idle steady state probability for the t-th iteration of the N queues. Mathematically, the steady state is realized once $||\hat{\mathbf{x}}^k - \hat{\mathbf{x}}^{k-1}|| < \varphi$, where φ is some predetermined tolerance. After steady state is reached, all temporal statistics are then gathered based on sufficiently large number of microscopic realizations. Finally, the whole process is repeated for sufficiently large number of macroscopic network realizations to ensure spatial ergodicity is reached.[5]

3.5.2 Prioritized Traffic Evaluation and Discussion

We start with the framework validation for all considered priority classes and proposed channel allocation strategies. Figure 3.5 shows thetransmission success probability (TSP) for three priority classes against the decoding threshold θ. The close matching between the theoretical and simulation results validates the developed spatiotemporal mathematical model. Moreover, focusing on a given channel allocation strategy and a priority class for low values of θ, the devices are able to empty their queues and go into idle state when operating below the overflow threshold. This leads to a lower network aggregate interference. As θ increases, the transmission success probability decreases, which leads in turn into having higher aggregate network interference. Based on the prioritized transmission and the assumption that $\alpha_j > \alpha_i$, $\forall j > i$, it is expected that $\mathrm{TSP}_j < d_i$. This is justified as lower priority packets are served only if all the higher priority queues are empty. In addition, it is clear that the SINR threshold θ, at which the system transitions from non-overflow to overflow operation depends on the priority class.[6]

To better assess the performance of the different allocation strategies, Figure 3.6 compares the considered strategies against the priority-agnostic strategy. First, it is observed that for lower values of θ, the dedicated equal-allocation strategy outperforms the shared strategy. This is attributed to the successful

[5] In point processes theory, a point process is said to be spatially ergodic if the spatial averages (across points) equal the ensemble averages (across realizations).

[6] Note that the overflow thresholds depict the point where the probability of queues overflow starts to dominate.

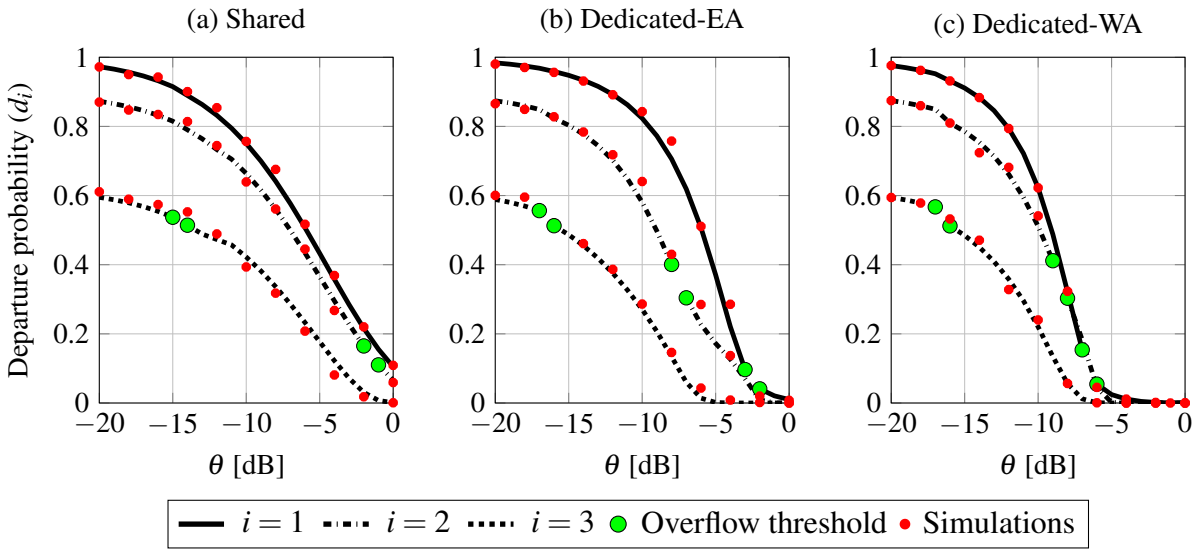


Figure 3.5: TSP for three priority classes as a function of the decoding threshold θ under (a) Shared (b) Dedicated equal allocation (c) Dedicated weighted allocation.

packets transmission attempts from the first priority class while benefiting from interference protection from lower priority classes. As θ increases, the shared strategy outperforms the dedicated one, which results from the head of queue effect of the higher priority packets. In the dedicated strategy, when several devices have high priority packets, they keep interfering on a subset of the available channels leaving other channels for lower priority packets underutilized. Moreover, the dedicated-weighted allocation strategy fails to provide gains in the high θ region, due to the strong interference experienced by the higher priority packets, that are allocated a smaller number of channels (i.e., compared to the dedicated equal-allocation strategy). Thus, hindering the transmission of lower priority packets, that are assigned larger pool of channels, due to the imposed priority-aware transmission discipline. Additionally, the shared channel allocation strategy alleviates the additional overhead required for channel allocation procedures, that is essential for the dedicated strategies. For the priority-agnostic scheme, we observe the performance deterioration experienced by the higher priority classes (e.g., first and second classes), which results from the priority-agnostic negligence of higher priority traffic. For the priority-agnostic scheme, a given packet is granted service depending on its arrival time, not its priority. Accordingly, depending on the arrival probability, transmission probability is larger for traffic with higher arrival probabilities (i.e., third class has larger transmission probability compared to second and first classes). For the third priority class, due to the FCFS nature of the priority-agnostic scheme, it outperforms the priority-aware strategies. Accordingly, the TSP values depict a flipped behavior among the higher and lower priority classes.

To further investigate the prioritization effect, the average packet delay is shown in Figure 3.7.[7] Due to its priority negligence of the priority-agnostic strategy, the packets belonging to the three classes experiences nearly the same waiting time with different values of θ. This is attributed to the inter-class FCFS discipline of the priority-agnostic. However, for the priority-aware strategies, high priority packets experience lower packet delays when compared to lower priority packets. The figure also highlights the traffic prioritization cost on lower priority packets, which is due to the service interruption upon

[7]The delay is defined as the time elapsed from packet generation at the device until its successful reception at the BS.

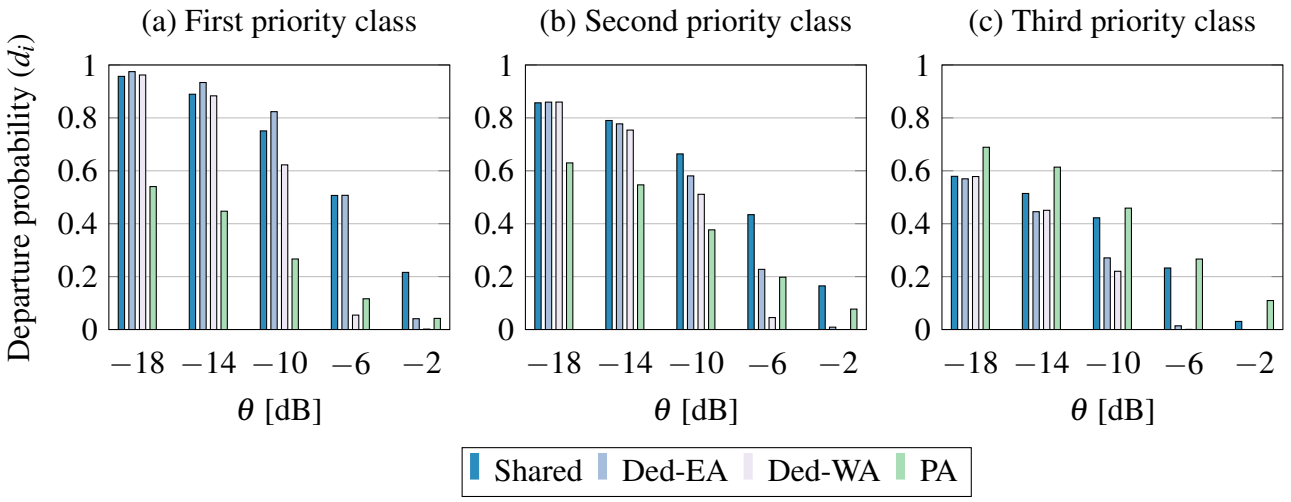


Figure 3.6: Comparison of different allocation strategies for three priority classes.

higher priority packets arrival. Hence, it is important to ensure that the prioritized transmission offers a differentiated service that meets the QoS requirement for all priority classes.

Throughout the rest of this section, we will focus on assessing the shared and dedicated equal-allocation strategies due to their promised performance superiority as shown in Figure To this end, 3.6. Figure 3.8 showcases different KPIs under the mentioned strategies. As a common behavior in all the sub-figures, we observe a large performance superiority of the shared allocation strategy over the dedicated equal-allocation one in the high θ regime. As θ increases, packets transmission is subjected to a more stringent requirement on the achieved SINR. This leads to increased retransmissions, thus, increasing the aggregate network interference. Furthermore, it can be interpreted that for the low θ regime, head of the queue is determined by the arrival priority, whereas for the high θ regime, head of the queue is determined by the prioritized-based preemption discipline. In details, Figure 3.8(a) presents the average number of packets, where it is observed that the shared strategy results in lower number of packets residing in the queues at the high θ regime. Within a given channel allocation strategy, as the priority of the queue gets lower, its average number of packets increases. Packets residing in a given queue will have to wait until all the higher queues are served, while new packets might arrive and accumulate in the queues. The figure also highlights the effect of the queue's priority on the overflow threshold. The transmission availability is presented in Figure 3.8(b). For the first priority class, such a metric equals one as highest priority packets will be served upon their arrival. However, for lower priority packets, the transmission availability decreases. Fig .3.8(c) demonstrates the transmission delay, where it can be observed the superiority of the shared over the dedicated strategy. In addition, Fig .3.8(d) presents the average queueing delay distribution over the first five time slots. The queueing delay distributions is dependent on the prioritization and the allocation strategy. In specific, the distribution tail decays for higher priority classes, whereas for the lower classes, it takes longer to dispatch their packets. We observe also a larger tail for the dedicated strategy, when compared to the shared one over the considered priority classes. Finally, Fig .3.8(d) shows the PAoI. We observe a flipped behavior between the first and second priority classes when considering a given allocation strategy. This is justified based on the PAoI sensitivity to the inter-arrival delays (recall $\alpha_1 = 0.1$ and $\alpha_2 = 0.25$). Such a behavior is expected, since

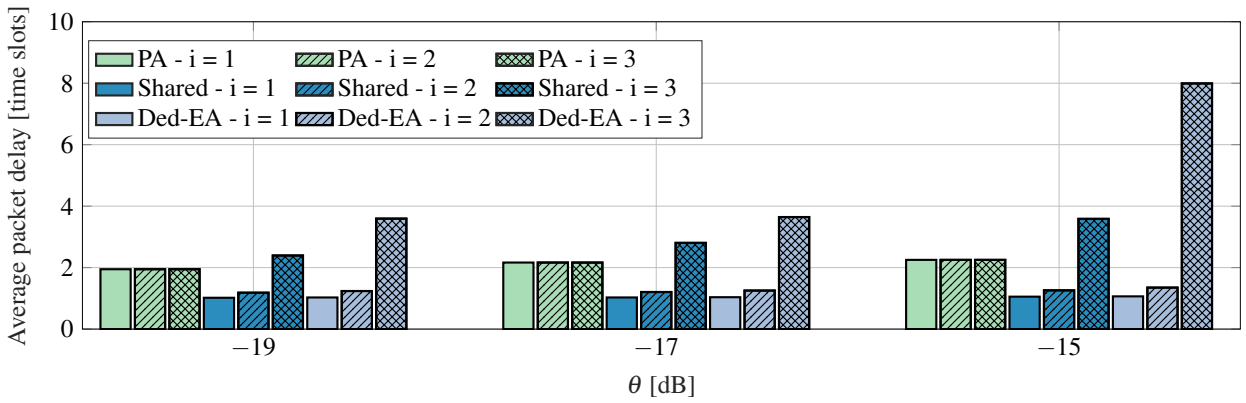


Figure 3.7: Average packet delay for priority agnostic, shared, and dedicated-equal strategies.

PAoI is lower when packets with low queueing delays are delivered regularly. Thus, larger inter-arrival times increases the PAoI. As θ increases, the queueing delays start to dominate the PAoI, yielding the queues eventually in an overflow state. Finally, via observing the reported results in Figure 3.8, it can be concluded that the exclusive resource partitioning for prioritized grant-free uplink traffic in IoT systems is outperformed by the shared channel allocation strategy.

In Figure 3.9, we investigate the effect of network scalability and devices densification of the first two priority queues under shared and dedication-equal-allocation allocation strategies. The considered values of κ represent a network with 640 and 5120 device/KM^2, given that $\lambda = 10$ BS/KM^2 and $C = 64$ channels. First, focusing on the first priority class (c.f. Figure 3.9(a)), we observe a slight superiority of the dedicated equal-allocation over the shared strategy over $\theta \in [-20, -6]$ dB. As mentioned earlier, such performance superiority is attributed to the successful packets transmission attempts from the first priority class while benefiting from interference protection from lower priority classes. Such a behavior is also reflected for $\kappa = 8$ within the range $\theta \in [-20, -14.8]$ dB. Furthermore, as κ increases, a given device experiences stronger interference which degrades the TSP and shifts the overflow-region threshold to lower values of θ. For the second priority class (c.f. Figure 3.9(b)), we observe the superiority of the shared over the dedicated equal-allocation strategy for the two values of κ. This is due to the fact that lower priority classes experience head of the queue problem more severely under the dedicated equal-allocation strategy. Finally, since κ implicitly considers the number of deployed channels at every BS, such a study can help in deriving the minimum number of channels required to meet a targeted requirement.

To showcase the network's stability regions, Figure 3.10 presents the non-overflow region frontiers under shared and dedicated equal-allocation strategies for different system parameters. Such regions ensure queues operating below the overflow threshold, which is represented via the filled area under the curves. The dark (solid lines) and light shaded (dashed lines) represent the shared and dedicated equal-allocation allocation strategies, respectively. First, Figure 3.10(a) shows the relation between the arrival probability of the two highest priority classes (α_1 and α_2) and the decoding threshold θ. As explained in Figure 3.5, larger values of θ leads to higher aggregate network interference, thus, supporting lower traffic arrivals to operate within the non-overflow regions. We observe that for low values of θ, the gap between the shared and dedicated equal-allocation allocation diminishes, since the devices are able to empty their queues nearly easily even under strong mutual interference. As θ increases, the shared

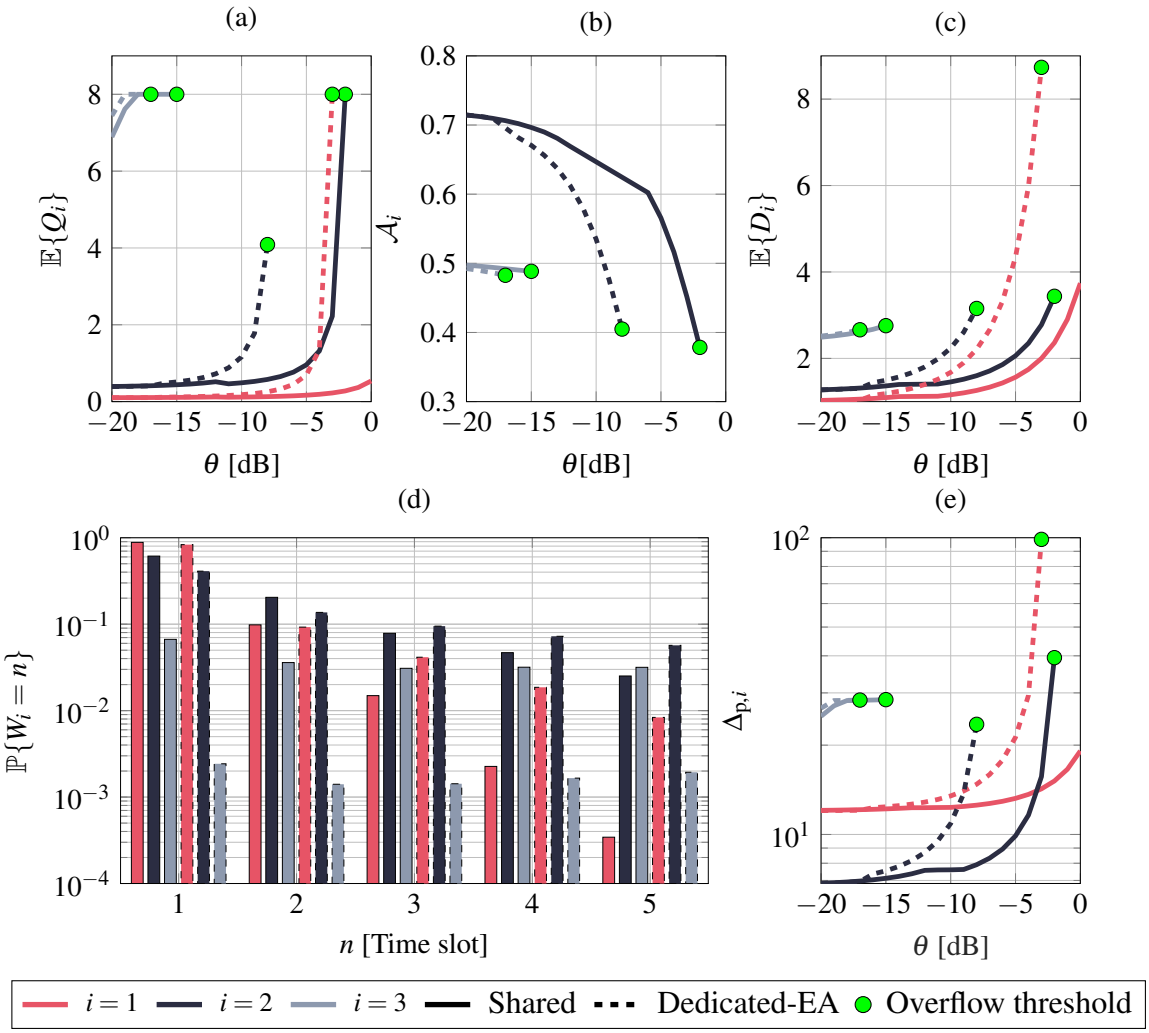


Figure 3.8: Performance evaluation for shared and dedicated-equal allocation strategies (a) average number of packets (b) transmission delay (c) transmission availability (d) waiting time distribution for $\theta = -10$ dB (e) peak age of information.

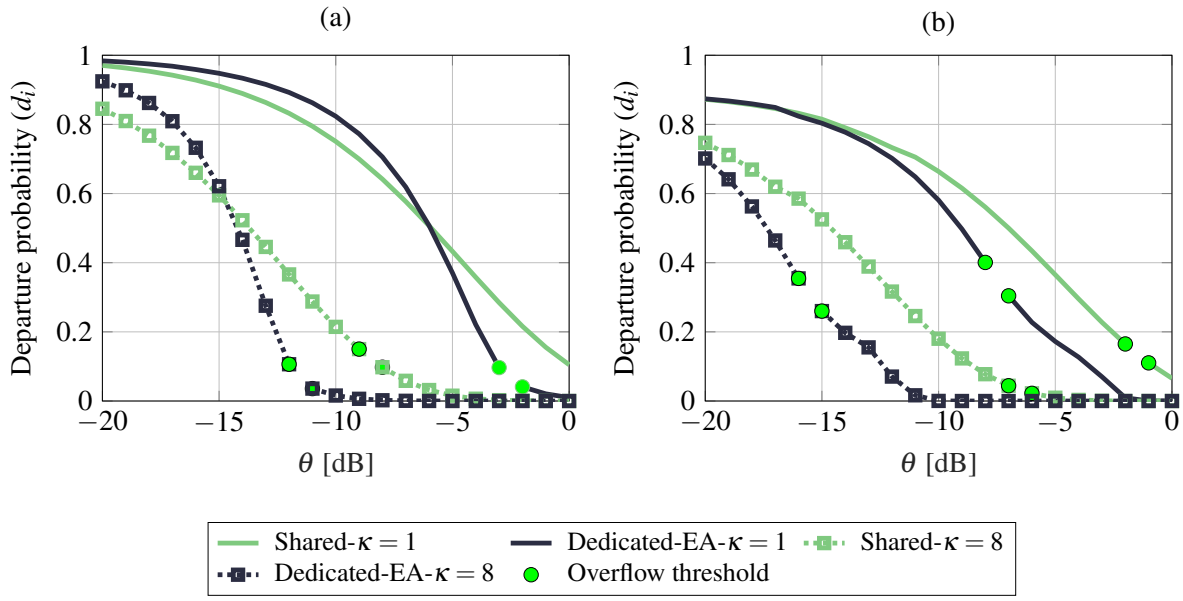


Figure 3.9: Effect of devices densification on the (a) first (b) second priority class.

strategy outperforms the dedicated equal-allocation, since more channels are available for each device for the former strategy. Similarly, Figure 3.10(b) highlights the effect of increasing the third priority packets arrival probability, where the overflow region decreases with larger arrival probabilities. Such a figure can provide interesting insights when studying the relation between different classes of traffic in order to ensure a stable network. The performance comparison between the shared and the dedicated equal-allocation strategies follows Figure 3.10(a). Finally, Figure 3.10(c) focuses on the relation between θ, uplink power control threshold ρ and κ. For a given κ, we can expect that as the uplink transmission can operate under higher thresholds (i.e., higher probabilities), the feasible set of θ ensuring non-overflow operation increases till saturation is reached. This follows from the system transitioning from the noise limited to the interference limited scenario, which is governed by the value of σ^2. On the other hand, as κ increases, the non-overflow region diminishes, which is due to the increased interference within the network. It is important to notice that the shared and dedicated equal-allocation strategies provide similar (θ, ρ) frontiers when considering the network's parameters, since the main dynamics affecting this frontier is radio-related and is oblivious to the adopted resource allocation strategy.

3.6 Conclusion

This chapter presents a tractable and scalable spatiotemporal mathematical framework for large scale uplink prioritized multi-stream traffic in IoT networks. The network is modeled via network of interacting vacation queue, where at the spatial macroscopic scale, interactions occur between different devices due to the mutual interference. At the spacial microscopic scale, interactions among different priority packets occur as the uplink channel can only be utilized by the highest priority packets at the device and is not available to any of the lower priority packets, which is denoted as service vacation. The developed spatiotemporal model is used to assess and compare three priority aware channel allocation strategies; namely dedicated-equal allocation, dedicated weighted allocation and shared allocation

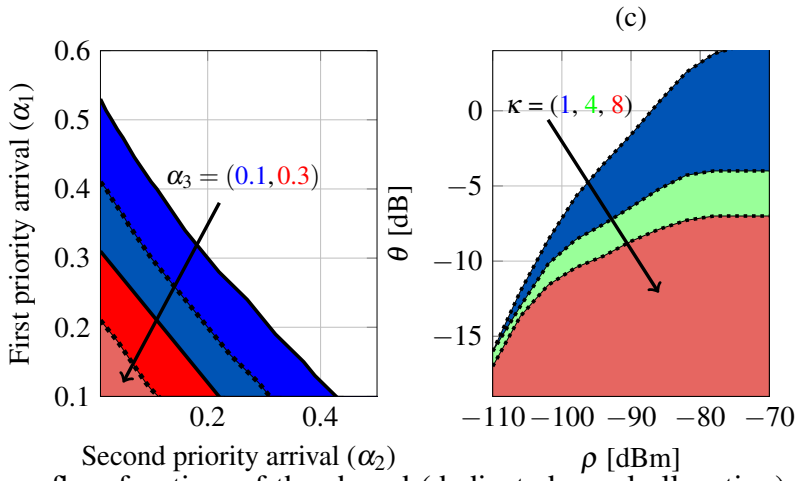


Figure 3.10: Non-overflow frontiers of the shared (dedicated equal-allocation) strategy for different system parameters represented by solid (dashed) lines.

strategy. Numerical evaluations showcase the performance of each priority class in terms of transmission success probability, average queue length, average-delay, delay distribution, and peak age of information. Furthermore, a multi-class priority agnostic scheme is used to benchmark the gains and costs of traffic prioritization on the different priority classes in terms of transmission success probability and average packet delay. The stability of the IoT network is assessed via the Pareto-frontiers of the non-overflow regions. Finally, results indicate the superiority of the shared channel allocation strategy over the dedicated ones, since the former offers higher pool of channels, enabling interference diversification.

Chapter 4

Time and Event-triggered Traffic: An Information Freshness Perspective

In the previous chapter, we considered prioritized multi-stream traffic in large scale networks. Utilizing similar methodology and analytical tools, in this chapter, we introduce a novel spatiotemporal framework that captures the PAoI for large scale IoT uplink network under time and event-triggered traffic. As mentioned in Chapter 2, timely message delivery is a key enabler for IoT and cyber-physical systems to support wide range of context-dependent applications, where conventional time-related metrics (e.g. delay and jitter) fails to characterize the timeliness of the system update. In the foreseen large-scale IoT networks, mutual interference imposes a delicate relation between traffic generation patterns and transmission delays. Numerical evaluations are conducted to validate the proposed mathematical framework and assess the effect of traffic load on the PAoI. The results unveil a counter-intuitive superiority of the event-triggered traffic over the time-triggered one in terms of PAoI, which is due to the involved temporal interference correlations. Insights regarding the network stability frontiers and the location-dependent performance are presented. Key design recommendations regarding the traffic load and decoding thresholds are highlighted.

To this end, information freshness background and our contributions are discussed in Section 4.1. Section 4.2 presents the system model, the underlying physical and MAC parameters, and the PAoI definition. Sections 4.3 and 4.4 discuss the location-dependent characterization of the network-wide interference and the queueing models along with the PAoI characterization, for the two traffic models, respectively. In Section 4.5, various simulation results and observations are discussed. Finally, Section 4.6 summarizes this chapter and draw final conclusions.

4.1 Introduction

Information freshness allows the devices to communicate with proximate devices and learn from their surrounding environment. One key characterization of IoT is the traffic generated by the IoT devices, which governs many of the system key performance indicators [83]. Therefore, it is important to provide a mathematical framework that can characterize the information freshness within large scale uplink IoT networks under different traffic models. IoT traffic can be categorized into time-triggered and event-

triggered traffic [84]. Time-triggered events generate periodic traffic as in vehicular communications, smart grids, and wireless sensor networks [85, 8]. As an example, one may consider a smart monitoring application where devices send timely-based updates to the network's server, resulting in uncoordinated time-triggered (i.e., periodic/deterministic) traffic. In such segments, a central entity collects status updates from multiple nodes (e.g., sensors, vehicles and monitors) through wireless channels. On the other hand, event-triggered traffic arises in scenarios where devices transmit their packets based on detecting random events [86]. Such scenarios can be observed as an example in a given area where power outage occurs. Thousands of devices report their status before the outage occurrence. The IoT network support for the two considered traffic models is crucial to maintain network functionality and attain the required QoSs [87]. Throughout this chapter, we address the critical challenge of how to maintain timely updates within an IoT uplink network under the two aforementioned traffic models.

To position our contribution in context, we first discuss a series of key prior works that studied the AoI and its variants. Authors in [48] consider the system where a sensor generates and transmits update packets to its destination under a FCFS discipline and derive the expression of average AoI for different queueing models. The work in [48] is extended to out-of-order packet delivery in [88]. Last come first serve queue discipline, with and without service preemption, is studied and contrasted to FCFS in [49, 89]. The AoI is characterized in [82] for prioritized packet delivery and in [50] for deterministic traffic. In summary, the previously mentioned works consider only a single sensor scenario.

In addition, a number of works has considered the information freshness in IoT networks with multiple sensors [90, 91, 52]. In particular, authors in [90] consider that one transmitter sends status update packets generated from multiple sensors to the destination, and analyze the average AoI for updates allowing the latest arrival to overwrite the previous buffered ones. In [91], the authors provided an optimization framework to analyze the optimal sampling and updating processes under energy constraints for single and multiple IoT devices. The authors in [92, 93] propose a new metric, namely PAoI, that characterizes the maximum value of the age achieved immediately before receiving a new packet. Focusing on the PAoI, [52] analyzes the system performance by considering a general service time distribution, and optimizes the update arrival rates to minimize its defined PAoI-related cost function. In [94], the authors investigate the role of an unmanned aerial vehicle as a mobile relay to minimize the PAoI. The joint effects of data pre-processing and transmission procedures on the PAoI under Poisson traffic model are investigated in [53].

While the aforementioned works characterize the AoI at the microscopic device level, they overlook the macroscopic impact of aggregate network interference between multiple devices. In the foreseen massively loaded IoT networks, the mutual interference between the active transmitters, trying to utilize the set of finite resources, might hinder timely updates of a given link of interest [62]. Capitalizing on the spatiotemporal perspective, delay and AoI can be characterized and assessed in large scale IoT networks. For instance, lower and upper bounds for the average AoI are proposed under a stochastic geometry framework in [95]. Additionally, AoI under a spatiotemporal framework has recently been investigated in [96], where the authors investigate different scheduling techniques to optimize the PAoI under a spatiotemporal framework. However, the work in [96] focuses on ad hoc networks with Bernoulli traffic arrivals, which is a special case of the event-triggered traffic considered our proposed framework.

To the best of our knowledge, PAoI has not been yet investigated under either of the time-triggered and generalized event-triggered traffic variants in uplink large-scale IoT networks.[1] In addition, the macroscopic and microscopic network scales are addressed through the proposed framework. For the macroscopic aspect, stochastic geometry is utilized to characterize the mutual interference among active devices (i.e., position dependent). In addition, tools from queueing theory are adopted to account for the microscopic queue evolution at each device under the time-triggered and event-triggered traffic models. In summary, the main contributions of this chapter compared to the previously stated works are summarized as follows:

- Develop a novel and tractable mathematical framework, based on stochastic geometry and queueing theory, that characterizes the spatiotemporal interactions under time-triggered and generalized event-triggered traffic models;
- develop a framework that integrates DTMCs and stochastic geometry to characterize and assess the PAoI in large-scale IoT networks under time-triggered and generalized event-triggered traffic models; and
- showcase the Pareto frontiers that characterize the network's stability regions.

4.2 System Model

4.2.1 Spatial & Physical Layer Parameters

An uplink cellular network is considered in this chapter where the BSs are deployed based on a PPP Φ with spatial intensity λ. The IoT devices follow an independent PPP Ψ, such that within the Voronoi cell of every BS $b_i \in \Phi$, a device is dropped uniformly and independently. All devices and BSs are equipped with single antennas. Let r be the distance between a device and its serving BS and $\eta > 2$ be the path-loss exponent, an unbounded path-loss propagation model is considered such that the signal power attenuates at the rate $r^{-\eta}$. Multi-path Rayleigh fading is assumed to characterize the small-scale fading. Additionally, h and g denote the intended and interference channel power gains, and are exponentially distributed with unit power gain. Spatial and temporal independence is assumed for all the channel gains. Fractional path-loss inversion power control is considered at the devices with compensation factor ε. Accordingly, the transmit power of a device positioned r meters is given by $\rho r^{\eta\varepsilon}$, where ρ is a power control parameter to adjust the average received power at the serving BS [73]. In this chapter, a fixed, yet arbitrary network realization of the network is considered to account for the much smaller time scale of the channel fading, packet generation, and transmission when compared to the spatial network dynamics.[2]

[1] When compared to the average AoI, PAoI is considered throughout this chapter because it is more suited to provision QoS and for min-max network design objectives [52, 96].

[2] To analyze the location-dependent performance of the network, we consider a static network where for a generic network realization, Φ and Ψ remain static over sufficiently large time horizon, while device activities, channel fading, and queue states vary each time slot.

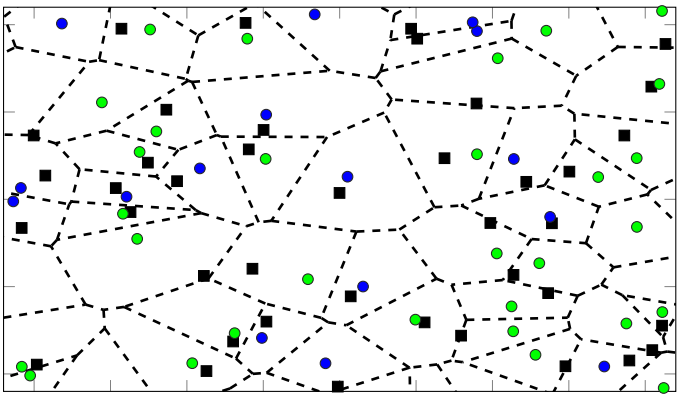


Figure 4.1: A network realization for $\lambda = 10^{-6}$ BS/ km^2. Black squares depict the BSs while green and blue circles represent devices with empty queue and devices with non-empty queues.

4.2.2 Temporal & MAC layer parameters

The proposed framework studies a discretized, time slotted, and synchronized system in which a new packet is generated at a generic device based on time-triggered or event-triggered traffic. For the time-triggered traffic, we consider an asynchronous homogeneous periodic packet generation scheme with duty cycle T and time-slot offset ν. That is, each device in the network generates a packet (e.g., measurement or status update) periodically every T time slots. However, it is not necessary that all devices in the network are synchronized to the same time slot for packet generation. Instead, it is assumed that the offset of the devices $\nu_i \in \{0,1,\cdots,T-1\}, \forall i \in \Psi$ are independently and uniformly distributed among the time slots within the duty cycle T, i.e., $\mathbb{P}\{\nu_i = \tau\} = 1/T,\ \tau \in \{0,1,\cdots,T-1\}$. For a generalized event-triggered traffic, the PH type distribution is employed to capture a wide range of different traffic variants as will be shown in more details in Section 4.4. An FCFS discipline is considered at each device, where failed packets are persistently retransmitted till successful reception. In particular, a packet residing at a generic device is successfully decoded if the received SIR is larger than a detection threshold θ at its serving BS.

In Figure 4.1, a spatiotemporal realization of the network is shown. At a given time slot, two different states of devices can be observed i) active due to non-empty queue and ii) idle due to empty queue. Note that for the time-triggered traffic, all devices with the same offset are synchronized together and become active at the same time slot. Furthermore, two devices with different offsets may become simultaneously active in case of retransmission, where the probability of simultaneous activity depends on the relative offset values between the devices and the decoding threshold θ.

4.2.3 Age of Information in Large Scale Networks

As previously mentioned, AoI quantifies the freshness (i.e., timeliness) of information transmitted by the devices within the network [48]. For any link within the considered time slotted system, the metric $\Delta(t)$ tracks the AoI evolution with time as shown in Figure 4.2. Assume that the i-th packet is generated at

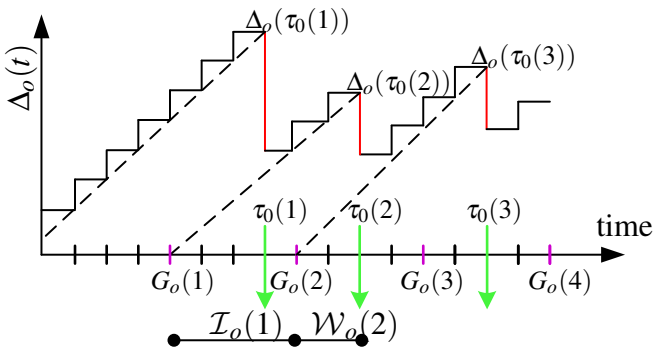


Figure 4.2: AoI evolution of a generic (o-th) device. The time stamps $G_o(i)$ and $\tau_o(i)$ denote the time at which the i-th packet was generated and successfully delivered. $\mathcal{I}_o(1)$ and $\mathcal{W}_o(2)$ denote the inter-arrival time and the waiting times.

time G_i, then $\Delta_i(t+1)$ is computed recursively as

$$\Delta_i(t+1) = \begin{cases} \Delta_i(t)+1, & \text{transmission failure,} \\ t - G_i + 1, & \text{otherwise} \end{cases} \tag{4.1}$$

Through this chapter, we consider the peak AoI, termed through the subsequent sections PAoI, which is defined as the value of age resulted immediately prior to receiving the i-th update [52]. The increased focus on the PAoI stems from the guaranteed system performance insights it unveils. In addition, the minimization of the PAoI may be required for time critical applications [53]. To this end, conditioned on a fixed, yet generic spatial realization, the spatially averaged PAoI,[3] as observed from Figure 4.2, is computed as

$$\mathbb{E}\{\Delta_{\mathrm{p}}|\Psi\} = \mathbb{E}^{!}\Big\{\mathcal{I}_o + \mathcal{W}_o \big| \Psi, \Phi\Big\}, \tag{4.2}$$

where $\mathbb{E}^{!}\{.\}$ is the reduced Palm expectation [22], $\mathcal{I}_o$ and $\mathcal{W}_o$ denote the inter-arrival time between consecutive packets and the waiting time of a generic packet at the o-th device, respectively. As observed, the evaluation of the waiting time is required to evaluate the PAoI. The waiting time depends on, among other parameters, the considered traffic model, queue distribution and network-wide aggregate interference. Throughout the following sections, we provide a spatiotemporal mathematical framework to characterize the PAoI.

4.3 Spatial Macroscopic Analysis

Throughout this section, a novel characterization of the network-wide aggregate interference will be presented for the time and event-triggered traffic models. Such characterization depends on the meta distribution of the network-wide SIR, which will be explained in the following subsection. Afterwards, we consider the time-triggered analysis followed by the event-triggered analysis in subsections 4.3.2 and 4.3.3, respectively.

[3] It is noteworthy to mention that the considered PAoI in this chapter incorporates temporal and spatial averaging.

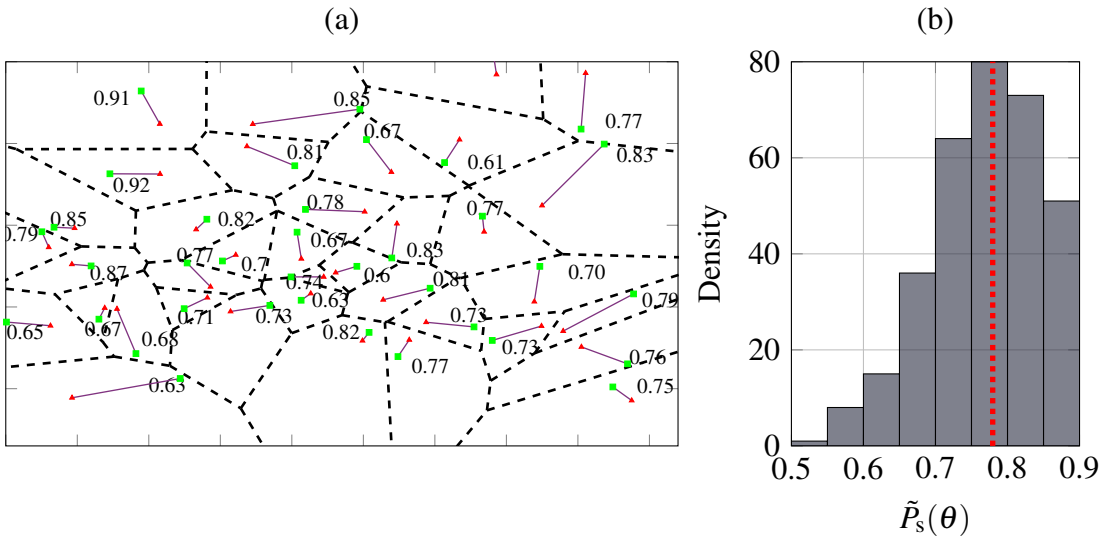


Figure 4.3: For $\theta = 0$ dB (a) network visualization with devices and BSs represented by green squares and red triangles, respectively, whereas the per-device achieved transmission success probability is shown. (b) Transmission success probability histogram with mean depicted via the dashed red line.

4.3.1 The Meta Distribution of the SIR: A Fine-grained Analysis

The meta distribution of the SIR has been firstly proposed for Poisson bipolar networks and downlink cellular networks in [78] and for uplink networks with power control in [97, 79]. Questions such as *What fraction of users in a network can achieve a target link reliability ξ given a required SIR threshold θ* or *How is the transmission success probability of individual devices is distributed with a network realization* are crucial to answer in order to meet the targeted diverse requirements. Moreover, the discrepancies among the users performance, which as an example, can be captured via the 5% user performance, that is the performance level that 95% of the users achieve or exceed, provide network operators with insights about the network performance and the delivered QoS. Quantitative answers for the above questions can be revealed from the meta distribution of the transmission success probability, whereas the traditional (mean) standard transmission success probability analysis provides virtually no information about it. Before delving into the meta distribution of the SIR, we briefly introduce some formal definitions that will help understand the connection between the typical-user and the transmission success probability analysis carried out in Chapter 3 and its meta distribution. Let SIR_o denotes the SIR of the typical receiver, the CCDF of the SIR, which characterizes the events that the typical receiver achieves an SIR above a given decoding threshold θ, is given by

$$P_s(\theta) = \bar{F}_{\mathsf{SIR}}(\theta) = \mathbb{P}^o\{\mathsf{SIR}_o \geq \theta\}, \tag{4.3}$$

where $\mathbb{P}^o\{\cdot\}$ is the Palm probability of the point process, which represents the probability of an event given that the transmitter point process contains a point at some location [22]. To this end, the conditional transmission success probability P_s is evaluated from the transmission success probability, conditioned on a fixed, yet generic spatial realization as follows

$$\tilde{P}_s(\theta) = \mathbb{P}^o\{\mathsf{SIR}_o > \theta | \Psi, \Phi\}. \tag{4.4}$$

It is worth noting that the expression in (4.3) entails randomness resulting from the channel fading, the channel access, and the point processes, whereas the expression in (4.4) is conditioned on the point processes (i.e., transmitter and receiver locations are given). Thus,the conditional transmission success probability unveils the transmission success probability of each link for a certain realization of the network point processes, and it characterizes the SIR performance (reliability) of a given link within the network. It is straightforward to obtain the standard transmission success probability from the conditional transmission success probability via the spatial averaging of $\tilde{P}_{\mathrm{s}}$, i.e., $(P_{\mathrm{s}}(\theta) = \mathbb{E}\{\mathbb{P}\{\mathsf{SIR}_o > \theta|\Psi,\Phi\}\})$. In order to provide a visualized understanding of the difference between (4.3), (4.4) and its meta distribution, we resort to Figure 4.3. The individual link achieved conditional transmission success probabilities P_{s} are plotted next to each transmitter in Figure 4.3(a), whereas the histogram of such values along with the network averaged transmission success probability $P_{\mathrm{s}}(\theta)$ are plotted in Figure 4.3(b). One can observe that the transmission success probability fails to capture the link achieved performance discrepancies, which is rather captured by the conditional transmission success probability. To this end, the meta distribution of the conditional transmission success characterizes the the location dependent success probability $\tilde{P}_{\mathrm{s}}$. Formally, it can be mathematically expressed as [78, 80]

$$\bar{F}_{\tilde{P}_{\mathrm{s}}}(\theta,\xi) = \mathbb{P}^{o}\{P_{\mathrm{s}}(\theta) > \xi|\Psi,\Phi\}, \tag{4.5}$$

where $\xi \in [0,1]$ denotes the percentile of devices within the network that achieves an SIR equals to θ. To this end, $\tilde{P}_{\mathrm{s}}$ can be interpreted as the reliability, i.e., the conditional success probability of the link in consideration given the SIR threshold θ, whereas the meta distribution corresponds to the fraction of links in each network realization that achieve an SIR of θ with reliability at least ξ. In other words, the meta distribution of the transmission success probability evaluates the distribution of achieved probabilities shown in Figure 4.3. Since an exact expression for the meta distribution from (4.5) is mathematically impossible, different approaches have been developed to utilize the moments of the meta distribution to reveal the high order statistics of P_{s} [98]. The b-th moment of $\tilde{P}_{\mathrm{s}}$ with respect to the Palm measure is defined as

$$M_b(\theta) = \mathbb{E}^{o}\left\{\tilde{P}_{\mathrm{s}}{}^{b}\right\}, \quad b \in \mathbb{N}. \tag{4.6}$$

Utilizing this definition and recalling that $\tilde{P}_{\mathrm{s}} \in [0,1]$, we get

$$M_b(\tilde{P}_{\mathrm{s}}) = \int_0^1 \xi^b \mathrm{d}F_{\tilde{P}_{\mathrm{s}}}(\xi) = \int_0^1 b\xi^{b-1}\bar{F}_{\tilde{P}_{\mathrm{s}}}(\xi)\mathrm{d}\xi. \tag{4.7}$$

For the average transmission success probability $P_{\mathrm{s}}(\theta)$, we have $P_{\mathrm{s}}(\theta) = M_1(\theta)$. Different statistical inequalities (e.g., Markov, Chebyshev, and Chernoff) have been proposed in the literature to evaluate the moments as shown in [78, 97]. Throughout this thesis, we adopt the tractable approach presented in [78, 79], which utilizes the beta distribution to approximate the meta distribution by mapping

first and second moments M_1 and M_2 of $\tilde{P}_{\mathrm{s}}$ to the mean and variance of the beta distribution as follows

$$\bar{F}_{\tilde{P}_{\mathrm{s}}}(\theta,\xi) \approx I_{\xi}\left(\frac{M_1\left(M_1-M_2\right)}{\left(M_2-M_1^2\right)}, \frac{\left(1-M_1\right)\left(M_1-M_2\right)}{\left(M_2-M_1^2\right)}\right), \tag{4.8}$$

where $I_{\xi}(a,b) = \int_0^{\xi} t^{a-1}(1-t)^{b-1}\mathrm{d}t$ is the regularized incomplete beta function. Since this approach only uses the first and second moments, it incurs low computational complexity compared to other techniques (e.g., Gil-Pelaez approach) [98]. In addition, the beta approximation method provides very good accuracy to the simulations, as will be shown in Section 4.5. In the next subsection, we will dive into the time-triggered analysis to quantify the meta distribution of the transmission success probability. Hereafter, the subscript $\tilde{P}_{\mathrm{s}}$ in $\bar{F}_{\tilde{P}_{\mathrm{s}}}(\theta,\xi)$ will be dropped for easier readability.

4.3.2 Time-triggered Traffic: Spatial Analysis

Due to uplink association, the devices point process Ψ is a Poisson Voronoi perturbed point process with intensity λ [99, 79, 100, 101]. The periodic time-triggered traffic can be incorporated to the devices point process via the notion of marked point process. That is, let $\tilde{\Psi} = \{x_i, \nu_i\}$ be a marked point process with points $x_i \in \Psi$ and time offset marks ν_i drawn from the uniform distribution $\mathbb{P}\{\nu_i = \tau\} = 1/T$, $\tau \in \{0,1,\cdots,T-1\}$. In addition, let $\tilde{\Psi}_\tau = \{(x_i,\nu_i) \in \tilde{\Psi} : \nu_i = \tau\}$ be the point process where all the devices have identical time offset. Due to the independent and uniform distribution of the time offsets, the intensity of $\tilde{\Psi}_\tau$ for each $\tau \in \{0,1,\cdots,T-1\}$ is $\frac{\lambda}{T}$. Note that, all the devices within the same $\tilde{\Psi}_\tau$ have synchronized packet generation every T time slots, and hence, always interfere together in their first transmission attempt. On the other hand, two devices within different sets $\tilde{\Psi}_{\tau_1}$ and $\tilde{\Psi}_{\tau_2}$ for $\tau_1 \neq \tau_2$ may only interfere together due to retransmissions. A pictorial illustrations of the transmission and mutual interference of four devices in the time-triggered traffic model is shown in Figure 4.4.

Focusing on a fixed, yet arbitrary, spatial realization of Φ and $\tilde{\Psi}$, let $(u_o,\nu_o) \in \tilde{\Psi}$, $b_o = \mathrm{argmin}_{b\in\Phi}||u_o - b||$ and $r_o = ||u_o - b_o||$ define, respectively, the location, time offset, serving BS, and association distance of a randomly selected o-th device, where $||.||$ is the Euclidean norm. For the ease of notation, we define the set $\tilde{\Psi}_{o,\kappa} = \{r_i = ||x_i - b_o|| : (x_i,\nu_i) \in \tilde{\Psi}, \nu_i = \kappa\}$ that contains the relative distances to the serving BS of the o-th device from all devices with time offset $\nu = \kappa$. Due to the adopted time-triggered packet generation and persistent transmission scheme, the SIR exhibits a regular time slot dependent pattern that is repeated every T time slots. In particular, let $\ell \in \mathbb{Z}$ be an integer and $\tau \in \{0,1,\cdots,T-1\}$ be a generic time slot within the duty cycle T, then the SIR of the o-th device at the $(\tau+\ell T)$-th time slot is given by

$$\mathsf{SIR}^{T}_{o,\tau+\ell T} = \frac{\rho h_o r_o^{\eta(1-\varepsilon)}}{\underbrace{\sum_{r_i\in\tilde{\Psi}_{o,\tau}} P_i g_i r_i^{-\eta}}_{\text{determistic for each } \tau} + \underbrace{\sum_{\kappa\neq\tau}\sum_{r_m\in\tilde{\Psi}_{o,\kappa}} 1_{\{a_\kappa^{(m)}(\tau+\ell T)\}} P_m g_m r_m^{-\eta}}_{\text{probabilistic retransmissions}}}, \tag{4.9}$$

where h_o is the intended channel power gain, $a_\kappa^{(m)}(\tau+\ell T)$ is the event that the m-th device with offset $\nu_m = \kappa$ has a non-empty queue at the $(\tau+\ell T)$-th time slot, $1_{\{\cdot\}}$ is an indicator function that is equal to 1 if the event $\{\cdot\}$ is true and zero otherwise, P_i (P_m) and g_i (g_m) denote the i-th (m-th) uplink transmit power and its channel power gain, respectively.

Let $p^{(m)}_{\kappa,\tau} = \mathbb{E}\left\{1_{\{a_\kappa^{(m)}(\tau+\ell T)\}} | m,\kappa,\tau\right\}$ be the probability that the m-th device with time offset $\nu_m = \kappa$ has a non-empty queue at the τ-th time slot within any cycle $\tau+\ell T$. Then the intensity of the interfering

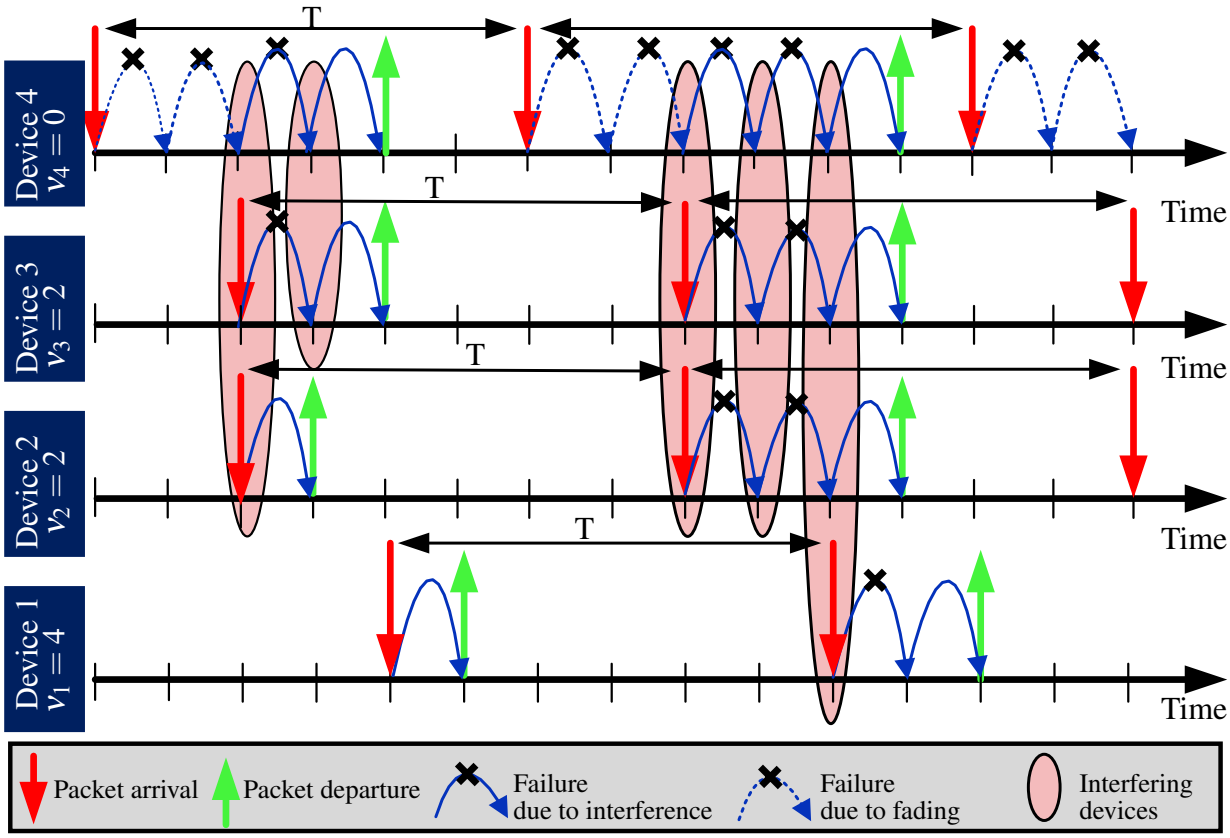


Figure 4.4: Packets generation, departure process, and mutual interference between devices undertime-triggered traffic with duty cycle $T = 6$.

devices within the τ-th time slot is given by

$$\lambda_\tau = \frac{(1+\Theta_\tau)\lambda}{T}, \tag{4.10}$$

where $\Theta_\tau \in [0, T-1]$ is given by $\Theta_\tau = \sum_{\kappa \neq \tau} \mathbb{E}_{\tilde{\Psi}_\kappa}\{p^{(m)}_{\kappa,\tau}\}$. Recalling that the intensity of the devices at each distinct time offset is $\frac{\lambda}{T}$, it is clear that Θ_τ depicts the aggregate percentiles of devices with time offsets $\kappa \neq \tau$ that are active at time slot τ. At the extreme case of flawless transmissions, $\Theta_\tau = 0$ and $\lambda_\tau = \frac{\lambda}{T}$, where only synchronized devices with newly generated packets mutually interfere together. On the other extreme, assuming backlogged queues due to poor transmission success probability, $\Theta_\tau = T-1$, and hence, $\lambda_\tau = \lambda$, where all devices are always active and mutually interfere together. In realistic cases, $0 \leq \Theta_\tau \leq T-1$, which is the focus of the current analysis.

As mentioned earlier, devices are only active when they have non-empty queues. A packet at the queue of a generic o-th device departs from its queue in the time slot $\tau \in \{0, 1, \cdots, T-1\}$ if $\mathbb{P}\{\mathsf{SIR}^T_{o,\tau} > \theta\}$. Since a packet is generated every T slots, it is required that $\mathsf{SIR}^T_{o,\tau}$ exceeds the threshold θ at least once for any of the time slots $\tau \in \{0, 1, \cdots, T-1\}$. Once the generated packet departs and the queue is empty, the device remains idle for the rest of the cycle until the next packet generation (cf. Figure 4.4). Otherwise, the departure rate is not sufficient to cope with the periodic packet generation and packets keep accumulating in the device's queue. Such devices are never idle and are denoted hereafter as unstable devices.

As illustrated from (4.9) and (4.10), the activities of interfering devices, and consequently, $\mathsf{SIR}_{o,\tau}^{T}$ are location and time slot dependent. Due to the fixed realization of the network, the static time offsets, and the periodic generation of packets, each device experiences a location and slot-dependent pattern of $\mathsf{SIR}_{o,\tau+\ell T}^{T}$ for $\tau \in \{0,1,\cdots,T-1\}$ that is repeated every cycle ℓT, $\forall \ell \in \mathbb{Z}$. Despite the randomness in the channel gains and the probabilistic interference of devices with different offsets, the network geometry and the periodic packet generation with static offsets have the dominating effect that highly correlates $\mathsf{SIR}_{o,\tau+\ell T}^{T}$ for each τ across different cycles. Such location and time slot dependence of the SIR yields intractable analysis. Furthermore, there is no known tractable exact analysis for Poisson Voronoi perturbed point process [99, 79, 100, 101]. Hence, for th sake of analytical tractability, we resort to the following two approximations.

Approximation 2. *The location and time slot dependent TSPs $\mathbb{P}\{SIR_{o,\tau}^{T} > \theta\}$ of the BSs in Φ and devices in $\tilde{\Psi}$ are approximated by the location-dependent TSPs $\mathbb{P}\{\hat{SIR}_{o}^{T} > \theta\}$ where each BS in Φ sees a fixed panorama of always active interfering devices constituting a fixed, yet arbitrary, PPP $\hat{\Psi}$ with intensity function*

$$\lambda_T(x) = \frac{(1+\Theta_T)\lambda}{T}(1-e^{-\pi\lambda x^2});\ 0 \leq \Theta_T \leq T-1. \tag{4.11}$$

Remark 3. *Approximation 2 can be regarded as approximating the success probability $\mathbb{P}\{SIR_{o,\tau}^{T} > \theta\}$ of each device across different time slots within the same cycle T by an approximate mean value $\mathbb{P}\{\hat{SIR}_{o}^{T} > \theta\} \approx \mathbb{E}_\tau\{\mathbb{P}\{SIR_o,\tau\}^T > \theta\}$ to alleviate the time-slot dependence. The approximating PPP $\hat{\Psi}$ is assumed to be static to account for the temporal correlations between different cycles, and hence, capture the location dependent performance of the devices. Note that the intensity function in* (4.11) *is sensitive to the effect of unsaturated time-triggered traffic through the parameter $\frac{(1+\Theta_T)\lambda}{T}$. Furthermore,* (4.11) *is also sensitive to the uplink association through the factor $(1-e^{-\pi\lambda x^2})$ [99, 79, 100, 101]. It is worth noting that the validity of such approximation is validated via independent Monte-Carlo simulations in Section 4.5.*

Approximation 3. *The transmission powers of the interfering devices are uncorrelated.*

Remark 4. *Approximation 3 ignores the correlations among the sizes of adjacent Voronoi cells, which lead to correlated transmission powers of devices due to the adopted fractional path-loss inversion power control scheme. Such approximation is widely utilized in the literature to maintain mathematical tractability [99, 79, 100, 101]. We further validate Approximation 3 via independent Monte-Carlo simulations in Section 4.5.*

By virtue of Approximation 2, the time slot indices κ and τ are dropped hereafter. Furthermore, exploiting Approximations 2 and 3 along with the mapping and displacement theorems of the PPP [22], the effect of the power control and path-loss can be incorporated to the intensity function of the approximating PPP. That is, the PPP of the interfering devices $\hat{\Psi}$ can be mapped to a 1-D PPP with unit transmission powers and inverse linear path-loss function. After mapping and displacement, following [79, Lemma 2], the intensity function in (4.11) becomes

$$\tilde{\lambda}_T(\omega) = \frac{2(1+\Theta_T)(\pi\lambda)^{1-\varepsilon}\rho^{\frac{2}{\eta}}}{T\eta\omega^{1-\frac{2}{\eta}}}\gamma\Big(1+\varepsilon,\pi\lambda(\omega\rho)^{\frac{2}{\eta(1-\varepsilon)}}\Big), \tag{4.12}$$

$$\tilde{M}_{b,T} = \int_0^{\infty} \exp\left\{ -z - \Big(\frac{(1+\Theta_T) 2 z^{1-\varepsilon}}{T\eta} \int_{1\{\varepsilon=1\}}^{\infty} y^{\frac{2}{\eta}-1} \Big(1 - \Big(\frac{y}{y+\theta}\Big)^b\Big) \gamma\Big(1+\varepsilon, z y^{\frac{2}{\eta(1-\varepsilon)}}\Big) dy \Big) \right\} dz. \tag{4.15}$$

where $\gamma(a,b) = \int_0^b t^{a-1} e^{-t} \mathrm{d}t$ is the lower incomplete gamma function. Using the intensity function in (4.12), the transmission success probability in the time-triggered traffic model is defined as

$$\begin{aligned} P_{\mathrm{s}} &= \mathbb{P}^{!}\left\{ \mathsf{SIR}_o^T > \theta | \hat{\Psi}, \Phi \right\}, \\ &= \prod_{\omega_i \in \tilde{\Psi}_T} \mathbb{E}^{!}\left\{ \left(\frac{1}{1 + \frac{\theta r_o^{\eta(1-\varepsilon)}}{\rho \omega_i}} \right) \middle| \hat{\Psi}, \Phi \right\}, \end{aligned} \tag{4.13}$$

where $\mathbb{P}^{!}\{\cdot\}$ is the reduced Palm probability ,$\hat{\Psi}_T = \{\omega_i = \frac{r_i}{P_i},\ \forall r_i \in \hat{\Psi}_o\}$, and the set $\hat{\Psi}_o$ contains all relative distances from the approximating PPP $\hat{\Psi}$ to the serving BS of the o-th device. The computation in (4.13) follows from the exponential distribution of h_o and h_i. The meta distribution of the transmission success probability for the time-triggered traffic $F_T(\theta, \xi)$ is approximated as

$$F_T(\theta, \xi) \approx I_{\xi}\left(\frac{M_{1,T}\left(M_{1,T} - M_{2,T}\right)}{\left(M_{2,T} - M_{1,T}^2\right)}, \frac{(1 - M_{1,T})\left(M_{1,T} - M_{2,T}\right)}{\left(M_{2,T} - M_{1,T}^2\right)} \right), \tag{4.14}$$

where $M_{1,T}$ and $M_{2,T}$ are the first and second moments of $\tilde{P}_{\mathrm{s}}$ under the time-triggered traffic model. The approximate moments of $\tilde{P}_{\mathrm{s}}$ for the time-triggered traffic model are given via the following lemma.

Lemma 3. *The moments of the TSPs in an uplink IoT network under time-triggered traffic with duty cycle T are approximated by $M_{b,T} \sim \tilde{M}_{b,T}$ as given in (4.15).*

Proof. See Appendix A.3. ■

Network Categorization

It is observed from Lemma 3 that the macroscopic network-wide aggregate characterization depends on the parameter Θ_T. Before delving into the details of characterizing Θ_T, we first discretize the meta distribution of $\tilde{P}_{\mathrm{s}}$ through uniform network partitioning [39]. Categorizing each devices within the network into a distinctive QoS class is not feasible due to the continuous support of $\tilde{P}_{\mathrm{s}} \in [0,1]$. Consequently, the transmission success probability is quantized into $\tilde{N}$ QoS classes.[4] The network categorization process of the distribution in (4.14) for the n-th class is conducted as follows

$$F_{\tilde{P}_{\mathrm{s}}}(\omega_n) - F_{\tilde{P}_{\mathrm{s}}}(\omega_{n+1}) = \int_{\omega_n}^{\omega_{n+1}} f_{\tilde{P}_{\mathrm{s}}}(\omega) d\omega = \frac{1}{N}, \tag{4.16}$$

[4]The continuous random variable $\tilde{P}_{\mathrm{s}}$ with distribution $f_{\tilde{P}_{\mathrm{s}}}(\omega)$ is quantized to an equally-probable uniform random variable $\tilde{\mathbf{d}} = [\tilde{d}_1\ \tilde{d}_2\ \cdots\ \tilde{d}_{\tilde{N}}]$.

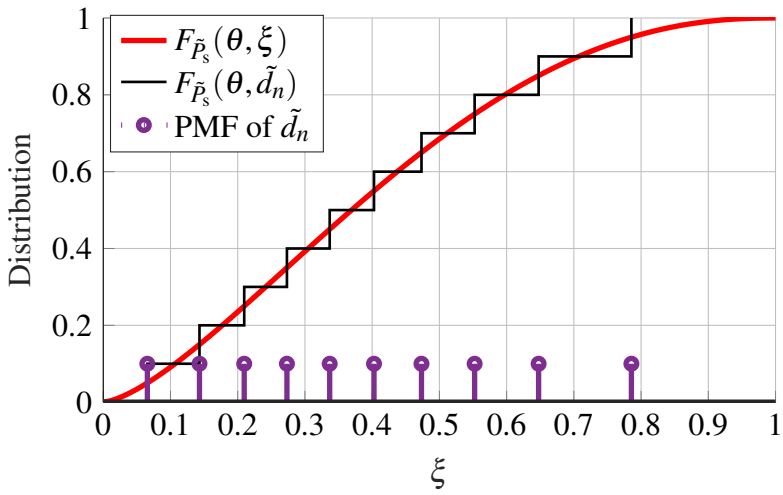


Figure 4.5: Quantized meta distribution for $\tilde{N} = 10$, hypothetical $\Theta_T = 0.5$ and $\theta = 5$ dB.

where $n \in \{1,2,\cdots \tilde{N}\}$. Afterwards, the discrete probability mass function $\tilde{d}_n$ (i.e., $F_{\tilde{P}_s}(\tilde{d}_n) = \frac{1}{\tilde{N}}$) can be evaluated using the bisection method as

$$\int_{\omega_n}^{\tilde{d}_n} f_{\tilde{P}_s}(\omega)d\omega = \int_{\tilde{d}_n}^{\omega_{n+1}} f_{\tilde{P}_s}(\omega)d\omega. \tag{4.17}$$

The computation of $\tilde{d}_n$, $\forall n$ via (4.16) and (4.17) quantizes the meta distribution of $\tilde{P}_s$ into $\tilde{N}$ equiprobable classes as shown in Figure 4.5. The queue's departure rate of a device belonging to the n-class is determined by $\tilde{d}_n$. Now we are in position to characterize the time-triggered traffic parameter Θ_T.

Θ_T Characterization for time-triggered Traffic

As mentioned earlier, for a given set of synchronized devices (i.e., with equal time offset), Θ_T depicts the aggregate percentiles of retransmitting devices from all other distinct time offsets. The first step to characterize Θ_T is to determine the set of always active devices, if any. Particularly, a QoS class that imposes a departure rate less than the packet arrival rate yields always active devices that continuously interfere with other devices irrespective of their relative time offsets. Stable and unstable QoS classes are discriminated via a transmission success probability threshold equal to the packet arrival probability [75]. Consequently, $\mathcal{S}_s = \{\tilde{d}_n \geq \frac{1}{T-1} \mid n \in \{1,2,\cdots,\tilde{N}\}\}$ is the set of stable QoS classes (i.e., devices belonging to this class can empty its queue within the duty cycle T). Visually, Figure 4.4 depicts an example scenario with four devices, each belongs to a given QoS class. Device 1, belonging to the lowest performing class (i.e., one with lowest $\tilde{d}_n$), requires more time slots to successfully transmit its packet. It is noteworthy to mention that transmission failures might occur due to the mutual interference between active devices or fading and path-loss effect. In accordance, $\mathcal{U}_u = \{\tilde{d}_n < \frac{1}{T-1} \mid n \in \{1,2,\cdots,\tilde{N}\}\}$ is the set of unstable QoS classes. Devices belonging to $\mathcal{U}_u$ are not able to empty their queues within the packet generation duty cycle T. Thus, their queues will have infinite size and become unstable. For mathematical tractability, we adopt the following approximation in our work.

Approximation 4. *Queues employed at the devices are QoS-aware but have temporally-independent departures.*

Remark 5. *The temporal correlation of interference is captured by the static QoS class of each device. That is, the departure probability of a device belonging to n-th QoS class remains $\tilde{d}_n$. Once the QoS class is fixed, the departures from the same device across different time slots are considered to be independent due to the randomness introduced by the channel fading and interfering devices activity profiles.*

Let $r_{\mathcal{S}_{sn},k}$ be the probability that a device belonging to a stable n-th QoS is active for k-constitutive time slots. Recall that every device has a new generated packet every T time slots and that stable devices, on average, are able to empty their packets within each duty cycle T. Leveraging the temporal independence between the time slots given by Approximation 4, $r_{\mathcal{S}_{sn},k}$ is evaluated as $r_{\mathcal{S}_{sn},k} = (1-\tilde{d}_n)^k$. The k-consecutive time slots activity due to transmission failures is illustrated in Figure 4.4. The characterization of the spatially averaged aggregate percentiles of retransmitting devices Θ_T for the time-triggered traffic is given in the following lemma.

Lemma 4. *Consider a time-triggered traffic model with duty cycle T. For each set of synchronized devices, the spatially averaged aggregate percentiles of retransmitting devices from all other distinct time offsets is given by*

$$\Theta_T = \frac{1}{\tilde{N}} \sum_{\tau=1}^{T-1} \Big(|\mathcal{U}_u| + \sum_{j=1}^{|\mathcal{S}_s|} r_{\mathcal{S}_{sj},k} \Big), \tag{4.18}$$

where $\mathcal{S}_s = \{\tilde{d}_n \geq \frac{1}{T-1} \mid n \in [1,2,\cdots,\tilde{N}]\}$ and $\mathcal{U}_u = \{\tilde{d}_n < \frac{1}{T-1} \mid n \in (1,\tilde{N})\}$ denote the set of stable and unstable QoS classes, respectively.

Proof. First, the devices belonging to a QoS class that is unstable are always contributing to the aggregate interference. Accordingly, for each distinct time offset, $\frac{|\mathcal{U}_u|}{\tilde{N}}$ percentile of the devices will always be interfering every time slot within the window T. Second, the set of stable devices with time offset k slots away from a given transmission will only interfere if they have encountered k-consecutive transmission failures. Considering all stable QoS classes within each set of devices with distinct time offset, the percentiles of devices that are k-slots active can be characterized as $\frac{\sum_{j=1}^{|\mathcal{S}_s|} r_{\mathcal{S}_{sj},k}}{\tilde{N}}$. Combining the two components (i.e., stable and unstable devices) together and considering all other distinct $T-1$ time offsets within the duty cycle, the lemma is obtained. ■

Iterating through Lemmas 3 and 4, one can evaluate Θ_T and the meta distribution $\bar{F}_T(\theta,\xi)$. In particular, for any feasible initial value of Θ_T, the moments and the TSPs for each QoS class can be calculated via (4.15), (4.16), and (4.17). Then, the value of Θ_T can be updated via (4.18). Repeating such steps, the aforementioned system of equations converges to a unique solution by virtue of fixed point theorem [81]. After convergence to a unique solution, the waiting time, a generic packet spends in the system till its successful transmission, can be evaluated based on the analysis that will be provided in Section 4.4.1.

$$\tilde{M}_{b,E} = \int_0^{\infty} \exp\left\{ -z - \frac{2z^{1-\varepsilon}}{\eta} \int_{1\{\varepsilon=1\}}^{\infty} y^{\frac{2}{\eta}-1} \Big(1 - \Big(\frac{y+\theta\bar{\Theta}_E}{y+\theta}\Big)^b\Big) \gamma\Big(1+\varepsilon, z y^{\frac{2}{\eta(1-\varepsilon)}}\Big) dy \right\} dz. \quad (4.20)$$

4.3.3 Event-triggered Traffic: Spatial Analysis

Following the same methodology that was presented for the time-triggered traffic analysis, the $\mathsf{SIR}_{o,\tau}$ of the o-th device at the τ-th time slot under event-triggered traffic is

$$\mathsf{SIR}^{E}_{o,\tau} = \frac{P_o h_o r_o^{\eta(1-\varepsilon)}}{\sum_{u_i \in \Psi \setminus u_o} 1_{\{a_i\}} P_i h_i r_i^{-\eta}}, \quad (4.19)$$

where a_i is the event that a generic device has a non-empty queue at steady state. Due to the randomized packet generation and departure, the interference in the event-triggered traffic does not exhibit regular repetitive pattern as in the time-triggered case. Hence, (4.19) is independent of the time slot index τ, which will be dropped hereafter. Analogous to Approximations 2 and 3, let $\hat{\Psi}_E$ be a PPP with an intensity function $\lambda_E(x) = \lambda(1 - e^{-\pi\lambda x^2})$ that approximates the interference from $\{\Psi \setminus b_o\}$. Exploiting the mapping and displacement theorems, the interfering PPP seen at a generic BS $b_o \in \Phi$ can be mapped to a 1-D inhomogeneous PPP $\hat{\Psi}_{E,o} = \{s_i = \frac{||x_i - b_o||^{\eta}}{P_i}, \forall x_i \in \hat{\Psi}_E\}$ with the following intensity function

$$\tilde{\lambda}_E(s) = \frac{2(\pi\lambda)^{1-\varepsilon} \rho^{\frac{2}{\eta}}}{\eta s^{1-\frac{2}{\eta}}} \gamma\Big(1+\varepsilon, \pi\lambda(s\rho)^{\frac{2}{\eta(1-\varepsilon)}}\Big). \quad (4.21)$$

Hence, the conditional transmission success probability for the event-triggered traffic model is expressed as

$$\tilde{P}_{\mathrm{s}} = \prod_{s_i \in \tilde{\Psi}_E} \mathbb{E}^{!} \left\{ \left(\frac{\Theta_E}{1 + \frac{a_i \theta r_o^{\eta(1-\varepsilon)}}{\rho s_i}} + \bar{\Theta}_E \right) \middle| \Psi, \Phi \right\}, \quad (4.22)$$

where Θ_E denotes the spatially averaged active probability (i.e., the probability that a device has a non-empty queue) under the event-triggered traffic at steady state. Different from its time-triggered counterpart in (4.13), the transmission success probability for the event-triggered in (4.22) depicts the varying set of interfering devices through the probability of empty queue [78, 79]. In particular, the higher probability of empty queues, the less correlated interference across time slots, and vice versa. The approximations of $\tilde{P}_{\mathrm{s}}$ moments for the event-triggered traffic model are given via the following lemma.

Lemma 5. *The moments of the transmission success probabilities in uplink network with event-triggered traffic model with arrival probability α are approximated by $M_{b,E} \sim \tilde{M}_{b,E}$ as given in (4.20), where Θ_E is the spatially averaged active probability.*

Proof. The proof follows similar steps as Lemma 3. ■

After the computation of the approximated moments under event-triggered traffic $\tilde{M}_{b,E},;\ b=\{1,2\}$, the meta distribution $F_E(\theta,\xi)$ is evaluated based on (4.14) after plugging the computed $\tilde{M}_{b,E}$. In addition, the network categorization procedure is carried out in a similar way as explained in Section 4.3.2.

Θ_E Characterization for event-triggered Traffic

The spatially averaged interfering intensity for the event-triggered traffic is equivalent to the percentage of devices which have packets to be transmitted in their respective queues at steady state. To this end, the idle probability of the n-th QoS class $x_{0,n}$ captures such activity. Resorting to the mean field theory, Θ_E is computed by averaging over the $\tilde{N}$ classes temporal idle probabilities as

$$\Theta_E = 1 - \frac{1}{\tilde{N}} \sum_{n=1}^{\tilde{N}} x_{0,n}. \tag{4.23}$$

It is clear that to evaluate $F_E(\theta,\xi)$, one needs first to compute $x_{0,n}$. Such inter-dependency between the network-wide aggregate interference and the queues characterization highlights the cross-relation between the microscopic and macroscopic scales in the network. The characterization of $x_{0,n}$, which is required to evaluate the waiting times and the PAoI will be discussed in Section 4.4.2.

4.4 Temporal Microscopic Analysis

The mathematical model for the microscopic scale (i.e., queue evolution) will be presented in this section. As discussed earlier, the device's location-dependency is captured via its departure probability (i.e. QoS class dependent), which remains unchanged over long time horizon. In this chapter, a geometric process is adopted to model the packets departure from each device. It is important to note that the geometric departure is an approximation that capitalizes on the negligible temporal correlation of the departure probabilities once the location-dependent QoS class is determined as mentioned in Approximation 4. Focusing on the time-triggered and event-triggered traffic in Sub-sections 4.4.1 and 4.4.2, respectively, tractable expressions to characterize the temporal evolution and the spatiotemporal PAoI will be presented.

4.4.1 Time-triggered Traffic: Temporal Analysis

We utilize a degenerate PH type distribution to mimic the time-triggered traffic generation at every device.[5] In particular, the utilized PH type distribution works as a deterministic counter that generates a packet every T time slots. A pictorial illustration of the DTMC with deterministic arrival of packets every $T=4$ time slots is shown in Figure 4.6(a). The PH type distribution is defined as an absorbing Markov chain [29], where in the context of time-triggered, absorption implies packet arrival. The utilized PH type distribution is represented by the tuple $(\boldsymbol{\rho},\mathbf{S})$, where $\boldsymbol{\rho}\in\mathbb{R}^{1\times T}$ is the initialization vector and $\mathbf{S}\in\mathbb{R}^{T\times T}$ is the sub-stochastic transient matrix [74]. In addition, the absorption vector $\mathbf{s}$ of the PH type distribution

[5] In probability theory, a degenerate distribution is a distribution that supports a single deterministic outcome. That is, a random variable with zero variance boils down to a deterministic value and its distribution is said to be a degenerate distribution [102].

can be evaluated as $\mathbf{s} = 1_T - \mathbf{S}1_T$. The matrix $\mathbf{S}$ is constructed to count exactly T time slots between two successive packet generations. Accordingly, there is no randomness in the packet generation process and the transition probabilities between the states equal 1. In order to mimic the periodic generation of a packet, $\mathbf{S}$ is formulated as

$$\mathbf{S} = \begin{bmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \cdots & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}, \tag{4.24}$$

and $\rho = [1\ \mathbf{0}_{T-1}]$. Based on the proposed PH type distribution for the time-triggered arrival process, we model the temporal interactions via an PH/Geo/1 queue [29]. The departure process is captured via a geometric process due to the adoption of Approximation 4. Figure 4.6(a) shows the proposed DTMC model for the time-triggered traffic, where the vertical and horizontal transitions depict transitions between levels and phases, respectively. Utilizing the previously mentioned PH type structure, one can provide a tractable model that captures the queueing temporal dynamics in the form of a QBD process [74]. The queue transitions for a device within the n-th class are captured through the QBD characterized via the following probability transition matrix

$$\mathbf{P}_n = \begin{bmatrix} \mathbf{B} & \mathbf{C} & & & \\ \mathbf{E}_n & \mathbf{A}_{1,n} & \mathbf{A}_{0,n} & & \\ & \mathbf{A}_{2,n} & \mathbf{A}_{1.n} & \mathbf{A}_{0,n} & \\ & & \ddots & \ddots & \ddots \end{bmatrix}, \tag{4.25}$$

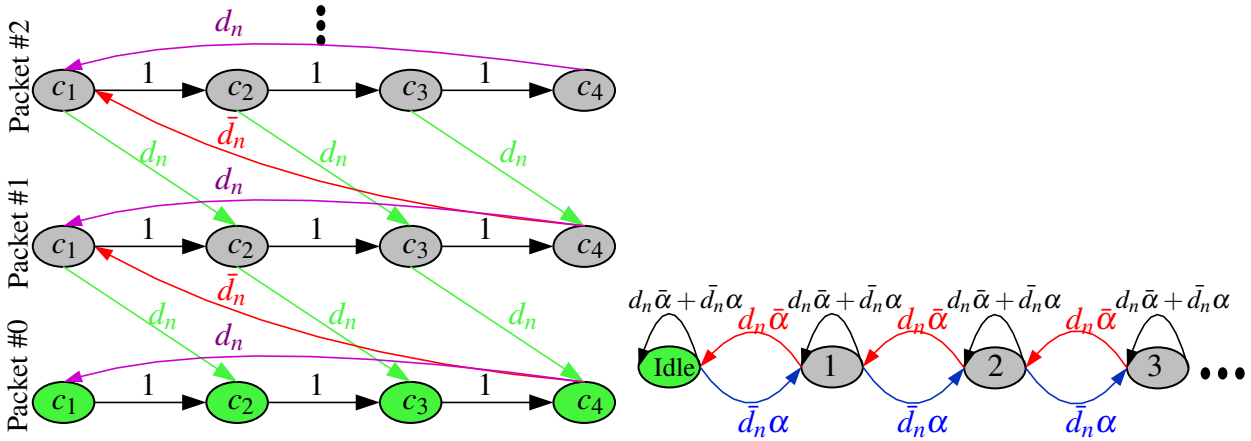
where $\mathbf{B} = \mathbf{S}, \mathbf{C} = \mathbf{s}\rho$ and $\mathbf{E}_n = \tilde{d}_n\mathbf{S} \in \mathbb{R}^{T\times T}$ are the boundary sub-stochastic matrices. In addition, $\mathbf{A}_{2,n} = \tilde{d}_n\mathbf{S}, \mathbf{A}_{0,n} = \bar{d}_n\mathbf{s}\rho$, and $\mathbf{A}_{1,n} = \tilde{d}_n\mathbf{s}\rho + \bar{d}_n\mathbf{S}$, where $\mathbf{A}_{2,n}$, $\mathbf{A}_{0,n}$, and $\mathbf{A}_{1,n} \in \mathbb{R}^{T\times T}$ represent the sub-stochastic matrices that capture the transition down a level, up a level, and in a fixed level within the QBD, respectively. In addition, $\mathbf{A}_{2,n}$, $\mathbf{A}_{1,n}$, and $\mathbf{A}_{0,n}$ are represented via the green, violet, and red arrows in Figure 4.6(a). As mentioned in the previous section, for the DTMC in (4.25) to be stable, the following condition must be satisfied [75]

$$\tilde{d}_n \geq \frac{1}{T-1}. \tag{4.26}$$

Let $\mathbf{x}_i = [\mathbf{x}_{n,0}\ \mathbf{x}_{n,1}\ \mathbf{x}_{n,2}\ \cdots]$ be the steady state probability vector, where $\mathbf{x}_{n,i} = [x_{n,i,1}\ x_{n,i,2}\ \cdots\ x_{n,i,T}]$, where $x_{n,i,k}$ is the probability that a device that belongs to the n-th QoS class has i packets and is in the k-th arrival state. In this context, the idle probability of device in the n-th class is evaluated as

$$x_{0,n} = \sum_{k=1}^{T} x_{0,n,k}. \tag{4.27}$$

Through this chapter, a mathematically tractable solution is sought to address the aforementioned DTMC employed at each device. Markov chains with QBD structure can be solved via utilizing the MAM [74],[29]. Based on the state transition matrix defined in (4.25), the following lemma derives the steady state distribution of the queues temporal evolution.



(a) TT traffic where $T = 4$ and C_i is the i-th counter.

(b) ET traffic highlighting the numbers of packets in the queue.

Figure 4.6: DTMCs modeling the temporal evolution. Green states represent idle states.

Lemma 6. *The steady state distribution of a device belonging to the n-th QoS class based on the state transition matrix* $\mathbf{P}_n$ *under time-triggered traffic with duty cycle* T *is given by*

$$\mathbf{x}_{i,n} = \begin{cases} \mathbf{x}_{i,n}\mathbf{B} + \mathbf{x}_{i+1,n}\mathbf{E}_n, & i = 0, \\ \mathbf{x}_{i-1,n}\mathbf{C} + \mathbf{x}_{i,n}(\mathbf{A}_{1,n} + \mathbf{R}_n\mathbf{A}_{2,n}), & i = 1, \\ \mathbf{x}_{i-1,n}\mathbf{R}_n, & i > 1, \end{cases} \tag{4.28}$$

where $\mathbf{R}_n$ *is the MAM matrix and is given by* $\mathbf{R}_n = \mathbf{A}_{0.n}(\mathbf{I}_T - \mathbf{A}_{1,n} - \omega\mathbf{A}_{02,n})^{-1}$. *The term* ω *is the spectral radius of* $\mathbf{R}$*, which can be evaluated by solving for* z *in* $z = \mathbf{s}(\mathbf{I}_T - \mathbf{A}_{1,n} - z\mathbf{A}_{2,n})^{-1}\mathbf{I}_T$. *In addition, (4.28) must satisfy the normalization* $\mathbf{x}_{0,n}\mathbf{1}_T + \mathbf{x}_{1,n}(\mathbf{I}_T - \mathbf{R}_n)^{-1}]\mathbf{1}_T = 1$.

Proof. See Appendix A.4. ■

Once the queue distribution is characterized, one can proceed with evaluating the waiting time distribution of a generic packet residing in a queue, which is the major component in computing the PAoI as explained in Section 4.2.3. Let $\mathcal{W}_n^T$ be the waiting time of a generic packet at a device belonging to the n-th QoS class in the queue under the time-triggered traffic and $\mathcal{W}_n^{m,T} = \mathbb{P}\{\mathcal{W}_n^T = m\}$. Also, let $\mathbf{q}_i^n = [q_{i,1}^n\ q_{i,2}^n\ \cdots\ q_{i,T}^n]$, where $q_{i,j}^n$ is the probability that an incoming packet at a device belonging to the n-th class will find i packets waiting and the next packet arrival has phase j. In accordance, $\mathbf{q}_i^n$ is evaluated as [29]

$$\mathbf{q}_l^n = \begin{cases} \sigma\left(\mathbf{x}_{i,n}\mathbf{s}\rho + \mathbf{x}_{i+1,n}\mathbf{s}\rho\tilde{d}_n\right), & l = 0 \\ \sigma\left(\mathbf{x}_{i,n}\mathbf{s}\rho\bar{d}_n + \mathbf{x}_{i+1,n}\mathbf{s}\rho\tilde{d}_n\right), & l \geq 1, \end{cases} \tag{4.29}$$

where $\sigma = \rho(\mathbf{I}_T - \mathbf{S})^{-1}1$. To this end, the waiting time distribution is calculated as

$$\mathcal{W}_n^{m,T} = \begin{cases} \mathbf{q}_0^n\mathbf{1}_T, & m = 0, \\ \sum_{v=1}^{i}\mathbf{q}_v^n\mathbf{1}_T\binom{i-1}{v-1}b^v(1-b)^{i-v}, & m \geq 1. \end{cases} \tag{4.30}$$

After computing the waiting time distribution for the time-triggered traffic model, one can proceed with the PAoI evaluation as presented in the following theorem.

Theorem 3. *The spatially averaged PAoI under time-triggered traffic with duty cycle T is given by*

$$\mathbb{E}\{\Delta_p|\Psi,\Phi\} = T + \frac{1}{N}\Big(\sum_{n=1}^{N}\sum_{j=0}^{\infty} j\mathcal{W}_n^{j,T}\Big). \tag{4.31}$$

Proof. The theorem is proven by plugging (4.30) into (5.20) and noting that $\mathbb{E}^!\Big\{\mathcal{I}_o|\Psi,\Phi\Big\} = T$. ■

4.4.2 Event-triggered Traffic: Temporal Analysis

Similar to (2.5) and leveraging the flexibility offered by the PH type distribution, as discussed in Chapter 2, a wide range of traffic models can be incorporated to model the event-triggered traffic [103, 29]. Following the construction of the PH type distribution to mimic the required event-triggered traffic model, the probability transition matrix $\mathbf{P}_n$ is constructed following (4.25). Afterwards, the MAM is utilized to compute the steady state probability vector $\mathbf{x}_n$ as presented in Lemma 6. For the special case of Bernoulli-based event-triggered traffic, with inter-slot packet arrival probability denoted by α, the per device DTMC is illustrated in Figure4.6(b) and is characterized in the following corollary.[6].

Corollary 1. *The probability transition matrix of a device in the n-th QoS class $\mathbf{P}_n$ under Bernoulli-based event-triggered traffic is given by*

$$\mathbf{P}_n = \begin{bmatrix} \bar{\alpha} & \alpha & & & \\ \bar{\alpha}\tilde{d}_n & \alpha\tilde{d}_n + \bar{\alpha}\bar{d}_n & \alpha\bar{d}_n & & \\ & \bar{\alpha}\tilde{d}_n & \alpha\tilde{d}_n + \bar{\alpha}\bar{d}_n & \alpha\bar{d}_n & \\ & & \ddots & \ddots & \ddots \end{bmatrix}. \tag{4.32}$$

In addition, the queue distribution is

$$x_{i,n} = R_n^i \frac{x_{0,n}}{\bar{d}_n}, \text{ where } R_n = \frac{\alpha\bar{d}_n}{\bar{\alpha}\tilde{d}_n}, \text{ and } x_{0,n} = \frac{\tilde{d}_n - \alpha}{\tilde{d}_n}, \tag{4.33}$$

where $x_{i,n}$ is the probability that a device belonging to the n-th class has i packets residing in its queue.

Proof. The probability transition matrix $\mathbf{P}_n$ is evaluated by setting $\mathbf{S} = \bar{\alpha}$ in (4.25), while the queue distribution follows that of a Geo/Geo/1 DTMC [29]. ■

Remark 6. *The DTMC presented in Corollary (1) is stable if the inequality $\frac{\alpha}{\tilde{d}_n} < 1$ is satisfied. For unstable DTMCs, the idle probability is naturally 0.*

Once the queue distribution is characterized, one can proceed with evaluating the spatially averaged idle probability Θ_E and the waiting time distribution of a generic packet within the considered queue. Similar to the time-triggered traffic, an inter-dependency exists between the network-wide aggregate

[6]The Bernoulli model for event-triggered traffic was adopted in this chapter due to its mathematical convenience [35, 36, 34, 104, 30] and practical significance as reported in the literature [84, 4, 105, 106].

Algorithm 2 Computation of $F_E(\theta,\xi)$

procedure $(\alpha,\varepsilon,\theta,N,\varphi)$
  initialize Θ_E.
  while $||{\Theta_E}^k - {\Theta_E}^{k-1}|| \geq \varphi$ **do**
    Compute the moments $\tilde{M}_{b,E}$ from Lemma 5.
    Evaluate $F_E(\theta,\xi)$ based on (4.14).
    Compute $\tilde{d}_n, \forall n = \{1,2,\cdots,N\}$ from discretized $F_E(\theta,\xi)$ based on (4.16)&(4.17).
    for $n = \{1,2,\cdots,N\}$ **do**
      if $\alpha < \tilde{d}_n$ **then** ▷ Stability condition
        Compute $x_{0,n}$.
      else
        Set $x_{0,n} = 0$.
      end if
    end for
    Compute Θ_E based on (4.23).
    Increment k.
  end while
  return $F_E(\theta,\xi)$
end procedure

interference (i.e., $F_E(\theta,\alpha)$) and the queues characterization (i.e., Θ_E). To solve such inter-dependency, Algorithm 2 is presented which provides a unique solution by virtue of fixed point theorem. To this end, the evaluation of the spatially averaged PAoI for the event-triggered traffic follows Theorem 3 with two minor modifications. First, the dimension of the vector (i.e., $\mathbf{1}_T$) in (4.30) equals the number of PH type distribution transient states m. Second, the average inter-arrival times $\mathbb{E}^!\Big\{\mathcal{I}_o|\Psi,\Phi\Big\}$ equals m/α, $\sum_{i=1}^{m}\frac{\rho_i}{\alpha_i}$ and $1/\alpha$ for the negative binomial, mixed geometric and Bernoulli distributions, respectively. For the special case of Bernoulli-based event-triggered traffic, the PAoI is characterized by the following corollary,

Corollary 2. *The spatially averaged PAoI for the Bernoulli-based event-triggered traffic with inter-slot arrival probability α is*

$$\mathbb{E}\{\Delta_p|\Psi,\Phi\} = \frac{1}{\alpha} + \frac{1}{N}\Big(\sum_{n=1}^{N}\sum_{j=0}^{\infty} j\mathcal{W}_n^{j,E}\Big), \tag{4.34}$$

where $\mathcal{W}_n^{j,E}$ is the waiting time of a generic packet of a device that belongs to the n-th QoS class and is evaluated as [29]

$$\mathcal{W}_n^{j,E} = \begin{cases} \frac{\tilde{d}_n-\alpha}{\tilde{d}_n}, & j=0, \\ \sum_{v=1}^{i} x_{v,n}\binom{i-1}{v-1}{\tilde{d}_n}^v(1-\tilde{d}_n)^{i-v}, & j\geq 1. \end{cases} \tag{4.35}$$

Proof. The corollary is realized based on (5.20) and setting $\mathbb{E}^!\Big\{\mathcal{I}_o|\Psi,\Phi\Big\} = \frac{1}{\alpha}$. ■

4.5 Numerical Results

In this section, different numerical insights are presented for the purpose of (a) validating the proposed mathematical framework for the two traffic models, (b) characterizing the information freshness within a large scale uplink IoT network, and (c) highlighting the influence of the system parameters on the network's stability. First, discussion of the simulation environment is presented to establish a clear understanding of the simulation framework.

4.5.1 Simulation Methodology

The established simulation framework involves deployment of BSs and devices as discussed in Section 4.2. Ergodicity is ensured via microscopic averaging, in which the temporal steady state statistics of the queues at each device are collected. The simulation area is 10×10 km^2 with a wrapped-around boundaries to eliminate the effect of the boundary devices within the network. Discretized, synchronized, and time-slotted system is considered, where during each time slot (i.e., microscopic run), independent channel gains are instantiated and packets are generated deterministically or probabilistically, depending on the traffic model. At the start of the simulation, for the time-triggered traffic, all the devices within the network are assigned an i.i.d. transmission offset ν_i from the distribution $f_\nu(\tau) = \frac{1}{T}$ for $\tau \in \{0,1,\cdots,T-1\}$, which depicts the time index of a packet generation event. A new packet is generated periodically following $\nu_o + \ell T$, $\forall \ell = 1,2,\cdots$. For the event-triggered traffic, a new packet is generated at each device every time slot with the probability α. Every device with packets residing in its queue attempts the communication of such packets with its serving BS based on a FCFS strategy. A packet is dropped from its queue if the realized uplink SIR at the serving BS is greater than the detection threshold θ.

To ensure a steady state operation of the queues, each queue's occupancy at each device is monitored. For initialization, all queues at the devices are initiated as being empty and then simulation runs for a sufficiently large number of time slots till steady-state is realized. Let $\hat{x}_0^t$ denotes the average idle steady state probability across all the devices within the network for the t-th iteration. Mathematically, the steady state behavior is reached once $||\hat{x}_0^k - \hat{x}_0^{k-1}|| < \varphi$, where φ is some predefined tolerance (e.g. 10^{-4}). Once steady state is reached, all temporal statistics are then collected based on adequately large number of microscopic realizations (e.g., 10000). Unless otherwise stated, we consider the following parameters: $\eta = 4$, $\rho = -90$ dBm, $\varepsilon = 1$, $T = 8$ and $\alpha = 0.125$.

4.5.2 Time and Event-triggered Results Discussion

In Figure 4.7, we consider the framework verification via the meta distribution of the transmission success probability for the time-triggered, Bernoulli-based and mixed geometric-based event-triggered traffic models with different traffic loads and detection threshold values. First, for all the considered traffic models, one can observe a close match between the simulation and the proposed analytical framework, which confirms the accuracy and flexibility of the proposed mathematical model and shows how the inter-dependency between the network-wide aggregate interference and the queues temporal evolution is well captured. For low values of θ, the devices are able to empty their queues and become idle. This leads to a lower network-wide aggregate interference, and thus increased percentile of devices achieving a given

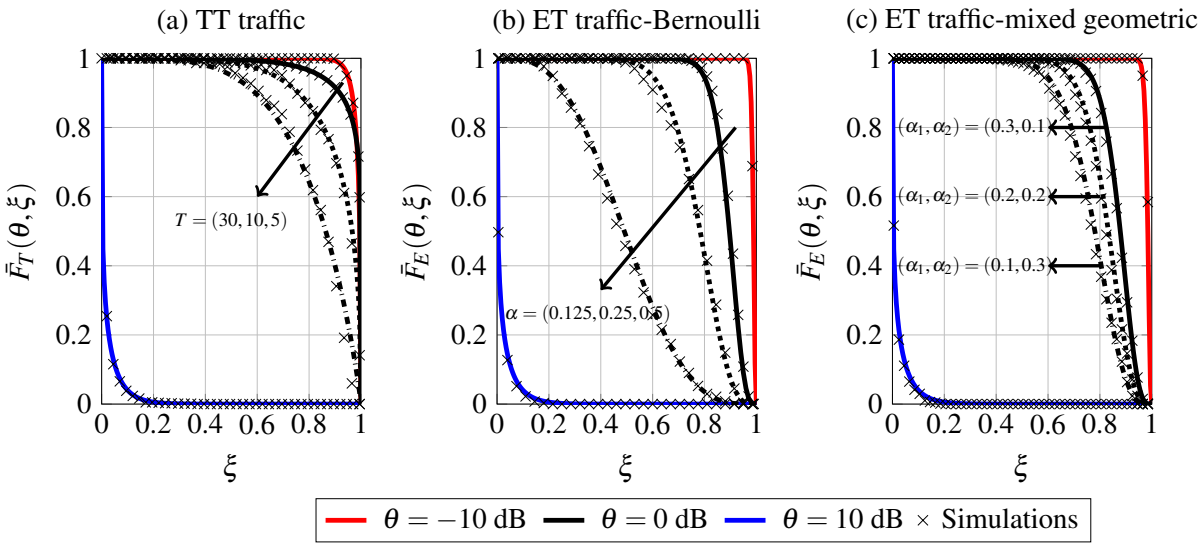


Figure 4.7: Framework verification for (a) time-triggered (b) Bernoulli-based event-triggered (c) mixed geometric-based event-triggered traffic.

reliability ξ. As θ increases, the probability of successful transmission attempts for a generic device decreases, which aggravates the aggregate network interference. Consequently, more devices are active within the network and the achieved reliability to meet the targeted θ decreases. Figure 4.7(a) presents the time-triggered traffic patters for different values of cyclic duration. It is observed that as T decreases, the percentile of active devices increases within the network. Decreasing T increases the packet generation rate, shortens the time required to dispatch generated packets, and increases the number of synchronized devices. Accordingly, the network interference increases, which deteriorates the transmission success probabilities. Such a consequential effect of increased traffic load affects the percentile of devices within the network to achieve a given transmission success probability, as illustrated via the meta distribution. In addition, Figure 4.7(b) presents the Bernoulli-based event-triggered traffic model with different arrival probabilities. Similar to time-triggered case, as α increases, the percentile of active devices increases within the network, thus affecting the reliability to achieve a targeted decoding threshold θ. More insights comparing the time-triggered to the event-triggered models will be discussed in Figure 4.11. Finally, Figure 4.7(c) shows the mixed geometric-based event-triggered traffic model with two traffic arrival classes α_1 and α_2 and initialization vector $\boldsymbol{\rho} = [0.3\ 0.7]$. It is observed that the effect of the initialization vector values on the network-wide performance. Recall that ρ_1 (ρ_2) denotes the probability of a given device gets assigned to α_1 (α_2). Based on the selected values of $\boldsymbol{\rho}$, we observe the performance gap between the different traffic loads for $\theta = 0$ dB, which is attributed to the effect of higher arrival rates of packets captured via α_1 and α_2. After the verification of the proposed spatiotemporal framework, throughout the rest of this section, we will focus on the Bernoulli-based event-triggered traffic model in order to benchmark it against the time-triggered traffic.

Figure 4.8 shows the spatially averaged PAoI along with average waiting time for versus the cycle duration T for the time-triggered traffic model. As explained in Section 4.2.3, the PAoI is sensitive to the inter-arrival and system waiting times of a randomly selected packet within the queue. First we investigate the effect of θ. As θ increases, packets transmission success is subjected to a more stringent requirement on the achieved SIR. This leads to increased retransmissions, thus, increasing the mutual interference

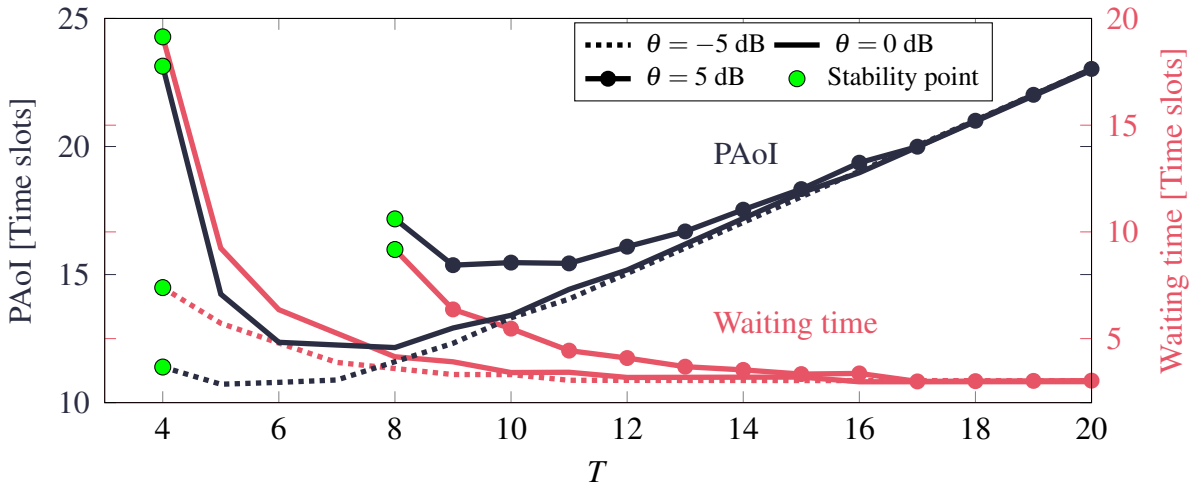


Figure 4.8: PAoI (left) and average waiting time (right) for time-triggered traffic with increasing duty cycle T and different θ.

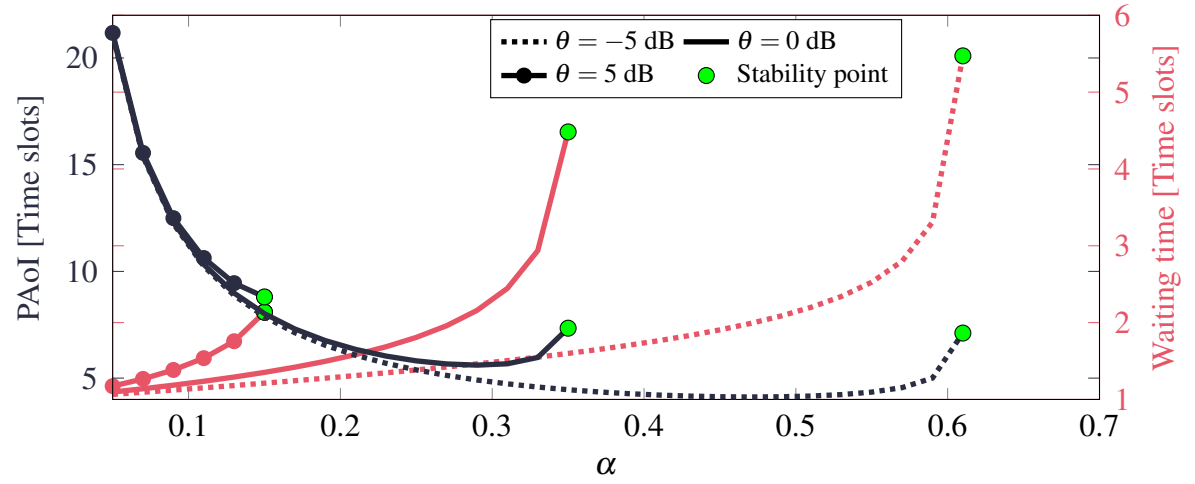


Figure 4.9: PAoI (left) and average waiting time (right) for event-triggered traffic with increasing arrival probability (α) and θ.

due to lower idle probabilities. The increased mutual interference hinders the successful departure of the packets from their respective queues and lead to queue instability in some devices, yielding instability (i.e., infinite waiting times and PAoI). The figure also shows the effect of the cycle times. For high values of T, the large inter-arrival times is the dominant factor, yielding high values of PAoI, while the waiting time is low. Low values of waiting times are the result of having sufficient time to transmit a residing packet, before the event of a new packet arrival. As T decreases, the waiting times dominates, yielding an increase in the PAoI till point of queue instability, as indicated by the stability point. Consequently, adopting a time-triggered traffic with duty cycle $T < 4$ results in an unstable system and infinite PAoI. The effect of θ on the stability frontiers can be explained in a similar fashion to that of Figure 4.7, where increasing θ diminishes the stability region due to the increased network-wide aggregate interference. While reduced traffic arrivals reliefs network interference and reduces delay, it is not the case for AoI because it prolongs the updates duty cycle. Hence, there is an optimal duty cycle that minimizes the PAoI by balancing the trade-off between frequency of updates and the aggregate network interference.

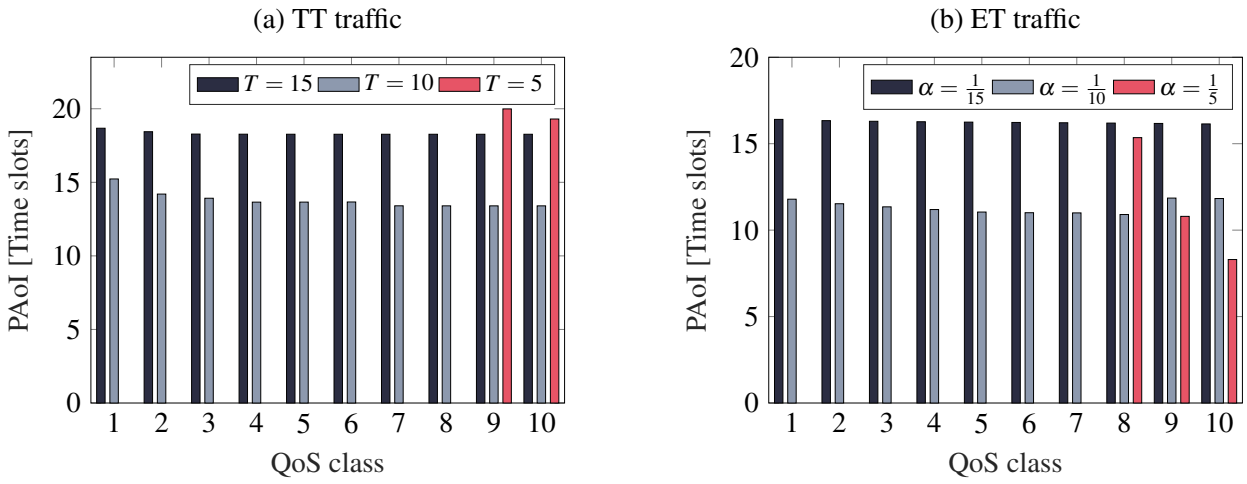


Figure 4.10: PAoI for $N = 10$ QoS classes and $\theta = 5$ dB.

Similar to the time-triggered traffic, Figure 4.9 shows the PAoI along with average waiting time for the event-triggered traffic with increasing arrival probability α. For low values of α, the inter-arrival component dominates, yielding high values of PAoI, while the waiting time is low. For low arrival probabilities, the network-wide aggregate interference is low, yielding higher probabilities for a packet to be successfully transmitted without large number of retransmissions. However, as α increases, the waiting times dominates, yielding an increase in the PAoI till point of queue instability, as indicated by the stability point.

Figure 4.10 presents the per-QoS class PAoI among the different QoS classes within the network. The shown classes are sorted in an ascending order with respect to $\tilde{d}_n$ (i.e., a device belonging to class i is spatially located closer to its serving BS compared to a device belonging to class j, such that $i > j$). The time-triggered and event-triggered traffic models are shown in Figure 4.10(a) and Figure 4.10(b), respectively. For $T = 15$ ($\alpha = 1/15$), the inter-arrival times dominates the PAoI, leading to a nearly-constant PAoI over all the classes. The location-dependency is more clear as $T(\alpha)$ decreases (increases). for $\alpha = 0.15$ and $\alpha = 0.25$. Consequently, classes with lower indices experience large PAoI due to their larger waiting times (i.e., effect of the location dependency captured via the meta distribution). For large traffic load (i.e., $T = 5$ ($\alpha = 1/5$)), all except last two and three classes are unstable, for the time-triggered and event-triggered traffic models, respectively. As mentioned earlier, unstable queues results in infinite PAoI.

Next we assess the time-triggered and event-triggered traffic based on their PAoI, meta distribution, and TSPs for three different traffic loads. Figure 4.11 presents different performance comparisons between the two traffic models. First, Figure 4.11(a) shows the PAoI as a function of increasing detection threshold θ. It is observed that the event-triggered traffic provides lower PAoI for all the considered set of traffic loads. Although, it was shown in [50] that periodic packet generation minimizes the age for the FCFS queues, considering the network-wide aggregate interference into the age analysis provides another perspective. As mentioned in Section 4.3, the time-triggered traffic model imposes a spatial and temporal correlation between the devices. In particular, each device sees the same set of active (i.e., interfering) devices in each transmission cycle T. Such correlation is alleviated in the event-triggered traffic model, in which the activity profiles are diversified among different time slots. This performance

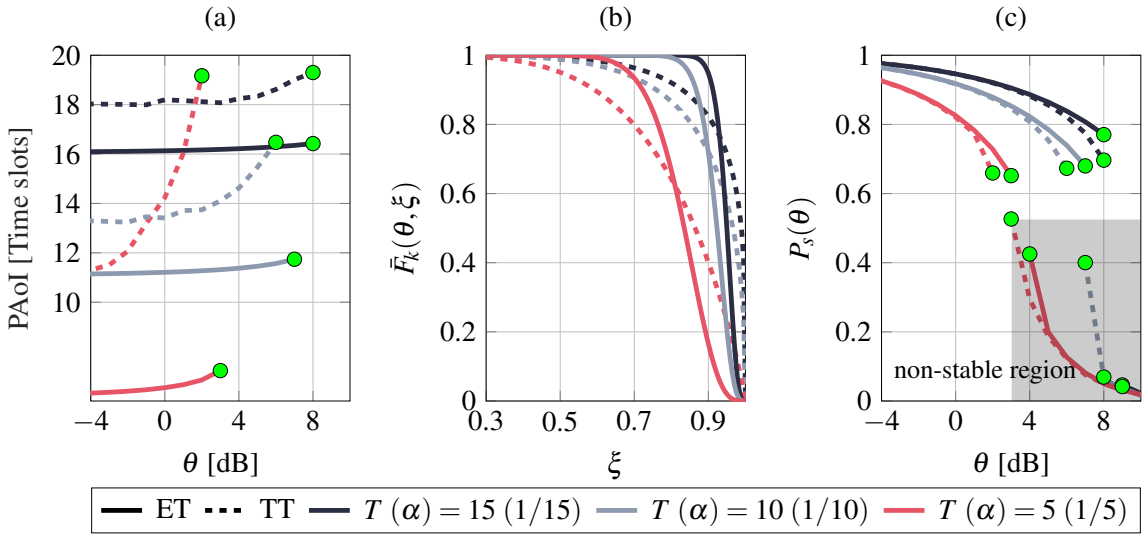


Figure 4.11: TT and event-triggered traffic models comparison based on (a) PAoI (b) meta distribution of the TSP for $\theta = 1$ (c) TSP.

gap between the time-triggered and event-triggered traffic is larger for low duty cycles (or high arrival probabilities), due to the stronger interference correlation in such scenarios. As the activity profiles are more relaxed (i.e., $T(\alpha)$ increases (decreases)), the gap between the two traffic models decreases. Next, Figure 4.11(b) presents the meta distribution for the considered traffic loads. As the traffic load increases for the two traffic models, the percentile of devices achieving a given reliability (i.e., ξ) decreases as explained in Figure 4.7. In addition, one can observe the discrepancies between the time-triggered and event-triggered traffic considering the similar traffic load. Such discrepancies are hardly captured by the spatially averages P_s, which emphasizes the importance of the meta distribution as shown in 4.11(c). In addition, a sharper transition in the meta distribution implies less location-dependent performance (i.e., less temporal interference correlation) and that all devices tend to operate as a typical device. Due to the aforementioned explained correlation between the active devices, the time-triggered traffic provides lower TSPs for all the considered traffic loads. The stability point, depicted by green circles, represent the point at which the queues are unstable. Any operation beyond such a point yields in operating in the non-stable region.

Finally, Figure 4.12 presents the Pareto frontiers for the arrival intensity of the event-triggered and time-triggered traffic with the detection threshold over the N QoS classes. Pareto frontiers define regions where the queues are guaranteed to be operating within a stable region. First, Figure 4.12(a) shows the relation between the arrival probability and the detection threshold θ for the existing five QoS classes. Due to retransmissions, a higher θ implies lower idle probability, and hence, higher aggregate network interference allowing lower values of α to ensure stability. In addition, due to the favorable spatial locations of the higher QoS classes compared to the lower ones, the Pareto frontiers for those higher classes are covering a larger set of (θ, α) values. Similarly, Figure 4.12(b) presents the Pareto frontiers between θ and the cyclic time T. The curves explanation follows that of the event-triggered traffic, since T represents the arrival events, comparable to α.

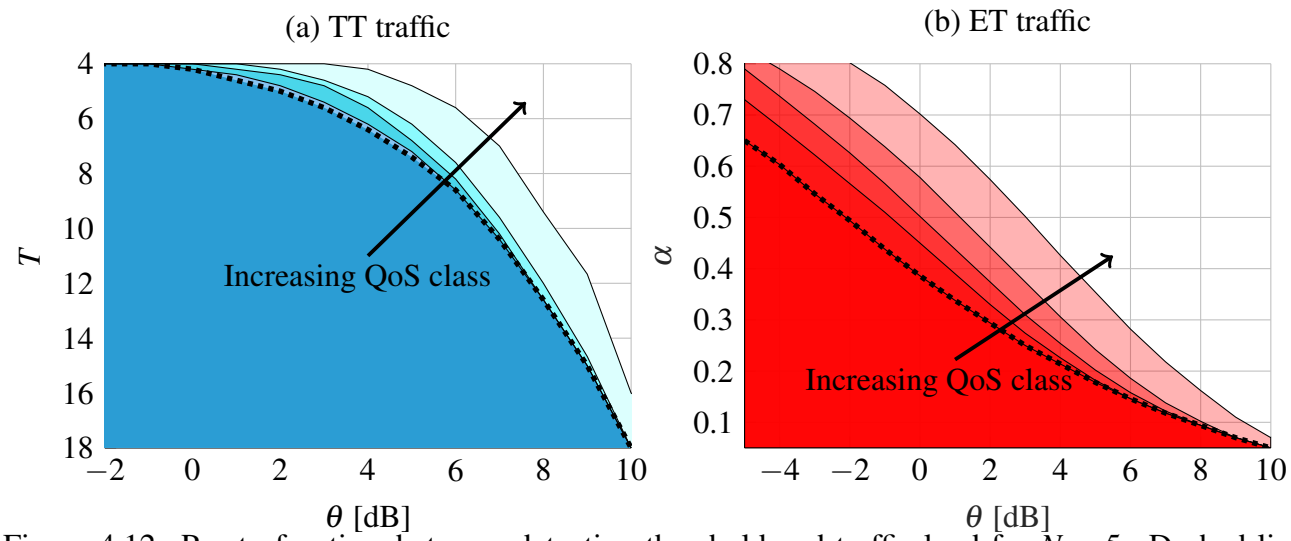


Figure 4.12: Pareto frontiers between detection threshold and traffic load for $N = 5$. Dashed lines represent the spatially averaged frontiers.

4.6 Conclusion

This chapter presents a mathematical spatiotemporal framework to characterize the PAoI in IoT uplink networks for time-triggered and event-triggered traffic. First, we leverage tools from stochastic geometry to analyze the location-dependent performance of the network under the two traffic models. Expressions for the network-wide aggregate interference are presented in the context of the location-aware meta distribution. Furthermore, we analyze the inter-dependency between the aggregate network wide-interference and the queues evolution at each device. Additionally, PH type distribution is leveraged to model the periodic and random traffic generation for the time-triggered and event-triggered traffic, respectively. In summary, both the time-triggered and event-triggered traffic models can be captured via unified queueing analysis. However, in large scale networks, both traffic patterns lead to different mutual interactions between the coexisting IoT devices, which is captured via the stochastic geometry analysis. To this end, the two traffic models are verified and compared in terms of transmission success probabilistic, delay, and PAoI. Simulation results are presented to validate the proposed framework. The results unveil the counter-intuitive lower PAoI of event-triggered traffic over the time-triggered traffic, which is due to the higher temporal interference correlations of the time-triggered traffic. In addition, the stability frontiers coupling the network's traffic load and decoding threshold are presented and their effect on the PAoI are discussed.

Chapter 5

Multi-access Edge Computing and Low Latency Communication

5.1 Introduction

Throughout this and the following chapter, we focus on the design of computation-aided wireless networks, in particular, leveraging the advents of MEC deployment. As mentioned earlier, the efficient design of 5G mobile networks is driven by the need to support the dynamic proliferation of several vertical market segments. Such verticals are nevertheless faced with new challenges such as, resources dimensioning, densification impact, network-wide mutual interference and others [7]. As a result, network design and resources dimensioning need to consider not only the communication aspect, but also the computation one and the interconnection between the two. Throughout this chapter, we investigate the following points:

- In a heterogeneous network deployment, where heterogeneity among different tiers is present in the communication and computation resources, questions related to the impact of the adopted cell association criterion on the experienced latency is our main focus. We propose in Section 5.2 a new, device-cell association metric, which takes into consideration the proximity of MEC resources to a device and investigate its effect on the device's end-to-end experienced latency.
- For a vehicular network, we showcase in Section 5.3 the latency and information freshness gains achieved from MEC deployment in comparison with a conventional, remote cloud-based cellular architecture. In order to carry out such evaluation, we model the different network components. Furthermore, the effect of different system parameters as well as the latency bottlenecks are discussed.

5.2 MEC-aware Cell Association for Future Networks

As mentioned earlier, the advent of MEC brings up the need to efficiently plan and dimension network deployment by means of jointly exploiting the available radio and processing resources [57, 58]. From this standpoint, advanced cell association of devices can play a key role for 5G systems [107]. Focusing

on a heterogeneous network , this Section proposes a comparison between state-of-the-art (i.e., radio-only) and MEC-aware cell association rules, taking the scenario of uplink task offloading as a use case. Numerical evaluations show that the proposed cell association rule provides nearly 60% latency reduction, as compared to its standard, radio-exclusive counterpart.

5.2.1 Background and Contributions

The applied rule for user-cell association plays a key role towards efficiently exploiting the entire set of resources [108]. Nevertheless, current mobile systems have been planned and deployed so far by following traditional paradigms of network planning (e.g., based on radio-only coverage). Unfortunately, this approach is not sustainable anymore, as current cell association rules completely discard the aforementioned availability of *processing* resources at the network's edge, hence, they fail to constitute cost-effective and flexible solutions for QoS provisioning [107]. To the best of our knowledge, current technical literature mostly sheds light on the problem of optimally allocating radio and computation resources to already connected devices, inherently assuming *conventional* cell association, where the device is connected to its serving BS based on the maximum reference signal received power (RSRP) rule.

In details, authors in [109] investigate task offloading in a multi-cell scenario, where they show an enhancement achieved by offloading to multiple BSs via benefiting from prior knowledge of radio statistics. In [110], the problem of radio and computation resource allocation over connected devices is investigated under time division multiple access (TDMA) and frequency division multiple access (FDMA) schemes. The authors optimize the joint allocation and show the achieved gains, as compared to a baseline round-robin scheme. An analytical framework that optimizes the offloading decision under task deadlines for a single device is presented in [111]. For multi-devices deployment, [112] studies the problem of joint radio and processing power allocation under an optimization framework, where the task completion time is minimized subject to energy consumption constraints. It is, thus, evident that none of the above works question the effectiveness of the applied cell association rule. With regards to the design of a cell association rule driven by performance requirements, in [113], a cross-layer, device matching problem was studied for a cloud-RAN. In this work, the authors proposed a joint matching scheme between the devices, cloud-RAN components and MEC hosts, aiming at meeting a task completion deadline at the device side. Nevertheless, this work did not exploit the multi-tier resource disparity expected in a heterogeneous network as well as reveal the practicality of the association procedure from a signaling overhead viewpoint. Given the above described research development and identified gaps, this section presents the following:

- Focusing on a MEC-enabled heterogeneous network , we introduce a new, device-cell association metric, which evaluates the proximity of MEC resources to a device.
- To highlight the benefits of the proposed association rule, we introduce a one-way packet latency budget metric, which is the latency consisting of the radio transmission time of an input packet between the device and the connected BS in the uplink, along with the execution time of a given task at a MEC host.

- We conduct numerical evaluation to compare the proposed association rule to the conventional RSRP rule, in terms of the packet latency budget performance for various inter-tier resource disparities, as well as for different network deployment densities.

5.2.2 Heterogeneous Communication and Computation Model

Throughout this section, a $\mathcal{K}$-tier cellular network is studied, where the BSs and devices locations are spatially randomized following a PPP deployment. According to this model, the locations of the BSs of the i-th tier are modeled through a homogeneous PPP $\Phi_i = \{x_i\}, i = 1,2,\cdots,\mathcal{K}$ of density λ_i. It should be noted that the $\mathcal{K}$ PPPs are mutually independent. On the other hand, the device positions are modeled via a different homogeneous PPP, Ψ, of density μ. Due to the network's heterogeneity, different tiers are distinguished by the transmit power, P_i, of their BSs, their spatial density, λ_i, and the total processing power, $\mathcal{C}_i$, of a MEC host co-located with an i-th tier BS. Cross-tier radio resource disparity can be adjusted by defining the ratio of the transmit powers of two consecutive tiers, $\mathcal{G}^{\mathrm{r}}{}_i$, (i.e., $\mathcal{G}^{\mathrm{r}}{}_i = \frac{P_i}{P_{i+1}} > 1$), as well as the ratio of processing powers of their MEC hosts, $\mathcal{G}^{\mathrm{c}}{}_i$, (i.e., $\mathcal{G}^{\mathrm{c}}{}_i = \frac{\mathcal{C}_i}{\mathcal{C}_{i+1}} > 1$). It should be noted that the mentioned ratios are always greater than 1 as a tier $i \in \{1,\cdots,\mathcal{K}\}$ is assumed to be overlaid with tiers of lower transmit power and processing capabilities. Note that $\mathcal{G}^{\mathrm{r}}{}_{\mathcal{K}} = \mathcal{G}^{\mathrm{c}}{}_{\mathcal{K}} = 1$.

The path-loss between a given device and its serving BS is modeled as inversely proportional to the distance r with a given path-loss exponent denoted by η, of common value for all tiers. Small-scale fading h, g is assumed to be Rayleigh distributed with unit average power and the fast fading effects are assumed non-correlated among the various links. Additionally, each device employs a fixed transmit power, P, which is greater than its serving BS sensitivity threshold. The target BS belonging to the i-th tier is assumed to be placed at the origin [114], thus, for uplink communication, the measured SINR at the k-th device associated to an BS in the i-th tier is

$$\mathsf{SINR}_{k,i} = \frac{P h_{k,i} r_{k,i}^{-\eta}}{I_{k,i} + \sigma^2}, \tag{5.1}$$

where σ^2 is the noise power and $I_{k,i}$ is the interference generated by other active devices as $I_k = \sum_{y_j \in \Psi \setminus y_k} P g_{j,i} ||y_j - z_i||^{-\eta}$, such that y_j is the location of an interfering device, $g_{j,i}$ is the channel power gain between the interfering device and the serving BS, $||.||$ is the Euclidean norm, and z_i is the device of interest serving BS's location. Finally, orthogonal channel allocation is assumed to avoid intra-cell interference.

As mentioned earlier, low latency access to cloud infrastructure is foreseen as a critical feature of 5G systems [115]. As a result, the experienced one-way packet latency budget, denoted by $\mathcal{E}$, at the device side during task offloading will be our metric of interest throughout this section. In Figure 5.1, the end-to-end experienced packet latency budget is provided. First, $\mathcal{T}^{\mathrm{dev}}$ represents the time needed for application initiation and packet generation at the device side, followed by time intervals for data transmission and task execution at the MEC host, denoted by $\mathcal{T}^{\mathrm{radio}}$ and $\mathcal{T}^{\mathrm{exc}}$, respectively. Throughout this section, $\mathcal{T}^{\mathrm{dev}}$ is implicitly modeled through T^{radio}, via random generation of packets, whereas, the back-haul, web and remote processing latencies, denoted by $T^{\mathrm{BH+CN}}$, T^{Web} and T^{Proc} respectively, are left to be addressed in the following Section. It is also assumed that the BSs and their corresponding

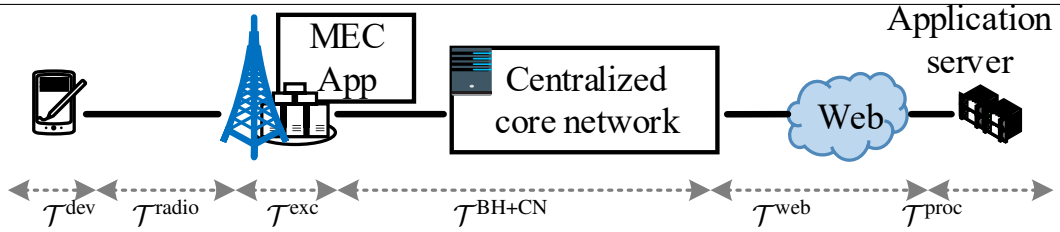


Figure 5.1: End-to-end latency overview.

MEC hosts are physically located at the same node and that all deployed devices concurrently offload their tasks to their chosen MEC host. As modeled in [110], the packet latency budget for the k-th device associated to an BS in the i-th tier is calculated as follows

$$\mathcal{E}_{k,i} = \mathcal{T}^{\text{radio}}_{k,i} + \mathcal{T}^{\text{exc}}_{k,i}, \tag{5.2}$$

where $\mathcal{T}^{\text{radio}}_{k,i}$ and $\mathcal{T}^{\text{exc}}_{k,i}$ stand for the radio propagation time and the task execution time at the MEC host, respectively. The radio propagation latency represents the time needed for a given packet of size of l_k bits to arrive at the serving BS [116], thus can be calculated as

$$\mathcal{T}^{\text{radio}}_{k,i} = \frac{l_k}{r_{k,i}} = \frac{l_k}{\mathcal{B}_{k,i}\log_2(1+\mathsf{SINR}_{k,i})}, \tag{5.3}$$

where $r_{k,i}$ is the achievable rate of the k-th device and $\mathcal{B}_{k,i}$ represents the bandwidth allocated to device k when served by an BS in the i-th tier. On the other hand, the execution time can be computed as

$$\mathcal{T}^{\text{exc}}_{k,i} = \frac{l_k f_k}{\mathcal{Y}_{k,i}\mathcal{C}_i}, \tag{5.4}$$

where f_k, measured in cycles/bit, is the number of processing operations per input bit for the task to be offloaded by the k-th device and $\mathcal{Y}_{k,i}$ represents the fraction of the total processing power of a tier-i MEC host dedicated to the k-th device. Throughout this chapter, we assume equal per-user allocation of radio bandwidth and computation (MEC) resources [117], as the design of a more sophisticated resource allocation scheme can be left to future work. Thus, for a given BS belonging to the i-th tier, the number of associated devices, which is obtained by means of applying a cell association rule, will determine the portion of bandwidth and computation resources dedicated to each connected device. In what follows, we present in detail the investigated association rules.

5.2.3 Computational Proximity Cell Association

Conventionally, considering a single tier of communication, the downlink RSRP rule determines the cell to which the device will be connected for both downlink and uplink communication. Nevertheless, employing such a connectivity criterion in a highly heterogeneous network consisting of multi-tier BSs with diverse capabilities leads to load imbalance among the different tiers [118, 99]. Moreover, the uplink cell association is achieved based on an BS proximity criterion, hence, leading to the minimum path-loss experienced by the device. A depiction of such a miss-match between the downlink and uplink coverage regions for two-tiers network is shown in Fig 5.2, where it is observed the load imbalance between the

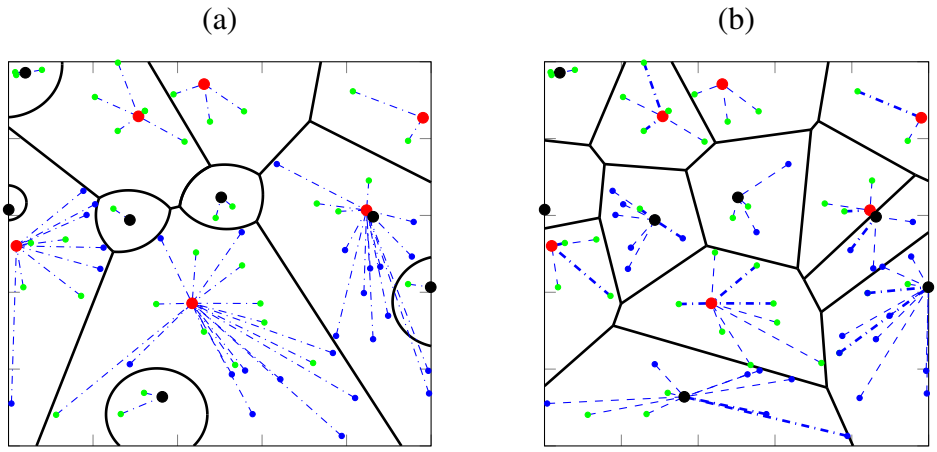


Figure 5.2: Coverage regions visualization for a) downlink and b) uplink communications for $\mathcal{G}^{\mathrm{r}}{}_1 = 40$. Macro and micro BSs are depicted via red and black circles, respectively. Devices with same and different serving BSs in downlink and uplink are shown by green and blue circles, respectively.

two tiers as well as the number of devices with decoupled access. As a result, downlink-uplink decoupled access has been proposed as a disruptive solution for an enhanced network performance, mainly giving the devices the flexibility to associate with the BS that provides the minimum path-loss, when it comes to uplink communication [119, 120]. To this end, we choose to revisit the aforementioned rules and propose a new, MEC-aware cell association rule, that aims at minimizing the execution time at the MEC host, along with ensuring connectivity to the closest BS. This is motivated through the added degree of freedom resulting from the MEC deployment, and questioning the optimality of the conventional, maximum downlink RSRP-based association rule, when it comes to the latency experienced by a device in a heterogeneous network . A mathematical representation of the association problem can be formulated as follows

$$x_i = \underset{x\in\Phi_i}{\arg\max}(\iota_i||x-y||^{-\eta}), \forall i = 1,2,\cdots,\mathcal{K}, \;\; x_o = \underset{x\in x_i : i=1,\cdots\mathcal{K}}{\arg\max}(\iota_i||x-y||^{-\eta}), \tag{5.5}$$

where x_i is the serving BS index and $\iota_i, i = 1,\cdots,\mathcal{K}$ represents a biasing factor for the i-th tier imposed to the devices, and y is the device's location. According to the RSRP cell association rule, a device is served by the BS providing it with the maximum RSRP in the downlink. This is equivalent to setting ι_i to be equal to P_i in (5.5). In a heterogeneous network with large radio disparity (i.e., $\mathcal{G}^{\mathrm{r}}{}_i >> 1$), the adoption of this rule leads to an imbalanced load among the multiple tiers and, as a result, to limited radio performance, since most of the devices will be associated to BSs of high transmit power. This problem is well-known and multiple solutions have been proposed, such as load-aware optimization [115] and cell range extension [121]. In order to quantify the number of devices associated with a tier-i BS, the

association probability of a given device to an BS of the i-th tier is calculated as [122]

$$\mathcal{A}_i^{\text{RSRP}} = \frac{\lambda_i}{\Xi_i^{\text{RSRP}}}, \tag{5.6}$$

$$\Xi_i^{\text{RSRP}} = P_i^{\frac{-2}{\eta}} \sum_{j=1}^{K} \lambda_j P_j^{\frac{2}{\eta}}. \tag{5.7}$$

Consequently, the average number of associated devices to an BS of the i-th tier, termed as $\hat{N}_i^{\text{RSRP}}$, will affect the experienced packet latency budget per device, as the amount of bandwidth and processing resources allocated per device is inversely proportional to the achieved packet latency budget. Mathematically, $\hat{N}_i^{\text{RSRP}}$ can be evaluated as

$$\hat{N}_i^{\text{RSRP}} = \frac{\mathcal{A}_i^{\text{RSRP}} \lambda_u}{\lambda_i} = \frac{\lambda_u}{\Xi_i^{\text{RSRP}}}. \tag{5.8}$$

Assuming equal resource allocation among the devices connected to an BS, the bandwidth and processing resources allocated to the k-th device associated to an BS of the i-th tier will be equal to

$$\mathcal{B}_{k,i} = \frac{\mathcal{B}}{\hat{N}_i^{\text{RSRP}}}, \;\; \mathcal{Y}_{k,i} = \frac{1}{\hat{N}_i^{\text{RSRP}}}, \tag{5.9}$$

where $\mathcal{B}$ represents the total bandwidth allocated to tier $i, i = 1, \cdots, \mathcal{K}$. At this stage, we can introduce the proposed MEC-aware cell association rule, according to which the serving BS is the one of the maximum *computation proximity*. In this context, computation proximity refers to the existence of a processing power source in the vicinity of a device of limited computation capabilities that chooses to offload a demanding task to this source. Such resources, as defined in Section 6.2.2 ($\mathcal{C}_i, \forall i = 1, \cdots, \mathcal{K}$) can be the same for all the tiers, thus, resulting in a homogeneous network from a MEC perspective, or can be varying across the tiers, resulting in a MEC heterogeneous network , thus, affecting the task offloading latency experienced by a device. As observed from (5.2), the overall packet latency budget is jointly affected by the proximity to the connected BS (i.e., radio part - $\mathcal{T}_{k,i}^{\text{radio}}$) as well as by the available processing power (i.e., MEC part - $\mathcal{T}_{k,i}^{\text{exc}}$). Our aim is to consider both resource domains through introducing a new association rule for uplink communication, by setting the bias factors ι_i as functions of the available computation resources (i.e., $\iota_i = \mathcal{C}_i$). As a consequence, the association probabilities and the average numbers of connected devices can be computed easily by replacing P_i by $\mathcal{C}_i$ and computing $\mathcal{A}_i^{\text{MEC}}$, Ξ_i^{MEC} and $\hat{N}_i^{\text{MEC}}$, accordingly.

Referring to the spatial network deployment, a critical factor affecting the performance of the proposed device-cell association rule is the ratio of radio/ MEC cross-tier disparities, which is defined as

$$\mathcal{X}_i = \frac{\mathcal{G}^{\text{r}}{}_i}{\mathcal{G}^{\text{c}}{}_i}. \tag{5.10}$$

In order to visualize the influence of parameter $\mathcal{X}$ on device connectivity, focusing on a two-tier network, Figure 5.3 presents a zoomed overview of a network realization, where devices are connected to their serving BSs/MEC hosts via the two discussed rules. One can observe that, assuming a large value of

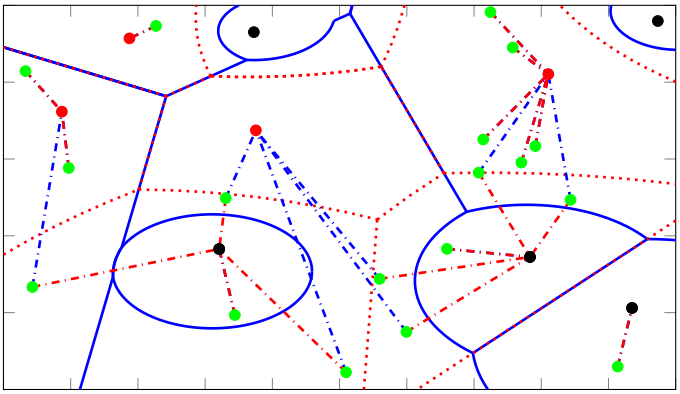


Figure 5.3: A spatial realization of a two-tier network consisting of macro, micro BSs and devices, represented by red, black and green circles, with $\mathcal{G}^{\mathrm{r}}{}_1 = 40$, $\mathcal{G}^{\mathrm{c}} = 2$, $\mathcal{X}_1 = 20$. The blue dashed-dotted lines represent device connectivity following the maximum downlink RSRP association rule, while, the red dashed lines represent device connectivity based on the proposed computation proximity-based rule.

parameter $\mathcal{X}_1$, for a fair number of devices, the maximum downlink RSRP association rule indicates a node for connectivity which is different from the one obtained by applying the proposed computation proximity-based association rule. This occurs because large cross-tier radio/ MEC disparities lead towards quite dissimilar radio/ MEC coverage areas. Such an observation paves the way towards a different insight on the network planning process, taking into account the available computation resources together with the radio transmission capabilities, since both of them directly affect the packet latency budget experienced by a given device, when the latter wishes to offload a demanding processing task to a MEC host. In the following subsection, we present various simulation results, highlighting key messages regarding the studied association rules, the role of cross-tier parameter disparities, as well as the effect of deployment densities on the achieved packet latency budget.

5.2.4 RSRP and MEC-aware Association Results

Our objective throughout this section is to provide insight on the packet latency budget improvements when applying the new proposed MEC-aware association rule, by means of numerical evaluation. A two-tier heterogeneous network is investigated, where the k-th device generates a random packet of size of l_k bits that is modeled as a uniform random variable taking values between $l_{\min}$ and $l_{\max}$. Additionally, the number of processing operations per input bit, f_k, is uniformly distributed, as well, between values $f_{\min}$ and $f_{\max}$. The amount of dedicated bandwidth and computation resources that each BS assigns to its associated devices is computed based on the applied association rules. A summary of the adopted simulation parameters is provided in Table 6.2, where the parameter values are fixed throughout the section, unless otherwise stated. It should be noted that, as a two-tier heterogeneous network is considered, the subscript of parameter $\mathcal{X}_1$ will be dropped for the sake of simplicity.

In Figure 5.4(a), the CCDF of the packet latency budget is shown for the two discussed association rules and for different values of $\mathcal{X}$. As previously explained, when $\mathcal{X}$ varies away from the value of one, the radio and MEC coverage areas become more dissimilar, hence, resulting in a selection divergence of the associating BS/ MEC host by a device. It is observed that, for values of $\mathcal{X}$ greater than one ($\mathcal{X} = 2$),

Table 5.1: Simulation parameters for MEC-aware cell association evaluations

Parameter	value
Number of tiers ($\mathcal{K}$)	2
BSs spatial intensity (λ_1, λ_2)	(0.5, 3) BSs/km
BS transmit power (P_1, P_2)	(46, 30) dBm
Deployment area ($\mathcal{A}_{\text{dep}}$)	10 km^2
Devices spatial intensity (μ)	30 devices/km
Device transmit power (P_{dev})	23 dBm
Noise power (σ^2)	-90 dBm
Packet size range ($l_{\min}, l_{\max}$)	(100, 300) kbits
Processing operations range ($f_{\min}, f_{\max}$)	(500, 1500) cycles/bit
Bandwidth ($\mathcal{B}$)	10 MHz
Path-loss exponent (η)	4
Number of realizations	10000

the proposed computation proximity association rule (denoted by MEC) provides a lower probability to violate a given packet latency budget threshold as compared to the maximum RSRP rule (denoted by RSRP), with nearly 60% packet latency budget reduction for the 50-th percentile of devices. This occurs due to the enhanced balance between the proximity and available computation resources at the MEC node. On the other hand, as $\mathcal{X}$ is lower than one ($\mathcal{X}$ = 0.5), the performance is turned over, as the RSRP rule provides a lower experienced packet latency budget of the same latency reduction. Consequently, we observe that having the two association metrics at hand, an adaptive, deployment-dependent cell association procedure can be envisioned, in order to fully capture the radio and MEC resource disparities for packet latency budget minimization. Under that framework, the device is ought to only acquire knowledge of the radio and MEC disparities of the heterogeneous network , in order to decide upon which association rule to consider. For the case of $\mathcal{X} = 1$, since the corresponding coverage areas obtained by the two rules will fully overlap, the experienced packet latency budget performance will be identical for the two rules.

With the aim of observing the effect of deployment density on the experienced packet latency budget, Figure 5.4(b) depicts the probability of violating a target of 0.4 seconds for an increasing ratio of micro-over-macro BS spatial densities when $\mathcal{X} = 2$. We observe a nearly constant association-based outage reduction in favor of the proposed MEC-aware association rule, similar to the latency reduction observed in Figure 5.4(a). The decreasing slope of the two curves is expected as the number of micro BSs over a unit area increases. This is due to the increasing probability for a device to be associated with a closer node, thus leading to lower delay values.

Finally, in Figure 5.5, the percentage of devices for which the maximum downlink RSRP and the proposed MEC-aware cell association rules provide different connectivity recommendations, is illustrated, as a function of the value of parameter $\mathcal{X}$. As anticipated, for the increase of cross-tier disparity between the radio and MEC capabilities (i.e. $\mathcal{X} \neq 1$), the two coverage areas become highly divergent, thus, leading to a higher probability of a device being present in this disjoint region (e.g., nearly 40% of UEs will reach different decisions upon associating to an BS/ MEC node for large disparities of $\mathcal{X} = 0.01$ or $\mathcal{X} = 80$). On the contrary, for the $\mathcal{X} = 1$ case, the radio and MEC coverage areas will be identical,

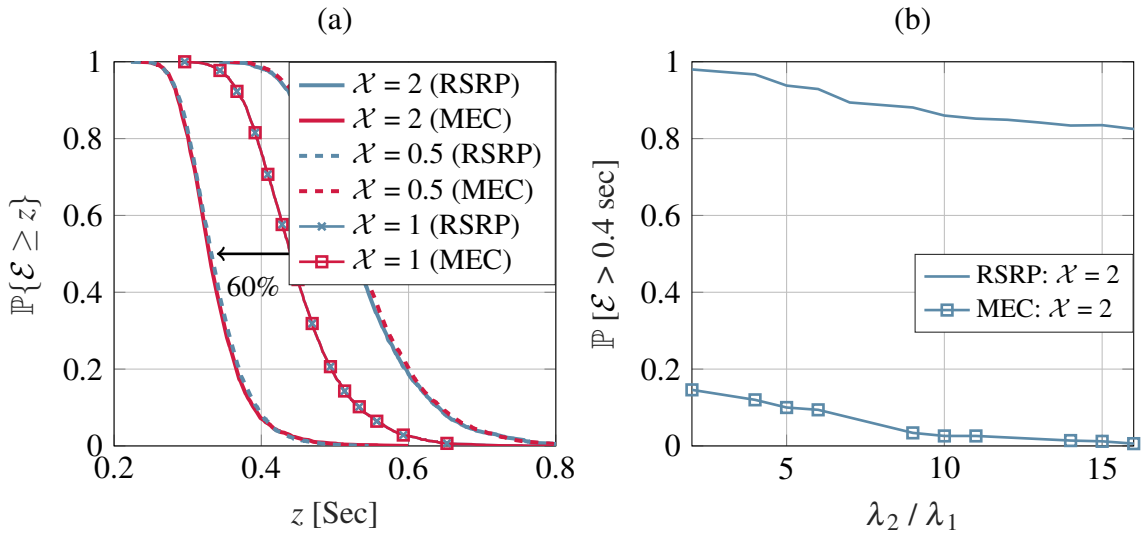


Figure 5.4: (a) CCDF of the packet latency budget under different radio and MEC resource disparities (b) Probability of a target packet latency budget as a function of the ratio of BS/ MEC deployment densities.

hence, the application of the two investigated association rules will provide the same preference for uplink connectivity.

Finally, it is evident the latency-related gains resulting from the proposed MEC-based cell association criterion in heterogeneous networks. Throughout the following section, we will continue showcasing MEC-deployment gains, via investigating a latency-critical vehicular use case in the automotive vertical.

5.3 MEC-Assisted End-to-End Latency Evaluations

Considering the automotive sector, different cellular-V2X (C-V2X) use cases have been identified by the industrial and research world, referring to infotainment, automated driving and road safety [123]. A common characteristic of these use cases is the need to exploit collective awareness of the road environment towards satisfying performance requirements. One of these requirements is the end-to-end latency when, for instance, vulnerable road users (VRUs) inform vehicles about their status (e.g., location) and activity, assisted by the cellular network [124]. We argue that, when it comes to safety-critical use cases, such as the one of VRU, additional metrics, such as the AoI can be more insightful when compared to traditional latency metrics. In particular, the impact of the packet inter-arrival time on the timeliness of VRU messages arriving at nearby vehicles can be directly assessed by exploiting the AoI metric. Accordingly, focusing on a freeway-based VRU scenario, we showcase in this section that in contrast to conventional, remote cloud-based cellular architecture, the deployment of MEC infrastructure can substantially prune the end-to-end communication latency as well as the experienced AoI. Our argument is supported by an extensive simulation-based performance comparison between the conventional and the MEC-assisted network architecture.

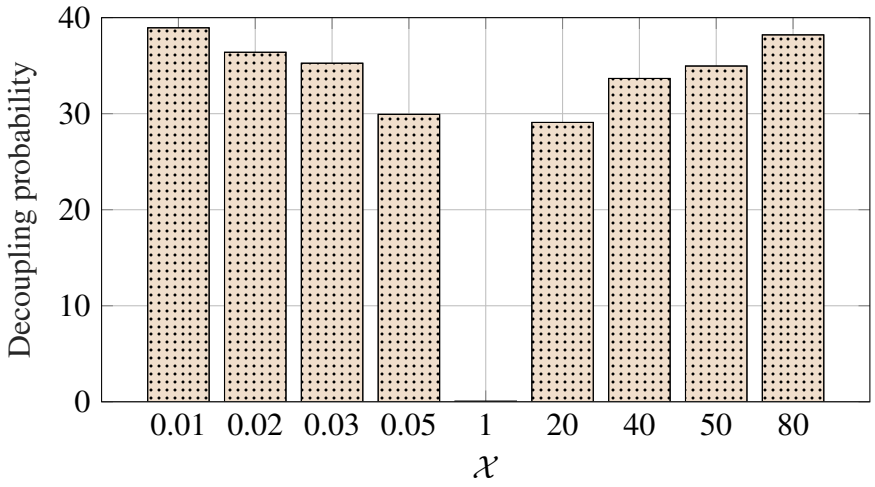


Figure 5.5: Fraction of devices reaching non-cohesive decisions upon cell association, as a function of cross-tier radio and MEC disparity.

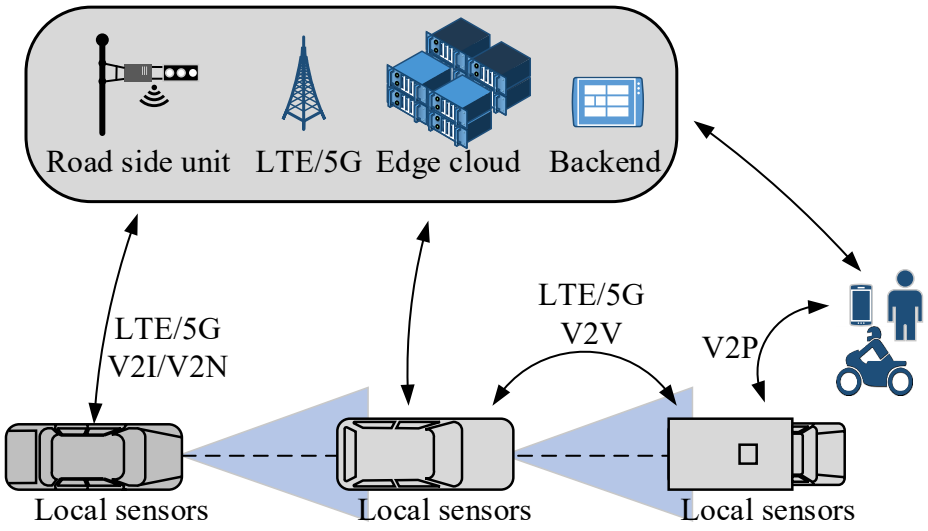


Figure 5.6: Envisioned 5G C-V2X system.

5.3.1 Background and Contributions

vehicle-to-everything (V2X) communication paves the way for a drastically improved road safety and driving experience via reliable and low latency wireless services [125]. The efficient V2X system development is based on a plethora of reliably-functioning sensors, which provide an enhanced environmental perception by means of exchanging critical messages among vehicles, pedestrians and road infrastructure [126]. Such a system, as depicted in Figure 5.6, incorporates different information exchange paths, namely, vehicle-to-infrastructure, vehicle-to-network, vehicle-to-pedestrian, and vehicle-to-vehicle. These signaling paths can be either established via direct short range communication, or, assisted by the cellular network providing coverage (C-V2X), or, through an inter-working of the two technologies [127].

Focusing on the C-V2X technology, the architecture of the cellular network is expected to have a vital impact on the support of delay-intolerant V2X services. This occurs, because the end-to-end latency of C-V2X signaling is limited by the quality and dimensioning of the cellular infrastructure, i.e., the capacity of back-haul connections, as well as the delays introduced by both the core network, as well as the transport network. As one would expect, these latency bottlenecks will be more prominent for high loads corresponding to coverage areas of high vehicular/ pedestrian densities. To cope with such requirements, extensive research has recently taken place to enhance the advent experience of V2X communication, with emphasis on latency shortening. For instance, in [128], the packet delivery latency and network utilization, focusing on an LTE system, are investigated for multimedia broadcast single frequency network. Furthermore, in [129], considering an LTE network architecture, core network gateway relocation is proposed for V2X latency improvement. Finally, with reference to implementation aspects, the authors in [130] investigate latency-reduction techniques such as transmission time shortening and self-contained sub-frames in C-V2X systems, whereas, in [131], a 5G implementation test-bed for autonomous vehicles based on software defined radio incorporating different solutions, was presented.

Nevertheless, in contrast to the above mentioned works, we argue that stringent latency requirements posed by the V2X system can be satisfied by introducing *MEC* technology to the cellular network architecture [132]. Leveraging its ability to provide processing capabilities at the cellular network's edge, an overlaid MEC deployment is expected to assist vehicles in achieving low packet delays, due to its close proximity to end devices, as shown in Section 5.2. In this section, concentrating on the VRU use case, which studies the safe interaction between vehicles and non-vehicle road devices (pedestrians, motorbikes, etc.) via the exchange of periodic cooperative awareness message (CAM) [133], we aim to reveal the latency-related benefits of introducing MEC system deployment over a state-of-the-art cellular network. Through extensive simulations, we show that the deployment of MEC infrastructure can substantially prune the end-to-end communication latency. Our study assumes V2X communication as it exploits the existing cellular infrastructure.

5.3.2 Spatial and Temporal Vehicular System Model

Throughout this section, a freeway road environment is assumed, consisting of one lane per direction, as shown in Figure 5.7. The vehicles are placed at the start of each system realization following a Matérn hard-core point process over one dimension [22]. The i-th vehicle's velocity is drawn from a uniformly

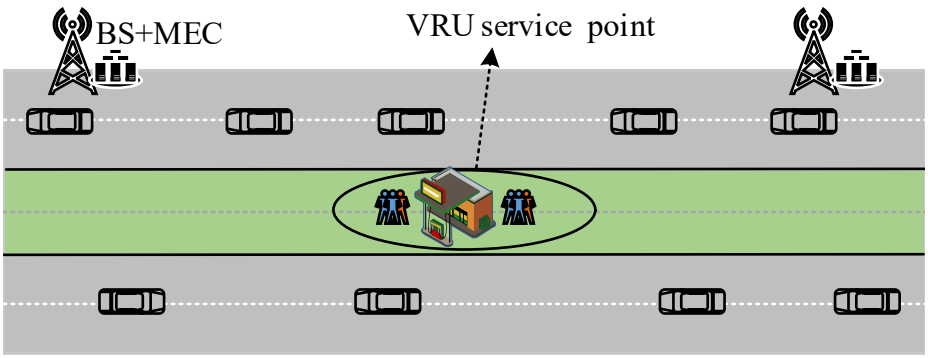


Figure 5.7: The investigated two-lanes freeway scenario.

distributed random variable (i.e., $\in U(v_{\min}, v_{\max})$), where $v_{\min}$ and $v_{\max}$, represent the minimum and maximum vehicle's velocity, respectively. To model the inter-vehicle distance, we have resorted to the hardcore parameter of the mentioned point process, which represents the repulsion between any two generated points. Additionally, a cluster of $\mathcal{N}_{\text{VRU}}$ VRUs is located on a pedestrian area between the two lanes; such a populated area can be mapped to real-world scenarios like gas stations or other service points across a freeway. At the network side, it is assumed that the focused freeway segment is under cellular coverage; given that, for brevity, we consider a continuous coverage scenario (i.e., occurrence of any handover events is not taken into account). The serving BS is assumed to be collocated with a MEC host of given processing capabilities, similar to the deployment considered in Section 5.2.

As mentioned earlier, a VRU is assumed to interact with vehicles and, possibly, other devices on the road. A straightforward example is the one of safety-related applications [134], in which periodically generated VRU messages (e.g., CAM) can be exploited for crash prevention purposes. In order to model the generation of those periodic messages, we assume that the k-th VRU generates data packets of size of $l_k \in \mathcal{U}(l_{\min}, l_{\max})$ bits at random starting time offsets, denoted as ν_k. Such CAM transmission randomness is used to model the nature of road-safety applications. Due to the CAM signaling periodicity, this cycle is repeated every T seconds with newly generated transmission offsets. A visualization of the messaging scheme for two VRUs is shown in Figure 5.8. It should be mentioned that, depending on the periodicity of packet generation and the number of VRUs existent at the focused service point, the available uplink radio resources are shared equally among the active VRUs. Once a given VRU transmits its CAM in the uplink exploiting the radio interface, the corresponding input packet will be processed by the MEC host collocated with the serving BS and then, the processed information (output packet) will be forwarded to vehicles in the vicinity of the VRU by means of downlink transmission.

According to the key results in [135], the main challenge in designing efficient C-V2X -CAM signaling is to serve the cell edge vehicles. Due to their low quality experienced conditions, such vehicles require a larger bandwidth, as compared to their cell-center counterparts. Therefore, accounting for the nature of CAM messages, where the end-to-end latency is dependent on the successful reception of the packets by the destined vehicles, we resort to the concept of *location-based vehicle clustering*. According to this approach and, based on location availability, each VRU defines a cluster of closest M vehicles, denoted by $\mathcal{H}_{M,k}$ for the k-th VRU, and a cluster-based multi-cast transmission takes place in the downlink.

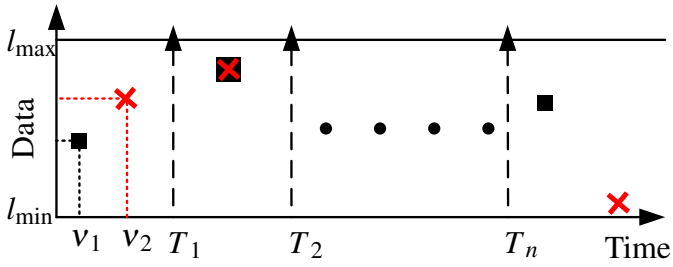


Figure 5.8: Packet generation procedure for two VRUs (black square and red cross, respectively) with random transmission timing offsets.

All the considered vehicles and VRUs are assumed to be served by an BS, based on the path-loss model adopted from the *WINNER+* project [136], as follows

$$\text{PL (dB)} = 22.7\log_{10}(r) - 17.3\log_{10}(\tilde{h}_{\text{BS}}) - 17.3\log_{10}(\tilde{h}_{\text{VRU}}) + 2.7\log_{10}(f_{\text{c}}) - 7.56, \tag{5.11}$$

where r is the distance between the transmitter and receiver, f_{c} is the center carrier frequency and $\tilde{h}_{\text{BS}}$ and $\tilde{h}_{\text{VRU}}$ represent the effective antenna heights at the BS and VRU, respectively. The latter quantities are computed as follows: $\tilde{h}_{\text{BS}} = h_{\text{BS}} - 1.0$ and $\tilde{h}_{\text{VRU}} = h_{\text{VRU}} - 1.0$, with h_{BS} and h_{VRU} being the actual antenna heights (i.e., in meters). Additionally, independent and identically distributed random variables are used to model the fast fading and shadowing-based attenuation phenomena. It should be noted that the scheduler employed in our work equally distributes the available bandwidth over all scheduled VRUs and vehicles. In what follows, a thorough end-to-end latency analysis is presented, focusing on both the proposed, MEC-assisted network architecture, as well as the conventional, *distant-cloud*-based cellular architecture, which will serve as a comparison benchmark for the numerical evaluations.

5.3.3 Latency Modeling of Network Components

As mentioned earlier, one objective of this section is to investigate the end-to-end latency gains achieved through MEC deployment within the network. Towards accomplishing this aim, in this section, we model the various latency components related to CAM transmission, routing and processing for both the proposed and conventional system approaches. Regarding the conventional cellular network architecture approach, which is depicted in Figure 5.9, the one-way CAM messaging latency is modeled as $\mathcal{T}_{\text{one-way}} = \mathcal{T}_{\text{UL}} + \mathcal{T}_{\text{BH}} + \mathcal{T}_{\text{TN}} + \mathcal{T}_{\text{CN}} + \mathcal{T}_{\text{Exc}}$, where $\mathcal{T}_{\text{UL}}$ is the radio uplink transmission latency, $\mathcal{T}_{\text{BH}}$ is the back-haul network latency, $\mathcal{T}_{\text{TN}}$ is the transport latency, $\mathcal{T}_{\text{CN}}$ is the core network latency and $\mathcal{T}_{\text{Exc}}$ is the CAM processing latency. Consequently, the end-to-end latency for the k-th VRU is expressed as

$$\mathcal{T}^k_{\text{E2E, C}} = \mathcal{T}^k_{\text{UL}} + \underbrace{2(\mathcal{T}^k_{\text{BH}} + \mathcal{T}^k_{\text{TN}} + \mathcal{T}^k_{\text{CN}})}_{\text{Network latency}} + \mathcal{T}^k_{\text{Exc}} + \mathcal{T}^k_{\text{DL}}, \tag{5.12}$$

where, $\mathcal{T}^k_{\text{DL}}$ represents the downlink transmission latency. For the proposed, MEC-enabled network approach, the network latency can be avoided via processing the CAM packets at the MEC host,

collocated with the connected BS. Therefore, in this case, the average end-to-end latency is given by[1]

$$\mathcal{T}^{k}_{\text{E2E, MEC}} = \mathcal{T}^{k}_{\text{UL}} + \mathcal{T}^{k}_{\text{Exc}} + \mathcal{T}^{k}_{\text{DL}}. \tag{5.13}$$

As described in subsection 6.2.2, for a given messaging cycle, each VRU generates a packet for transmission at a random starting time instant. In this section, we assume fair resource allocation, where the available bandwidth is shared equally among the VRUs transmitting at the same time index. Thus, the number of these VRUs, denoted by $\hat{N}_k$, sharing the resources with the k-th VRU equals

$$\hat{N}_k = \sum_{i=1}^{\mathcal{N}_{\text{VRU}}} 1(\nu_i = \nu_k),\ \forall k = \{1,2,\cdots,\mathcal{N}_{\text{VRU}}\}. \tag{5.14}$$

Thus, the time required for the k-th VRU to transmit a packet of size of l_k bits to its serving BS is computed as

$$\mathcal{T}^{k}_{\text{UL}} = \frac{l_k}{r^{UL}_k},\ r^{UL}_k = \frac{\mathcal{B}}{\hat{N}_k}\log_2(1+\mathsf{SINR}_k), \tag{5.15}$$

where r^{UL}_k is the achievable uplink rate, $\mathcal{B}$ is the system's bandwidth and SINR_k represents the received SINR at the BS. Due to the periodic nature of message generation, the computation of shared resources is carried out for each time window (i.e., $[T_j, T_{j+1}]$, $\forall j = \{1,2,\cdots\}$). As mentioned in Section 6.2.2, for downlink transmissions, after successful packet processing at the host, we resort to the concept of cluster-based multi-cast transmission [135]. The main idea is to select a set of vehicles in the system for transmission, in order to avoid large latencies caused by cell-edge vehicles, which would not be of high criticality for the VRU, as the set of VRUs is assumed to be located close to the cell center. Consequently, the cluster of the vehicles for the k-th VRU denoted as $\mathcal{H}_{k,M}$, will consist of the M closest vehicles to that VRU. Thus, the downlink latency can be expressed as follows

$$\mathcal{T}^{k}_{\text{DL}} = \max_{(\forall i \in \mathcal{H}_{k,M})}\{\frac{l_k}{r^{DL}_k}\}, \tag{5.16}$$

where r^{DL}_k denotes the downlink rate of the k-th vehicle and the maximum operator is used to measure the farthest vehicle's packet reception delay in cluster $\mathcal{H}_{M,k}$. Regardless of the BS location, having the k-th VRU position as a reference, the maximum radio downlink latency serves as a cluster-wide metric, which is aimed to be minimized. As it will be shown later, the effect of the cluster size is significant, since the available radio resources in the downlink have to be shared among all vehicles within cluster $\mathcal{H}_{M,k}$.

As mentioned earlier, the following latency components are non-existent for the MEC-assisted CAM signaling case, since there is no involvement of the back-haul, core, and transport network components in CAM packet routing. The back-haul latency $\mathcal{T}_{\text{BH}}$ represents the time required for packets to be routed through the back-haul network, which has a finite capacity, denoted by C_{BH}. It is assumed that the back-haul capacity is equally shared among the $\hat{N}_k$ VRUs concurrently uploading their messages at time instant ν_k. As a result, assuming that the packet size is the same for all VRUs, the back-haul latency for

[1] Latency from the BS to the MEC host and vice versa is not considered and is left for future work.

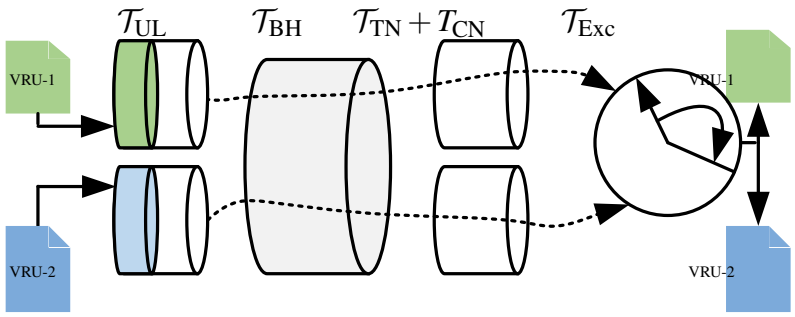


Figure 5.9: One-way signaling latency for two VRUs - conventional approach.

the k-th VRU is

$$\mathcal{T}_{\mathrm{BH},k} = \frac{l_k \hat{N}_k}{C_{\mathrm{BH}}}. \tag{5.17}$$

In order to provide realistic modeling of the transport and core network latencies, we resorted to the recent results reported in [137], where a proof-of-concept was implemented for an LTE environment with commercial terminals, running a real-time adaptive video streaming service routed through a MEC host and several BS agents placed at different geographical locations, as compared to the MEC host position. More details regarding the system setup and the methodology employed can be found in [137]. Consequently, inspired by the results presented in the mentioned work, the two latency components are assumed to be uniformly distributed, over a range of realistic values, as it will be shown in the numerical evaluation section.

Finally, we model the time required for processing a packet of size of l_k bits at a host, either collocated with the BS or at the distant cloud. Assuming that the input packet requires f_k cycles/bit for processing and the host has a processing capacity of F, the k-th VRU execution latency equals

$$\mathcal{T}_{\mathrm{Exc},k} = \frac{\hat{N}_k l_k f_k}{F}. \tag{5.18}$$

5.3.4 Information Freshness Quantification of the VRU Messages

Concentrating on the VRU use case, we argue that, apart from the end-to-end latency, the freshness of continuous status updates of nodes within a V2X system is another fundamental performance indicator to ensure efficient service functionality, especially for safety-critical situations. This implies continuous information update about the real-time state between a given source and its targeted destination [13]. The AoI metric proposed in [48] characterizes the freshness of information at the receiver and has recently received increased attention as it is a useful metric to evaluate the efficiency of technology solutions for various vertical industries, such as the automotive one. Consequently, for the examined use case, to ensure an almost real-time VRU awareness across the vehicles, it is the timeliness of VRU messages received by nearby vehicles that would rather need to be monitored and improved, e.g., by properly varying the VRU packet generation traffic. In relation to that, a critical challenge is how to maintain timely VRU status updates across all approaching connected vehicles [138].

For the considered vehicular time-slotted system, the AoI function, $\Delta_k(t)$, tracks the AoI evolution over time, t, at each of the cluster member vehicles aimed to be reached by the k-th VRU. Let G_k denote

the packet generation time stamp for the k-th VRU; then, focusing on a specific vehicle/ cluster member, the AoI at the $(t+1)$-st time slot, denoted by $\Delta_k(t+1)$, is computed recursively as follows

$$\Delta_k(t+1) = \begin{cases} \Delta_k(t)+1, & \text{if no update was received,} \\ t - G_k + 1, & \text{otherwise.} \end{cases} \tag{5.19}$$

In this section, focusing on a given VRU, we consider the cluster-wide PAoI, which is defined as the AoI observed at the farthest member of the vehicle cluster targeted by the VRU, when achieved immediately before this vehicle receives a new VRU message. As discussed before, the PAoI represents the temporally averaged peaks attained by the AoI function. As the PAoI provides insights on guaranteed system performance, we deem it as an important metric for the investigated VRU scenario. Mathematically, and based on the periodic nature of the VRU messages, the PAoI of the k-th VRU, when averaged over time, is expressed as follows

$$\Delta_{\mathrm{p},k} = \mathbb{E}_t \left\{ \mathcal{I}_k + \mathcal{T}^k_{\mathrm{E2E,\,j}} \right\} = T + \mathcal{T}^k_{\mathrm{E2E,\,j}}, \;\; k \in \{1, \cdots, \mathcal{K}\}, \tag{5.20}$$

where $j \in \{\mathrm{MEC}, \mathrm{C}\}$, $\mathbb{E}_t\{.\}$ is the temporal expectation operator, while, $\mathcal{I}_k$ and $\mathcal{T}^k_{\mathrm{E2E,\,j}}$ denote the inter-arrival time between consecutive VRU messages and the end-to-end latency of a given VRU message under a given network architecture, respectively. As highlighted earlier, the objective of this section is to investigate the VRU awareness timeliness performance achieved through collocated deployment of a MEC and cellular network infrastructure and compare it to the one of conventional cellular system architecture incorporating (distant) cloud infrastructure. Therefore, based on presented models in this section, the network-wide PAoI, averaged over all $\mathcal{N}_{\mathrm{VRU}}$ VRUs in the network is expressed as

$$\mathbb{E}_k\{\Delta_{\mathrm{p},k}\} = \frac{1}{K} \sum_{k=1}^{K} (T + \mathcal{T}^k_{\mathrm{E2E,\,j}}), \quad j \in \{\mathrm{C}, \mathrm{MEC}\}, \tag{5.21}$$

where $\mathcal{T}^{(k)}_{\mathrm{E2E,\,j}}$ represents the time-averaged end-to-end latency for the k-th VRU.

5.3.5 C-V2X Evaluation Campaign

In order to illustrate the latency improvements via MEC deployment within cellular systems for V2X communications, we provide different simulation scenarios by varying the values of two main system parameters; namely, the vehicles and VRUs spatial densities. Moreover, we also aim at observing the vehicles' cluster size impact on the experienced latency. For both the proposed and conventional cellular network architectures, the focused metric is the end-to-end latency, as well as its individual components as explained in (5.12). The values of all involved parameters are presented in Table 6.2, unless otherwise stated.

First, we look into the case of increasing VRUs. As explained in the previous sections, each VRU is assigned a random timing offset for transmission. Thus, the generated periodic message traffic increases accordingly with the VRUs. In Figure 5.10, the average end-to-end signaling latency with and without MEC host deployment is shown both as a whole and component-wise. Clearly, MEC utilization provides a lower end-to-end latency (the observed gains are in the range of 66%-80%), due to the exploitation of

Table 5.2: Simulation parameters for C-V2X evaluations

Entity	Parameter	Value
Vehicles	Velocity	$\sim \mathcal{U}(70, 140)$ km/h
	Inter-vehicle distance	10 m
	Vehicles spatial intensity (μ_{veh})	0.01 vehicles/m
	Cluster size (M)	5
VRU	Number of VRUs ($\mathcal{N}_{\text{VRU}}$)	100
	x-coordinates	$\sim \mathcal{U}(1200, 1800)$
	Transmit power	23 dBm
	Packet size (l_k)	$\sim \mathcal{U}(8, 12)$ kbits
	Processing per bit (f_k)	$\sim \mathcal{U}(100, 300)$ cycles/bit
BS / MEC host	Transmit power	46 dBm
	Bandwidth ($\mathcal{B}$)	9 MHz
	Back-haul capacity (C_{BH})	10 Mbps
	F	9×10^9 cycles/sec
General	Frequency (f_s)	5.9 GHz
	Number of lanes	2
	Lane length	3 km
	Lane width	6 m
	Path-loss exponent (η)	3
	Shadowing standard deviation	3 dB
	Fast fading standard deviation	4 dB
	Thermal noise power	-110 dBm
	Additional losses	15 dB
	Transport and core network latency	$\sim \mathcal{U}(15, 35)$ milliseconds

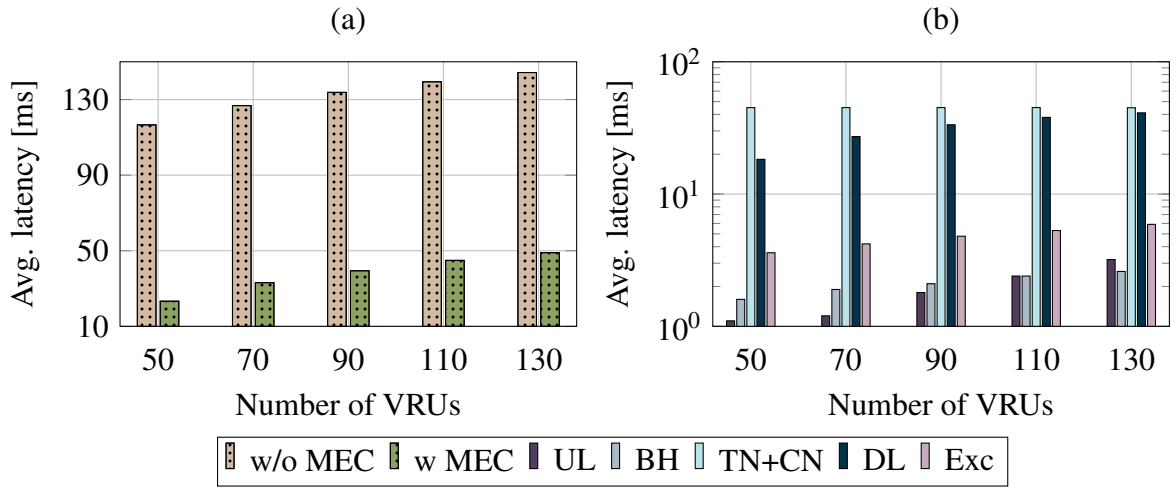


Figure 5.10: (a) Average end-to-end latency and (b) Component-wise latency.

processing resource proximity offered by the MEC host. Additionally, we observe an increasing behavior of the latency along with the VRU density, which is due to the increasing demand of the available resources. First, for the radio transmission latency components, as the number of VRUs increases, the available resources per VRU decrease, due to the equal allocation assumption. Similar explanations hold for the back-haul (BH) and the execution latencies. It should be noted that the transport and core network latencies were modeled as random variables, independent of the other system parameters values.

Regarding the network-wide PAoI behavior of the system for increasing VRU density, due to the periodic nature of VRU message generation, for an increased number of VRUs, the generated VRU message traffic per unit time within the network will increase as well, hence, resulting to less radio and processing resources allocated per VRU to transmit and process each VRU message, respectively. In Figure 5.11, assuming that T=100 milliseconds, the network-wide PAoI performance is illustrated, for both the MEC-enabled and conventional network architecture variants. Firstly, as expected, for both system architecture variants, we observe a monotonically increasing behavior of the PAoI as a function of the VRU load, owing to the increasing demand for radio and processing resources. Furthermore, it is observed that, for all considered values of $\mathcal{N}_{\text{VRU}}$, MEC infrastructure utilization provides a lower PAoI, thus, higher information timeliness, which, in its turn, is translated into better VRU awareness, compared to the conventional cellular architecture. As an example, for $K = 150$ VRUs, the achieved PAoI is equal to $\tilde{\Delta}^{\text{p}}_{\text{MEC}} = 160$ milliseconds, which is only a fraction of $\tilde{\Delta}^{\text{p}}_{\text{C}} = 258$ milliseconds achieved by the conventional network architecture. Such a, nearly 61%, reduction in PAoI, is due to the exploitation of MEC processing resource proximity.

To jointly evaluate the effect of VRU packet generation periodicity on system-wide timeliness and end-to-end delay performance, along with the performance gains provided by the existence of MEC infrastructure, assuming the existence of $\mathcal{N}_{\text{VRU}}$=100 VRUs in the system, we measure the network-wide PAoI together with the average E2E VRU message latency for various VRU packet inter-arrival times, $T \in [10\text{ms}, 100\text{ms}]$. Figure 5.12 depicts the numerical evaluation results, where, PAoI and average end-to-end delay values appear in the left and right hand side vertical axes of the figure, respectively. Apart from the clear MEC-related performance gains, one can identify two different performance behaviors with respect to the VRU packet inter-arrival time for both network architecture options.

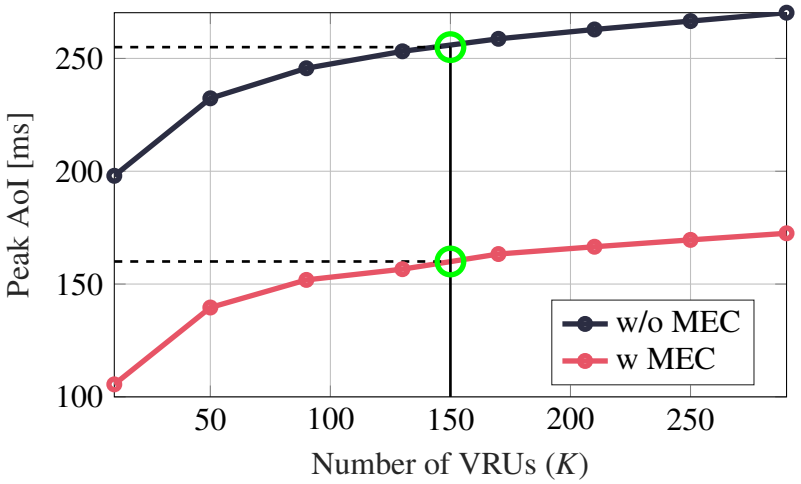


Figure 5.11: Network-wide PAoI with increasing VRU density for $T = 100$ ms.

When $T \in [10\text{ms}, 30\text{ms}]$, both the achieved PAoI and the average E2E latency performance curves are monotonically decreasing, as a function of T. Such a behavior is justified as, in this regime, in contrast to T, the average E2E delay, which dominantly contributes to the PAoI, progressively reduces due to the reducing congestion on the available resources; this PAoI regime can be labeled as a *resource stagnation-driven* one. On the contrary, when $T \in [30\text{ms}, 100\text{ms}]$, it is observed that, although the average end-to-end latency continues to decrease, as a function of T, the achieved PAoI starts to increase. This behavior divergence occurs, because, focusing on the end-to-end latency, the resource contention among the VRUs radically decreases, as the set of possible VRU transmission offsets becomes fairly larger, hence, leading to lower overall delay per VRU message. Nevertheless, larger values of T imply less frequent VRU status updates, resulting to higher values of the PAoI, as T now decisively contributes to it; this PAoI regime can be labeled as an *update scarcity-driven* one. In summary, we observe the limitations of considering the end-to-end latency as the sole objective of system design, with regards to time-critical applications of C-V2X communications. To alleviate these limitations, AoI minimization shall be the overall design objective when it comes to such applications and use cases.

In this part, an alternative scenario of fixing the number of VRUs and increasing the spatial density of the vehicles is studied, as per Figure 5.13. Since the VRUs in the investigated use case are the active agents and the vehicles are the passive ones, i.e., transmission is always initiated by the VRUs, the end-to-end latency is dependent on the vehicles' spatial density. As discussed in Section 6.2.2, the vehicles' density (i.e., μ_{veh}) only plays a role in the radio downlink latency. Since a location-based multi-cast transmission is employed, where the cluster size (i.e., $|\mathcal{H}_{M,k}|$) is fixed, as the number of vehicles increases, the probability to have the cluster closer to the VRU of interest will increase as well. Hence, as expected, the downlink latency decreases with increasing μ_{veh}.

Since the cluster size highly affects the end-to-end latency through its contribution to the downlink radio latency, the experienced downlink latency for increasing vehicle cluster sizes is simulated and presented in Figure 5.14. Due to the definition of the downlink latency (eq. (5.16)) and its dependence on the cluster's farthest vehicle to successfully receive the packet, as the cluster size increases, the probability of vehicles being far from the focused VRU will increase as well. As a result, this explains the increasing fashion of the radio downlink latency, which is as depicted in Figure 5.14.

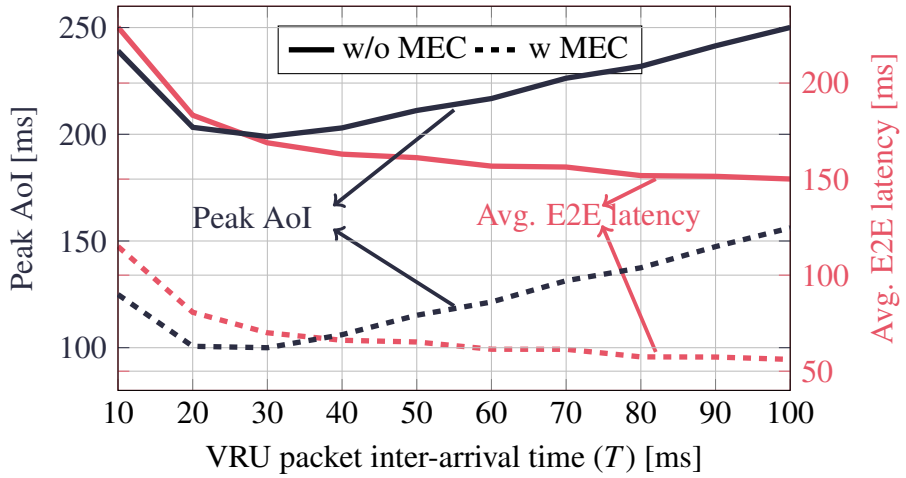


Figure 5.12: Peak AoI and average E2E latency for increasing VRU packet inter-arrical time with $K = 100$ VRUs.

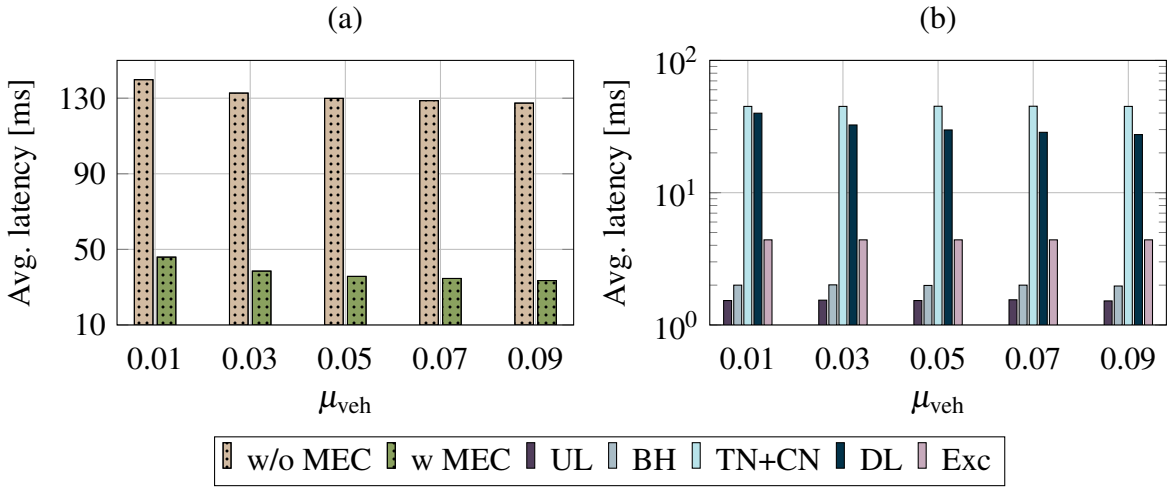


Figure 5.13: (a) Average end-to-end latency for increasing vehicles' deployment densities. (b) Component-wise latency breakdown.

5.4 Conclusion

In this Chapter, we leverage the MEC degree of freedom in planning and dimensioning wireless networks, via the joint investigation of the communication and computation resources. First, for a task offloading use case, a new association metric for uplink communication in a heterogeneous network is proposed, aiming at reducing the experienced packet latency budget of a device. Different scenarios spanning diverse radio and MEC cross-tier disparities are presented to highlight the cell association decision effect on system performance. It is shown that, for a range of disparities between radio and MEC capabilities between tiers, the proposed computation proximity rule provided gains in terms of latencies, as compared to the conventional maximum RSRP rule. This performance gain degrades as cross-tier radio/ MEC disparities become similar. Also importantly, we explore the case, in which, for different association rules, a device would favor associating to different BS/MEC hosts in the uplink.

In addition, we investigate the problem of improving the timeliness of collective road awareness, concentrating on the vehicular segment and focusing on an VRU use case under cellular network

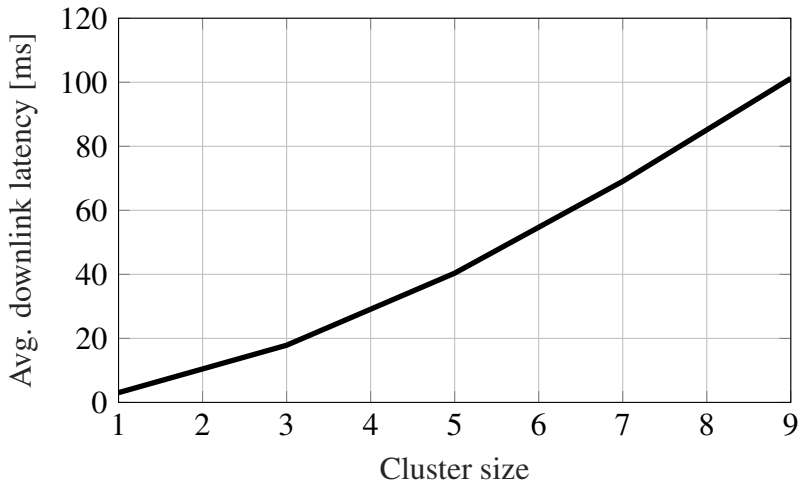


Figure 5.14: Average cluster-related radio downlink latency as a function of the vehicle cluster size.

coverage. With the aim of minimizing end-to-end signaling latency, we propose a MEC-assisted network architecture, according to which MEC hosts are collocated with BSs, thus, they can receive and process VRU messages at the edge of the access network. Towards quantifying the benefits of the new approach, we define the latencies related to radio transmission and message processing, driven by realistic assumptions. In addition, focusing on the PAoI as a means to quantify the information freshness of VRU messages, we quantify the achieved PAoI for both the proposed, MEC-assisted and the conventional network architectures. Via numerical evaluation for some of the investigated system parameterizations, the proposed overlaid deployment of MEC hosts offers up to 80% average gains in latency reduction, as compared to the conventional network architecture. In addition, we show that, for a given VRU load, the network-wide PAoI of the conventional system architecture can be reduced by nearly 61% when a MEC-enabled architecture is taken into account. It is interestingly shown that performance benefits remain significant for different vehicle/ VRU deployment densities, as well as for different inter-packet generation times.

Chapter 6

Dependable Computation Services in Wireless Systems

6.1 Introduction

As shown throughout the previous chapter, the deployment of MEC in 5G and beyond systems allows more efficient task execution, owing to the MEC hosts high computation power. Nevertheless, a major challenge for such systems is to provide dependable and ubiquitous computing services that meet the computing demands of devices running various heterogeneous applications [7]. On the other hand, wireless links are characterized by fluctuating quality, leading to variable packet error rates which are orders of magnitude higher than the ones of wired links [139]. Therefore, it is of paramount importance to develop tools and frameworks that provide insights regarding the limitations of using wireless links for IoT applications. Such frameworks represent a first step towards the realization of determinism of process flows *anytime* and *anywhere*. Based on a spatiotemporal approach, in this chapter, we provide novel definitions of dependability attributes for communication and computation services.

In details, Section 6.2 presents a novel spatiotemporal framework that utilizes stochastic geometry and continuous time Markov chains to jointly characterize the communication and computation performance of MEC-enabled wireless systems. Additionally, we evaluate the influence of various system parameters on dependability metrics such as (i) computation resources availability, (ii) task execution retainability, and (iii) task execution capacity. Our findings showcase that there exists an optimal number of virtual machines for parallel computing at the MEC host to maximize the task execution capacity. In Section 6.3 the availability and reliability of a given service, assuming a number of BSs and devices deployed over a fixed area, are quantified. In the space domain, we characterize spatially available areas consisting of all locations that meet a predefined performance requirement with given confidence. In the time domain, we propose a channel allocation scheme accounting for the spatial availability of a cell. With the aim to reveal the incurred space-time performance trade-offs, numerical results are presented, also highlighting the effect of different system parameters on the achievable service availability and reliability. Finally, Section 6.4 summarizes this chapter.

6.2 Dependable Task Execution Services in MEC-enabled Wireless Systems

6.2.1 Background and Contributions

The joint consideration of i) contention-based communications for task offloading and ii) parallel computing and occupation of failure-prone MEC processing resources (virtual machines), is envisaged, to properly understand the communication and computation chain [3]. To ensure efficient operation, the task offloading reliability, computation resources availability, and task execution retainability ought to be jointly quantified and optimized. In MEC-enabled networks, successful task execution at the MEC host is strongly tied to its resources availability and its resilience to failures [54]. In this context, various resilience and provisioning techniques are discussed in [140] with a focus on cloud computing infrastructures. Causes of service disruption due to physical machines and virtual machines (VMs) failures along with their analysis are provided in [141]. With regard to wireless-based task offloading, [142] examines the network scalability and identifies communication and computation performance frontiers. Analysis for heterogeneous networks is presented in [143], where the network-wide outage probability is derived for task offloading assuming different computation architectural variants. Authors in [144] proposed a transmission and energy efficient offloading algorithm based on a Markov decision process that accounts for the spatial and temporal network parameters.

However, the aforementioned works either consider a dependability view of the network[54, 140, 141], or a spatiotemporal one [142–144], where the problem of feasible and dependable task execution, considering the joint limitation of network-wide mutual interference and parallel task computing by failure-prone VMs, under a spatiotemporal framework, is still not addressed. Accordingly, we propose a spatiotemporal feasibility-assessment framework that entails network-wide mutual interference and temporal-based task arrivals/ processing in uplink MEC-enabled networks. Furthermore, we adopt an individual (i.e., per-task and per-device) task execution criterion that aims to exploit the computation resources at the MEC server if the radio conditions permit. Our analysis is then followed by the assessment of new service dependability-relevant KPIs that shed light on the system availability and task execution capability.

6.2.2 System Model

Network model

We consider a cellular uplink network, where the BSs and devices are spatially deployed in $\mathbb{R}^2$ according to two independent homogeneous Poisson point processes (PPPs), denoted by Φ and Ψ with intensities λ and μ, respectively. An unbounded path-loss propagation model is adopted such that the signal power attenuates at rate of $r^{-\eta}$, where r is the distance and η is the path-loss exponent. Wireless links are assumed to undergo Rayleigh fading, where the signal of interest h and interference channel power g gains are exponentially distributed with unit power gain. Full path-loss channel inversion power control is adopted, which implies that all devices adjust their transmit powers such that the received uplink power levels at the BS are equal to a predetermined threshold ρ [73].

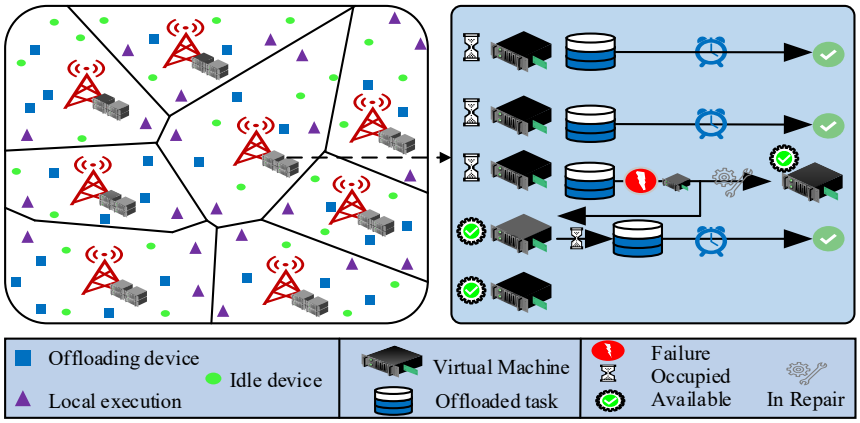


Figure 6.1: System setup involving a MEC host equipped with 5 VMs.

Offloading model

We consider a continuous time system where task arrivals at each device are modeled via an independent Poisson process with rate λ_a tasks/ unit time [142, 143]. In a proactive manner, devices attempt to offload generated tasks by sending instructions to the MEC host at the BS. In our system, grant-free access is assumed, where each device attempt to transmit its task instruction (i.e., not the whole task) using one of the available C uplink channels randomly and uniformly without a scheduling grant from the BS [145]. Furthermore, let $\kappa = \frac{\mu}{\lambda C}$ denote the average number of devices per BS per channel and T_s be the transmission time of a given task's instruction. A task instruction is successfully decoded at the BS if its received SINR is larger than a predefined threshold, θ. The offloading success probability of a generic device, which is denoted by $\mathcal{O}$, quantifies the probability of successful task offloading as $\mathcal{O} = \mathbb{P}\{\mathsf{SINR} > \theta\}$.[1] In the case of decoding failure (i.e., NACK is received), the device opts to compute its task locally. Accordingly, we adopt a coverage-based offloading feasibility criterion, in which the offloading success probability $\mathcal{O}$ governs the offloading feasibility, thus, the offloading decision problem and its underlying parameters are not considered and left for future work. Retransmissions at the devices are not considered in the proposed model to lower the aggregate network-wide interference [146].

Computing model

The MEC host residing at each BS is equipped with a single physical machine that encompasses $\mathcal{V}_{\text{MEC}}$ VMs for parallel task computing. To account for resource sharing among the VMs (e.g., buses for input/output (I/O), CPU, memory), I/O interference is observed within the physical machine at the MEC host. Thus, the parallel-operating VMs interfere with each other, leading to a degraded computation power [147]. For the case of a single VM deployment, the task's execution rate is modeled via a Poisson process with a rate of μ_o tasks/ unit time. However, to account for the I/O interference among the $\mathcal{V}_{\text{MEC}}$ VMs, the task's execution rate of a given VM depends on the total number of VMs as follows

$$\mu_{\text{MEC}} = \frac{\mu_o}{(1+d)^{\mathcal{V}_{\text{MEC}}-1}}, \tag{6.1}$$

[1] acknowledgment (ACK) and non-acknowledgment (NACK) transmission latencies are ignored as they incur negligible amount compared to T_s and the task's execution time.

where d is the computation degradation factor due to I/O interference among the $\mathcal{V}_{\text{MEC}}$ VMs [142, 147]. For local computation of tasks, devices are assumed to be equipped with a local physical machine that accommodates a single VM (i.e., $\mathcal{V}_{\text{loc}} = 1$, thus, no parallel processing), where the local computation rate is modeled via a Poisson process with rate μ_{loc}. Moreover, a task to be computed is blocked if no VM is idle (locally or at the MEC host in case of offloading). To investigate the relative ratio between the MEC and the local computation capabilities, we define $\mu_r = \mu_{\text{MEC}}/\mu_{\text{loc}}$ which denotes the relative computation rate such that $\mu_r >> 1$.

Failure & repair model

Due to possible hardware and software faults, the proposed model accounts for events of VM failures and their repairment times [148, 141]. The failure (repair) rate of a given VM is modeled via a Poisson process with rate $\mathcal{F}$ ($\mathcal{R}$) failure (repairment) events/ unit time.[2] VMs are prone to failure regardless of being idle or occupied. A failed idle VM is labeled as out of operation and cannot admit future tasks. Upon the failure of an occupied VM, the physical machine will handover the running task to an idle VM, if one exists. If not, the running task is discarded and the concerned device is notified via downlink signaling. The considered system model is visualized in Figure 6.1, where one can observe a plethora of devices belonging to three categories, namely, idle, offloading and local execution devices. Focusing on a selected cell that serves a number of offloading devices, a VM fails while being in service. Thus, the task being served by this VM is transferred to an idle VM to resume its execution. Meanwhile, based on the repair rate, the failed VM goes back into operation to serve newly incoming tasks.

6.2.3 Spatial System Analysis

Upon task generation, the task instructions are sent to the MEC host co-located with the BS by uplink transmissions. Those instructions are correctly decoded, and hence the task is successfully offloaded, if the received SINR is greater than θ. Otherwise, the device executes the task locally. To characterize the offloading feasibility within the network, the offloading success probability of a randomly selected device considering the network-wide mutual interference is

$$\mathcal{O} = \mathbb{P}\left\{\frac{\rho h_0}{\sum_{y_n \in \Psi \setminus y_o} a_n P_n g_n ||y_n - z_o||^{-\eta} + \sigma^2} > \theta\right\}, \tag{6.2}$$

$$\overset{(a)}{=} \exp\left\{-\frac{\sigma^2 \theta}{\rho}\right\} \mathcal{L}_{I_{\text{out}}}\left(\frac{\theta}{\rho}\right) \mathcal{L}_{I_{\text{in}}}\left(\frac{\theta}{\rho}\right). \tag{6.3}$$

where h_o is the channel gain between the intended device and its serving BS located at z_o, $||.||$ is the Euclidean norm, y_n is the n-th device location in the network excluding the intended device $\Psi \setminus y_o$, P_n is its transmit power, g_n is the channel power gain between this interfering device and the intended BS, σ^2 is the noise power and a_n equals one if the n-th device is transmitting on the same channel as the intended device, and zero otherwise. In addition, (a) results from the exponential distribution of h_o combined with the path loss inversion power control, where $\mathcal{L}_{I_{\text{out}}}(\cdot)$ and $\mathcal{L}_{I_{\text{in}}}(\cdot)$ represent the LT of the aggregate

[2]The Poisson model is adopted in our work for task-related parameters to provide a good compromise between practical consideration of real-time events and mathematical tractability [142, 143].

Table 6.1: State transitions $z=(x_I,x_O,x_F)$ of the VMs.

Event	Des. state	Rate	Condition
1- Task arrival and an idle VM is allocated	(x_I-1,x_O+1,x_F)	λ_v	$x_I>0$
2- Task execution at an occupied VM	(x_I+1,x_O-1,x_F)	$x_O\mu_v$	$x_O>0$
3- An idle VM fails	(x_I-1,x_O,x_F+1)	$x_I\mathcal{F}$	$x_I>0$
4- An occupied VM fails. Task is offloaded to another idle VM	(x_I-1,x_O-1,x_F+1)	$x_O\mathcal{F}$	$x_O,x_I>0$
5- An occupied VM fails and task is aborted	(x_I,x_O-1,x_F+1)	$x_O\mathcal{F}$	$x_O>0,x_I=0$
6- A failed VM is repaired	(x_I+1,x_O,x_F-1)	$x_F\mathcal{R}$	$x_F>0$

intra-cell and inter-cell interference, respectively. To provide an uplink tractable analysis, we assume that the spatial correlations between adjacent Voronoi cell areas are ignored, thus, the transmission powers of the devices are independent and identically distributed [73, 35]. The aforementioned approximations are validated in S subsection 6.2.5 against independent Monte Carlo simulations. In order to quantify the total arrival rate of offloaded tasks at the MEC host, the offloading success probability of each device is first calculated in the following theorem.

Theorem 4. *The offloading success probability for a generic device is given by*

$$\mathcal{O}\approx\frac{\exp\left\{-\frac{\sigma^2\theta}{\rho}-\frac{2\theta P_a\kappa}{(\eta-2)}{}_2F_1(1,1-2/\eta,2-2/\eta,-\theta)\right\}}{\left(1+\frac{\theta P_a\kappa}{(1+\theta)c}\right)^c}$$
$$\overset{(\eta=4)}{=}\frac{\exp\left\{-\frac{\sigma^2\theta}{\rho}-P_a\kappa\sqrt{\theta}\arctan\left(\sqrt{\theta}\right)\right\}}{\left(1+\frac{\theta P_a\kappa}{(1+\theta)c}\right)^c}, \tag{6.4}$$

where $P_a=1-e^{-(2T_s\lambda_a)}$ is the device's active probability within $[-T_s,T_s]$, $\kappa=\frac{\mu}{\lambda C}$, ${}_2F_1(\cdot)$ is the Gaussian hyper-geometric function and $c=3.575$. The approximation is due to the employed approximate probability distribution function (PDF) of the PPP Voronoi cell area in $\mathbb{R}^2$.

Proof. Proof can be shown following similar steps that were conducted for Theorem 1 in Chapter 3, while taking into consideration that $P_a\kappa$ denotes the portion of interfering device within the network. ■

Once $\mathcal{O}$ is evaluated, we can now define and evaluate the related task execution KPIs for the case of offloaded and locally executed tasks as explained in the following section.

6.2.4 Temporal Computational Analysis

As explained earlier, the offloading success probability provides an offloading feasibility assessment via controlling the aggregate load of tasks at the MEC host. That is, the total average arrival rate of tasks to be computed at the MEC host is $\lambda_{\text{MEC}}=\mathcal{O}\lambda_a\mathbb{E}\{\mathcal{N}_d\}=\frac{\mathcal{O}\lambda_a\mu}{\lambda}$. On the other hand, the average arrival rate of tasks to be locally computed is $\lambda_{\text{loc}}=\bar{\mathcal{O}}\lambda_a$ tasks/ unit time, where $\bar{\mathcal{O}}=1-\mathcal{O}$. To analyze the temporal occupancy of the VMs either locally or at the MEC host, we employ tools from queueing theory. To construct the proposed CTMC, we first determine the system's state space. A general state of our model is represented by the tuple $z=(x_I,x_O,x_F)$; where $x_i;i\in\{I,O,F\}$ represents the number

of VMs that are idle, occupied and failed, respectively. Let $\mathcal{S}_v = \left\{z | \sum_j x_j = M_v; j \in \{I,O,F\}\right\}$ denote the state space, where $v \in \{\text{MEC}, \text{loc}\}$ denotes the MEC and local systems. The steady state equations can be vectorized as $\tau_v = [\tau_1\ \tau_2\ \cdots\ \tau_\ell\ \cdots \tau_{|\mathcal{S}_v|}]$, where τ_ℓ is the probability of being in the ℓ-th state. For full temporal characterization, we need to construct the state transition matrix $\mathbf{Q}_v$. For each system v, $\mathbf{Q}_v$ constitutes the transition rates associated with different states. To systematically construct τ_v, while taking into account the different temporal events, Table I is utilized, which entails the transition rates and conditions among different system states.Focusing in this work on the steady state solution, the steady state probabilities are evaluated via solving

$$\tau_v \mathbf{Q}_v = 0, \quad \text{and} \quad \sum_{z \in \mathcal{S}_v} \tau_v(z) = 1. \tag{6.5}$$

Let 1 and $\mathcal{I}$ denote the all ones vector and the all ones matrix, with the appropriate sizes respectively, then, τ_v equals

$$\tau_v = 1(\mathbf{Q}_v + \mathcal{I})^{-1}. \tag{6.6}$$

Once the solution τ_v is obtained, several dependability-based KPIs can be assessed. First, we consider the communication resources availability. This metric quantifies the probability that an incoming device's task, either locally managed or offloaded to the MEC host, finds a vacant computation resource. First, let $\mathcal{N}_v = \{z | x_I = 0, z \in \mathcal{S}_v\}$ denote all states with no idle VMs. Then, the communication resources availability, denoted as $\mathcal{A}_t$, can be evaluated as

$$\mathcal{A}_t = \mathcal{O}\left(1 - \sum_{z \in \mathcal{N}_{\text{MEC}}} \tau_{\text{MEC}}(z)\right) + \bar{\mathcal{O}}\left(1 - \sum_{z \in \mathcal{N}_{\text{loc}}} \tau_{\text{loc}}(z)\right). \tag{6.7}$$

Another important KPI that quantifies the degree of successful task execution, is the task execution capacity. Let $\mathcal{C}_v = \{z | x_O > 0, z \in \mathcal{S}_v\}$ denote all states with at least a single occupied VM. This KPI considers such states to evaluate the system's capability to execute task successfully. Denoted by $\mathcal{C}_t$, the task execution capacity can be computed as

$$\mathcal{C}_t = \mathcal{O}\mu_{\text{MEC}} \sum_{z \in \mathcal{C}_{\text{MEC}}} x_O \tau_{\text{MEC}}(z) + \bar{\mathcal{O}}\mu_{\text{loc}} \sum_{z \in \mathcal{C}_{\text{loc}}} x_O \tau_{\text{loc}}(z). \tag{6.8}$$

Finally, we consider the task execution retainability, which is defined as the probability that a task, once assigned to a VM, will be computed successfully without interruption [149]. In order to evaluate the task execution retainability, let us first define the task execution forced termination rate, denoted by F_v, which represents the ratio between the mean forced termination rate of ongoing tasks and the effective rate in which a new task is assigned to an idle VM, denoted by Λ_v, which equals $\Lambda_v = \lambda_v \left(1 - \sum_{z \in \mathcal{N}_v} \tau_v(z)\right)$. Let $\mathcal{F}_v = \mathcal{C}_v \cup \mathcal{N}_v$ denote all states with at least a single occupied VM and no idle VMs. Tasks that are interrupted in those states, because of VM failures, are dropped. Mathematically, F_v and the task

Table 6.2: Simulation parameters for MEC-enabled dependable task execution.

Parameter	Value
Average number of devices per radio channel (κ)	20
Number of VMs ($\mathcal{V}_{\text{MEC}}$)	5
Number of radio channels (C)	10
Uplink power control threshold (ρ)	-90 dBm
Noise power (σ^2)	-90 dBm
Detection threshold (θ)	-10 dB
Task arrival rate per device (λ_a)	0.15 tasks/ unit time
Single VM execution rate (μ_o)	3 tasks/ unit time
Local execution rate (μ_{loc})	0.1 tasks/ unit time
VM repair rate ($\mathcal{F}$)	1 events/ unit time
VM failure rate ($\mathcal{R}$)	0.1 events/ unit time
VM I/O efficiency (d)	0.1

execution retainability equals

$$F_v = \mathcal{F} \sum_{z \in \mathcal{F}_v} (M_v - x_F)\tau_v(z), \tag{6.9}$$

$$\mathcal{R}_t = \mathcal{O}\left(1 - \frac{F_{\text{MEC}}}{\Lambda_{\text{MEC}}}\right) + \bar{\mathcal{O}}\left(1 - \frac{F_{\text{loc}}}{\Lambda_{\text{loc}}}\right). \tag{6.10}$$

6.2.5 Numerical Results

This subsection aims to numerically evaluate the proposed task execution service dependability KPIs focusing on the studied MEC-enabled network. Unless otherwise stated, the list of involved network parameters are summarized in Table 6.2.

Figure 6.2 shows the offloading success probability as a function of the decoding threshold θ for different active probabilities P_a. The close match between the simulation and the proposed analytical framework validates the analysis and justifies the considered approximations. For increasing values of θ, the offloading success probability decreases due to higher requirement on the link quality. For increasing values of P_a, the rate of task generation at the devices as well as their the probability to utilize the same uplink channel increases, thus network-wide mutual interference increases, hence, leading to lower achievable offloading success probabilities.

Focusing on the discussed KPIs in Section 6.2.4, Figure 6.3 showcases the system's performance for increasing values of θ with different system parameters. Generally, as θ increases, the offloading success probability decreases, thus, owing to the coverage-based offloading criterion, more devices opt to execute their tasks locally. Depending on $\mathcal{O}$, which depends on θ among other parameters, the network oscillates between an offloading-dominant and a local execution-dominant regime. In Figure 6.3(a), we observe that the communication resources availability keeps increasing till a cut-off threshold (i.e., $\theta = -6, -7$ and -8 dB for $\mu_r = 20, 40, 80$, respectively). Operating above these threshold values, the network transitions to the local execution-dominant regime. As μ_r decreases, the gap between the two regimes decreases, since the computation capabilities of the MEC host and device become comparable. Figure 6.3(b) presents the task execution retainability for different per-device task arrival rates. As λ_a

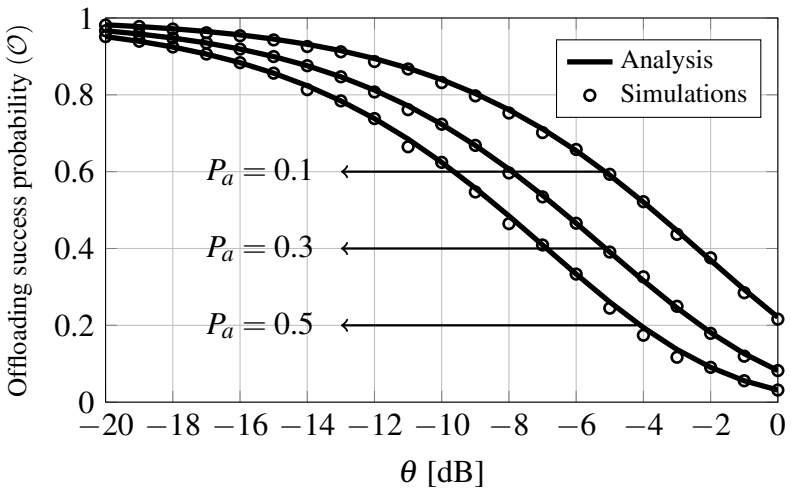


Figure 6.2: OSP model verification.

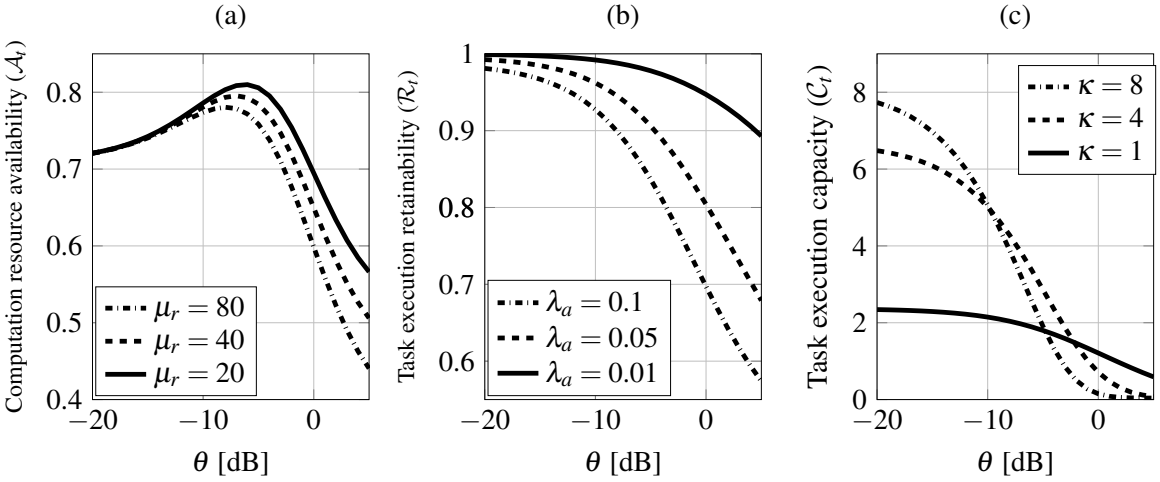


Figure 6.3: Steady state KPIs (a) computation resources availability, (b) task execution retainability, and (c) task execution capacity.

increases, the contention on the radio and the computation resources increases, leading to degradation in the task execution retainability. Figure 6.3(c) shows the task execution capacity for different densification ratios (i.e., average number of devices per BS per channel). In the offloading-dominant regime, high values of task execution capacity are achieved since the offloaded tasks leverage the computationally capable MEC host. However, in the local execution-dominant regime, task execution capacity degrades till it reaches zero. We observe also the effect of κ on the slope steepness of each curve.

The computation resources scalability is investigated via Figure 6.4 which shows the task execution capacity as a function of the number of MEC host VMs $\mathcal{V}_{\text{MEC}}$ and for three different values of the computation degradation factor d. The optimal number of deployed VMs for each value of d, calculated via Algorithm 3, which has a complexity of $O(\mathcal{V}_{\text{MEC}})$, is shown via red circles. It is worth mentioning that the values present in Table 6.2 result in $P_a = 0.25$ and $p = 0.83$. Thus, around 83% of the active devices will offload their generated tasks to the MEC host, thus, operating at the offloading-dominant regime. Nevertheless, due to the I/O interference between the employed VMs at the MEC host, increasing $\mathcal{V}_{\text{MEC}}$ beyond a given value, depending on the value of parameter d, leads to degradation in μ_{MEC} till

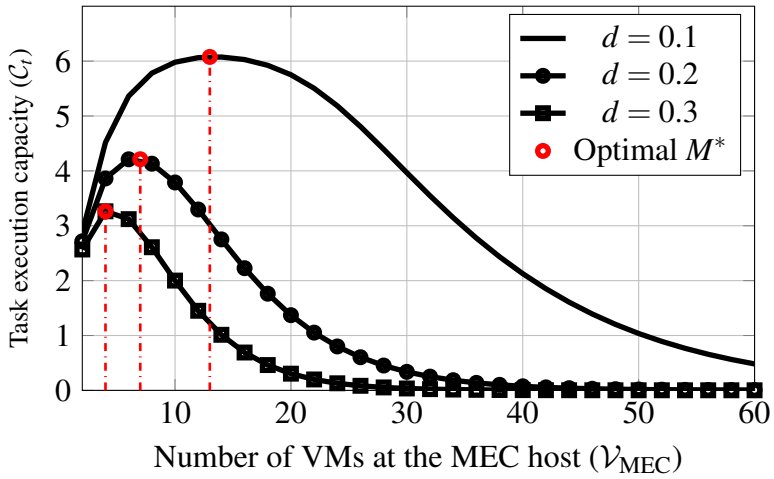


Figure 6.4: Task execution capacity as a function of number of VMs.

Algorithm 3 Optimal number of deployed VMs computation.

procedure $(P_a, \lambda_a, \lambda_b, \lambda_d, \mathcal{V}_{\text{MEC}}, \mathcal{V}_{\text{loc}}, \mu_o, \mu_{\text{loc}}, d, \mathcal{R}, \mathcal{F})$
  Set $m = 1, C(0) = -\infty$, and compute $C(m)$ $\triangleright$ $R(m)$ implies computing C in (6.8) with $\mathcal{V}_{\text{MEC}} = m$.
  while $C(m) > C(m-1)$ **do**
    Compute $C(m)$ from (6.8).
    Increment m.
  end while
  return ${\mathcal{V}_{\text{MEC}}}^* = m$ and $C^* = C({\mathcal{V}_{\text{MEC}}}^*)$.
end procedure

the VM I/O interference dominates and the task execution capacity approaches zero. Such behavior also explains why as d decreases, higher numbers of VMs are desirable. These performance results figure provide network operators with important insights regarding dimensioning the network's infrastructure.

Finally, Figure 6.5 shows the task execution retainability as a function of the repair rate $\mathcal{R}$ for different values of failure rate $\mathcal{F}$. For the extreme case of $\mathcal{F} = 0$, the task execution retainability equals 1, independent of $\mathcal{R}$, since no VM will ever fail. As $\mathcal{F}$ increases, we observe the impact of the repair rate on the task execution retainability, especially within the range $\mathcal{R} \in [0,1]$. For higher values of $\mathcal{R}$, the task execution retainability starts to saturate, owing to its superiority over $\mathcal{F}$, which yields it insignificant with respect to the task execution retainability.

In the following section, we will propose novel definitions of spatial and temporal availability and reliability of a given wireless-based service. Such definitions can be utilized further within different vertical segments (e.g., industrial and automotive).

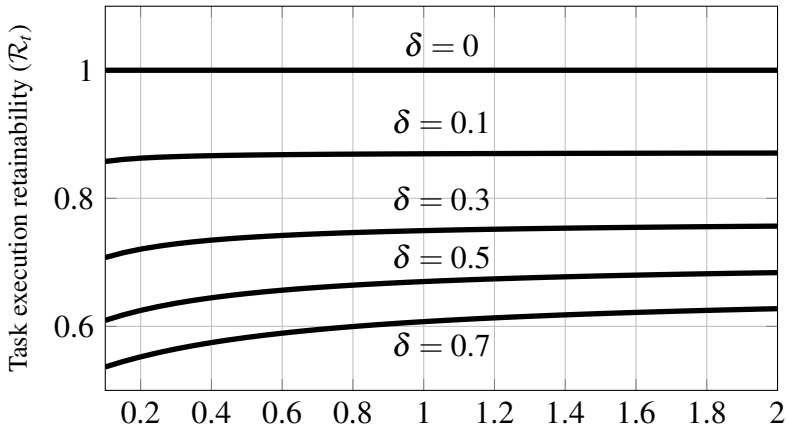


Figure 6.5: Task execution retainability as a function of repair rate.

6.3 Availability and Reliability of Wireless-based Services

6.3.1 Background and Contributions

Complementing on the computation-oriented KPIs adopted in the previous section and due to the challenging service requirements posed by verticals aiming to exploit 5G systems, enhancing existent KPIs and defining new ones is inevitable [150]. Two important performance requirements are the reliability and availability of a running service, which need to be formally understood and quantified. Consequently, a paradigm shift from the conventional network assessment is imminent [139]. Moreover, timely task execution has been mainly evaluated by means of metrics such as packet error ratio, latency and jitter [150]. These metrics, though fundamentally meaningful from the radio communication perspective, need to be looked collectively with the service demands from a vertical-specific point of view (e.g., availability of a service and reliability of its operation). Consequently, such service-specific metrics need to be first well defined, understood and then mapped to the wireless system's parameters, prior to evaluating system-wide feasibility of the focused service/operation. Conceptually, this novel system view aims to unlock the potential of running wireless services quasi-deterministically.

To the best of our knowledge, adopting definitions of service-tailored link availability and reliability for wireless-based systems has not yet been expressed adequately. In [151], the authors propose a new definition of spatial availability, as the ratio of the mean covered area to the geographical area of a given BS. Nevertheless, an interference-free scenario was considered and no insights on the time evolution of communication availability were provided. Additionally, in [152], the authors propose a reliability metric, consisting of two components: the temporal availability and the probability to overcome a received power threshold, however, for a single cell scenario. Furthermore, the authors in [153] summarize main definitions from reliability theory [56], and present an automation-based use case exploiting multi-link connectivity. Nevertheless, the spatial dimension of service availability was not considered at all. In addition, authors in [139], provide a tutorial-like overview, introducing different challenges and solution proposals for URLLC services. Although quite insightful, this work did not touch upon the concepts of time and space availability.

Motivated by the above, in this section, concentrating on both space and time domains, we make a first attempt to bridge the gap between traditional *radio-link* KPIs and *service-level* KPIs by providing an

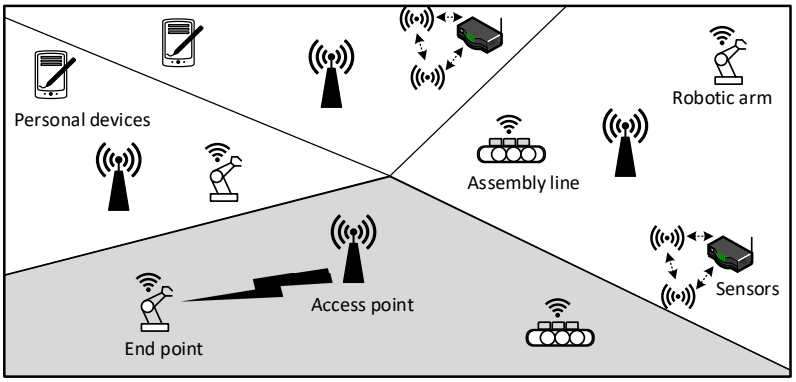


Figure 6.6: System model consisting with heterogeneous devices. The shaded area represents the connectivity region of a given BS.

insight on the availability and reliability of wireless links in 5G systems. The proposed framework aims to indicate which locations in a given area would overcome a performance threshold over a specific time window with a guaranteed level of confidence. In further detail, the contributions of this section are the following; i) proposal the definition of a new, stochastic quantity to measure the spatial availability of a wireless link given a service-specific confidence level, ii) capitalizing on the proposed definition of spatial availability, we present a novel resource allocation approach dependent on the spatially available area of a given BS and based on the concept of resource provisioning, and iii) we present numerical evaluations, highlighting the different effects of system parameter values on spatial availability, as well as on temporal availability and reliability. In addition, we show the relation between spatial and steady state temporal availability.

6.3.2 System Model

A downlink wireless system is considered which consists of BSs that are spatially deployed in $\mathbb{R}^2$ according to a homogeneous PPP, denoted by Φ with intensity λ. Single-antenna BSs of equal transmit power that are deployed over a two-dimensional bounded area $\mathcal{A}_{\text{dep}}$ (e.g., a factory floor). Without loss of generality, the proposed system model can be applied to different communication systems. Over the assumed area, a multitude of devices, like personal tablets, control units, sensors or actuators, are being served via wireless links, as shown in Figure 6.6. Each BS has access to C uplink resources that can be used for data transmission, so as for the devices to fulfill their service requests. At the device side, service requests form an arrival process which follows a Poisson distribution with an average arrival rate denoted by λ_a packets/ unit time, whereas the service time of a service at the BS follows an exponential distribution with an average service rate of μ_{ser} packets/ unit time. A frequency reuse factor of one is assumed in this work, which translates to the potential presence of inter-cell interference among all BSs. From a joint deployment and connectivity point of view, the cell's connectivity region (i.e., the shaded area in Figure 6.6, also named as Voronoi cell) represents the geographical area in which a wireless link can be established between a device and its closest BS. Equivalently, assuming that the path-loss exponent is of the same value over the whole bounded area, Voronoi cells are shaped by applying a BS-device connectivity rule based on a minimum path-loss criterion. Additionally, the SIR of a generic

device located at point i and served by the j-th BS (i.e., $j \in \Phi$) is computed as

$$\mathsf{SIR}_{i,j} = \frac{h_{i,j}||i-j||^{-\eta}}{\sum_{k\in\mathcal{I}_j} g_{i,k}||i-k||^{-\eta}}, \tag{6.11}$$

where P_j is the j-th BS transmit power, $h_{i,j}$, $g_{i,k}, k \in \mathcal{I}_j$ are the signal of interest and interfering channel gains, $||\cdot||$ is the Euclidean distance and $\mathcal{I}_j$ represents the set of all interfering BSs. Additionally, h and g are exponentially distributed with unit gain and are assumed non-correlated among the various links.

One key enabler towards ubiquitous and reliable computation services delivery over wireless links is to investigate the guaranteed performance of a given BS-device link. In this context, we introduce a new binary evaluation metric, $\Omega_{i,j}(\theta,\xi,\Phi)$, having as a decision criterion the probability for a wireless link to achieve a given SIR threshold θ with a predetermined confidence level ξ for a given BS deployment realization Φ.[3] This quantity is mathematically expressed as follows

$$\Omega_{i,j}(\theta,\xi,\Phi) = 1(\mathbb{P}[\mathsf{SIR}_{i,j} \geq \theta] \geq \xi). \tag{6.12}$$

This metric will be exploited in what follows for defining the spatial, service-relevant availability of a wireless link.

6.3.3 Availability Analysis: Spatial Domain

Utilizing the aforementioned metric, we aim now to project the well-established definitions of time-domain availability and reliability to the spatial domain. Temporally, instantaneous availability of a system is the probability of the system being operational at a given time instant [56], whereas, in the space domain, as introduced in [151], the spatial availability $\mathcal{A}_s$, defines the locations on a given Euclidean plane, where the system is operational. The region of operation was modeled as a circular coverage area in [151], due to the lack of interference from the surrounding BSs. We consider a new, service-related definition of spatial availability, taking into account the confidence level of surpassing a predefined SIR threshold. To formalize our contribution, we present the following definition:

(θ,ξ)-availability. *Any device located at i and served by an BS located at j is labeled as (θ,ξ)-available, if $\Omega_{i,j}(\theta,\xi,\Phi) = 1,\ j \in \Phi,\ i,j \in \mathbb{R}^2$, and non-available otherwise.*

Focusing on a given BS deployment Φ, and accumulating all devices possible locations $z; z \in \mathbb{R}^2$ which satisfy the spatial availability criterion $\Omega_{z,j}(\theta,\xi,\Phi)$ when connected to an BS located at point j, we obtain the following (θ,ξ)-available region $\mathcal{D}_j$ as follows

$$\mathcal{D}_j = \{z \in \mathbb{R}^2 | \Omega_{z,j}(\theta,\xi,\Phi) = 1,\ j \in \Phi\}. \tag{6.13}$$

The presence of (θ,ξ)-available region (i.e., $\mathcal{D}_j$) can be physically interpreted as the locations in which a specific QoS (θ) can be achieved with a given confidence level (ξ). The defined QoS can be extended in future work to accommodate further parameters. Accordingly, the spatial availability for an BS located at

[3] It is worth noting that ξ also represents the percentile of devices that achieve $\mathsf{SIR} \geq \theta$ as explained in details in Chapter 4.

j, can be defined as

$$\mathcal{A}_s(j) = \min\Big(1, \frac{\text{Area}(\mathcal{D}_j)}{\text{Area}(\mathcal{V}_j)}\Big) = \min\Big(1, \frac{|\mathcal{D}_j|}{|\mathcal{V}_j|}\Big), \tag{6.14}$$

where the minimum operator accounts for cases where the (θ, ξ)-available area is larger than geographical area of the BS (in such cases $A_s(j) = 1$) and $\mathcal{V}_j$ is the collection of points constituting the Voronoi cell of the j-th BS. In addition, (6.14) can be expanded as

$$\min\Big(1, \frac{\int_{z\in\mathbb{R}^2} 1_{\{z\in\mathcal{D}_j\}}\mathrm{d}z}{\frac{1}{2}|\sum_{\ell}^{g-1} x_\ell y_{\ell+1} + x_q y_q - \sum_{\ell}^{g-1} x_{\ell+1} y_\ell - x_q y_q|}\Big), \tag{6.15}$$

where g is the number of edges and (x_ℓ, y_ℓ) are the Cartesian coordinates of the ℓ-th vertex of the Voronoi cell. The denominator in (6.14) is obtained by applying the well-known shoelace algorithm that computes the area of a Voronoi polygon with g edges [151].

In order to compute the area of set $\mathcal{D}_j$, its boundary needs to be specified. In other words, focusing on an BS, all the points satisfying the SIR threshold θ with confidence level ξ are sought. One can thus expand (6.12), for a given spatial deployment (Φ is dropped for simplicity) as follows

$$\begin{aligned}
\Omega_{j,z}(\theta,\xi,\Phi) &= 1\left\{\mathbb{P}\left\{\frac{h_{j,z}||j-z||^{-\eta}}{\sum_{k\in\mathcal{I}_j} h_{z,k}||k-z||^{-\eta}} \geq \theta\right\} \geq \xi\right\}, \\
&\overset{(a)}{=} 1\left\{\mathbb{E}\left\{\exp\Big(\frac{-\theta}{||j-z||^{-\eta}}\sum_{k\in\mathcal{I}_j} h_{z,k}||k-z||^{-\eta}\right\} \geq \xi\right\}, \\
&\overset{(b)}{=} 1\left\{\prod_{k\in\mathcal{I}_j} \mathbb{M}_h\left\{\frac{-\theta}{||j-z||^{-\eta}} P_k||k-z||^{-\eta})\right\} \geq \xi\right\}, \\
&\overset{(a)}{=} 1\left\{\prod_{k\in\mathcal{I}_j} \frac{1}{1+\frac{\theta}{||j-z||^{-\eta}}||k-z||^{-\eta})} \geq \xi\right\},
\end{aligned} \tag{6.16}$$

where (a) follows since h is exponentially distributed and (b) is the moment generating function (MGF) of an exponentially distributed random variable. Since the BSs transmit with equal power, the final expression of $\Omega_{z,j}(\theta,\xi,\Phi)$ is oblivious to the transmission power [21].

Expression (6.16) can be utilized to define the boundary of Area($\mathcal{D}_j$) by substituting inequality by pure equality. However, since this boundary is hardly tractable in closed form, to obtain quantitative results, we resort to a bisection-based algorithm to approximate the size of this area to be then used in (6.14). A visualization of the computed regions is shown in Figure 6.7, where different combinations of (θ,ξ) are considered for a given BS. It is observable that in Figure 6.7(b), the $(-10\text{ dB}, 0.8)$ region is not convex, due to the interference imposed by the closest interfering BS, which reveals that, along with (θ,ξ), the number and location of deployed and, consequently, interfering BSs is expected to highly affect the spatial availability. A thorough investigation on the effect of (θ,ξ) on spatial availability for random BS deployments will be presented in subsection 5.3.5.

As one would expect,the selected parameter values highly affect the achieved spatial availability, thus, to ensure an average insight over all possible BS locations, spatial averaging over a large number of

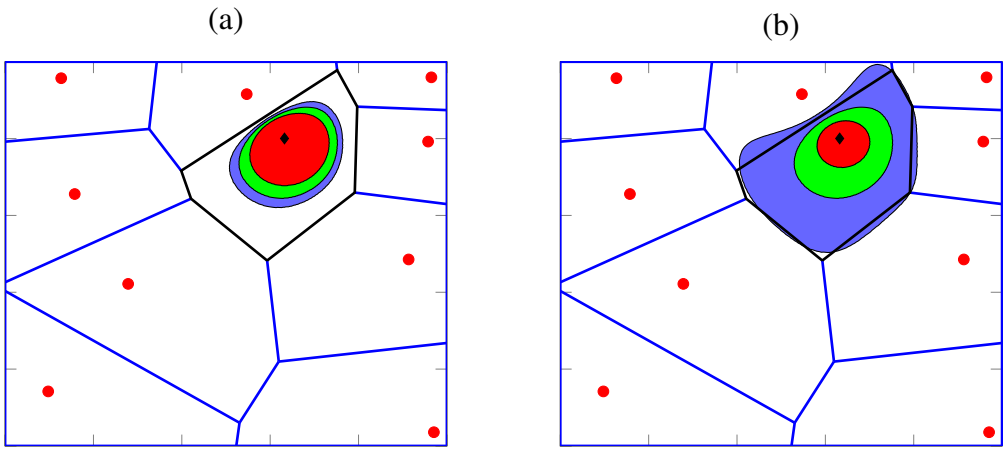


Figure 6.7: (θ,ξ,Φ)-available regions for a deployment with $\Phi = 0.1$ focusing on a generic BS (its Voronoi border drawn in black) (a) $\theta = 0$ dB and $\xi = (0.7, 0.8, 0.9)$ and (b) $\theta = (-10, 0, 10)$ dB and $\xi = 0.8$.

deployments was conducted. In Figure 6.8, we highlight the effect of parameters θ and ξ along with the BS's intensity on the spatial availability. First, in Figure 6.8(a), the spatial availability $\mathcal{A}_s$ of a randomly selected BSs is plotted as a function of θ for different confidence levels, ξ. As expected, for increasing values of θ (or ξ), the spatial availability of that BS decreases, as the equivalent (θ,ξ)-available region reduces.

Second, in Figure 6.8(b), $\mathcal{A}_s$ is shown as a function of λ over the fixed deployment area for two different confidence levels, when θ=0 dB. The monotonically increasing fashion of $\mathcal{A}_s$ as a function of Φ for a given value of ξ is explained as follows: as the system becomes more densified with BSs, the Voronoi area of each BS decreases, since the BSs become geographically closer. Also the accompanying (θ,ξ)-available region shrinks, due to the higher interference received from other BSs. However, the latter region is less affected compared to the former, due to the stochastic nature of the region forming criterion together with the applied bisection-based approach for computing the (θ,ξ)-available regions.

6.3.4 Availability Analysis: Temporal Domain

Having analyzed the spatial availability metric in the previous section, and since our objective is to propose a unified, space-time availability framework useful to URLLC systems, in this section we concentrate on the time domain. As explained in Section 6.2.2, each BS has $\mathcal{R}_t$ channels that can be accessible by the devices in its Voronoi region. To account for the spatial availability $\mathcal{A}_s$ as defined in subsection 6.3.3, we propose a spatial availability-proportional channel allocation scheme. According to this scheme, since $\mathcal{A}_s$ decomposes the Voronoi region of an BS located at point $\mathbf{j}$ into two regions, the number of channels to be utilized by devices located in the (θ,ξ)-available and non-available regions of this device can be, respectively, written as

$$C_a(j) = \lceil \mathcal{A}_s(j)C \rceil,\ C_n(j) = C - C_a(j). \tag{6.17}$$

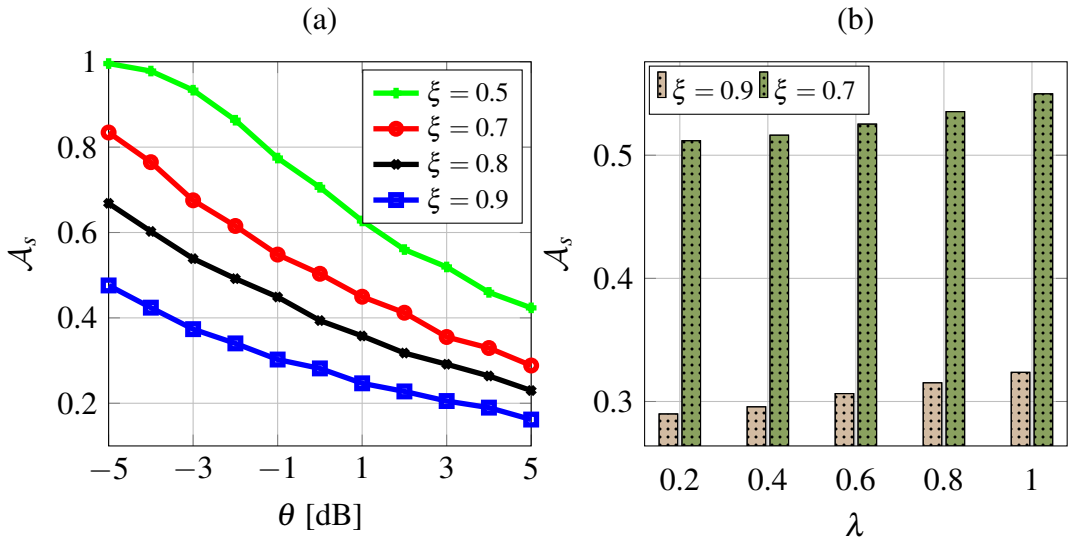


Figure 6.8: Spatial availability as a function of (a) (θ, ξ) parameters (b) number of BSs for $\theta = 0$ dB.

As a result of the proposed policy, assuming that service requests arrive uniformly in space, when the spatial availability ratio $\mathcal{A}_s$ is low, a few channels will be allocated to the few evolving requests coming from the (θ, ξ)-available region, while, the majority of channels will be allocated to the (possibly many) requests coming from the (θ, ξ) non-available region.

In order to model the resources status at a generic BS over time, we resort to a CTMC model that captures the number of idle/ occupied channels as time evolves. To also capture the decomposition of service requests into two sets (i.e., coming from spatially available/ non-available areas), a two dimensional CTMC is utilized, where one dimension represents the number of devices being served within the (θ, ξ)-available region of the BS, and the other dimension represents the number of devices in the rest of the coverage region. Figure 6.9 visualizes the proposed framework, where n_a and n_n are generic numbers of devices being served in the two mentioned regions. Such a model leads to a finite birth/ death Markov process, where the total number of states is limited by the total number of channels and all possible partitioning options. Thus, for a given channel allocation, the set of feasible states is represented as

$$\mathcal{Q} = \{(n_a, n_n) |\, 0 \leq n_a \leq C_a,\, 0 \leq n_n \leq C_n,\, C_a + C_n = C\}, \tag{6.18}$$

where the total number of states is $|\mathcal{Q}| = (C_a + 1)(C_n + 1)$, as a number of n channels will lead to $n+1$ states. Based on the above described model, the temporal availability is defined as the probability of at least one channel being available for a new request. As a result, the set of temporally available states for the (θ, ξ)-available and non-available regions are $\mathcal{A}_a = \{(n_a, n_n), n_a = \{0, 1, \cdots, C_a - 1\}\}$ and $\mathcal{A}_n = \{(n_a, n_n), n_n = \{0, 1, \cdots, Cn - 1\}\}$, respectively. The state equations can be vectorized as $\tau(t) = \{\tau_1(t), \tau_2(t), \cdots, \tau_{|\mathcal{Q}|}(t)\}$, where $\tau_\ell(t)$ is the probability of the system being in the ℓ-th state at time instant t. Resorting to the matrix notation and using the Kolmogorov forward equations [56], the state probabilities can be computed by solving the following equation

$$\frac{\mathrm{d}}{\mathrm{d}t}\tau(t) = \tau(t)\mathbf{Q}, \tag{6.19}$$

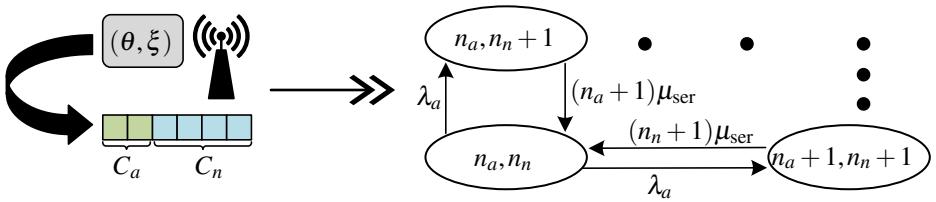


Figure 6.9: Resource partitioning based on $\mathcal{A}_s$ along with part of the two dimensional birth/death Markov process.

where the infinitesimal generator (i.e., transition rate) matrix is denoted by $\mathbf{Q}$, with dimension $|\mathcal{Q}| \times |\mathcal{Q}|$. To compute the system's temporal availability, one needs to solve (6.19). We adopted a similar approach as in [152], based on the uniformization method [154], where the solution of (6.19), for a given initial state probability (i.e., $t = 0$), denoted by $\tau(0)$, can be rewritten as

$$\begin{aligned}\tau(t) &= \tau(0)\mathrm{e}^{\mathbf{Q}t} = \tau(0)\sum_{i=0}^{\infty}\frac{(\mathbf{Q}t)^i}{i!},\\ &\overset{(a)}{=} \tau(0)\mathrm{e}^{-qt}\sum_{i=0}^{\infty}\frac{(qt)^i}{i!}\mathbf{J}^n,\end{aligned} \tag{6.20}$$

where (a) follows from the introduction of $\mathbf{J} = \mathbf{I}_{|\mathcal{Q}|\times|\mathcal{Q}|} + \frac{1}{q}\mathbf{Q}$, $\mathbf{I}_{|\mathcal{Q}|\times|\mathcal{Q}|}$ is the identity matrix and q is a number satisfying $q \geq \max(q_{ii})$, where q_{ii} are the diagonal elements of $\mathbf{Q}$. To numerically solve (6.20), the summation must be truncated at level N_c as shown in [152]. In order to obtain the temporal availability of a generic BS at a given time instant t, one needs to consider all the available states as follows

$$\mathcal{A}_t{}^u(t) = \sum_{i\in\mathcal{A}_u}\tau_i(t),\ u \in \{a,n\}, \tag{6.21}$$

where index $u \in \{a,n\}$ represents the (θ,ξ)-available and non-available regions, respectively.

Reliability Analysis

Another important metric for the temporal analysis is the system's temporal reliability $R(t)$ [153], which is defined as the probability that the system is operational during time interval $[0,t]$. Such a definition can be employed in the studied CTMC model, by forcing the system to remain in an unavailable state once it reaches one. In other words, the transition rate from any unavailable state is set to zero. Such a modification leads to a modified infinitesimal generator matrix $\hat{\mathbf{Q}}$ and (6.19) can be re-expressed as $\frac{\mathrm{d}}{\mathrm{d}t}\hat{\tau}(t) = \hat{\tau}(t)\hat{\mathbf{Q}}$, where $\hat{\tau}(t)$ corresponds to state probability of the modified CTMC. Accordingly, the system's temporal reliability is computed as

$$\mathcal{R}_t{}^u(t) = \sum_{i\in\mathcal{A}_u}\hat{\tau}_i(t),\ u \in \{a,n\}. \tag{6.22}$$

As it will be numerically shown later, the system reliability is always upper bounded by its time availability (i.e., $\mathcal{A}_t(t) \geq \mathcal{R}_t(t)$), since for a repairable system, transition rates from a failed state are non-zero.

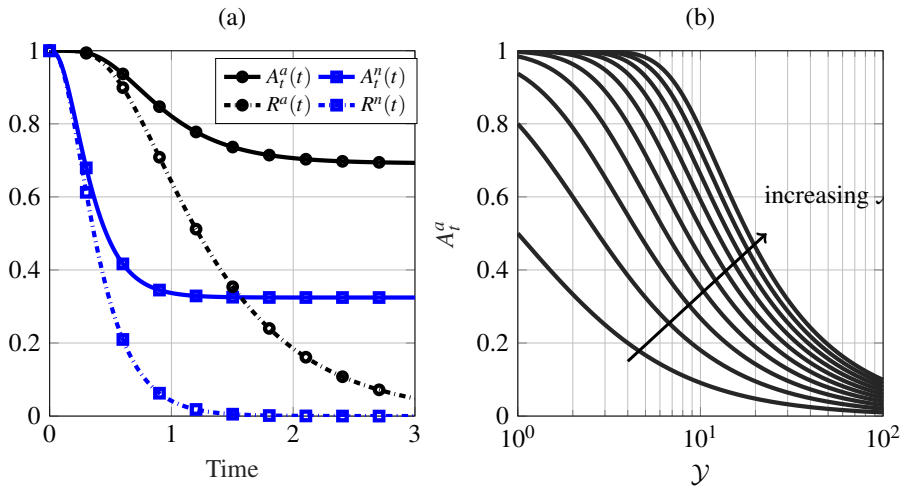


Figure 6.10: (a) Temporal availability and reliability of (θ,ξ)-available and non-available regions for $\mathcal{A}_s = 0.7$ (b) Steady state time availability for increasing values of spatial availability ($\mathcal{A}_s$) ranging from 0.1 to 1.

Steady State Analysis

Another interesting metric relevant to temporal analysis is the steady state time availability, which is time independent and can be interpreted as the average operating time [56]. Mathematically [153], it can be represented as

$$\mathcal{A}_t{}^u = \lim_{t\to\infty} \mathcal{A}_t{}^u(t) = \sum_{i\in\mathcal{A}_u} \tau_i = \sum_{i\in\mathcal{A}_u} \frac{\mathcal{Y}}{i!}\Big(1+\sum_{l=1}^{C_u} \frac{\mathcal{Y}^l}{l!},\Big), \tag{6.23}$$

where $u \in \{a,n\}$ and $\mathcal{Y} = \frac{\lambda_a}{\mu_{\text{ser}}}$ represents the arrival to service rate ratio. Based on the presented metrics, we investigate, in what follows, the temporal availability and reliability for the proposed access scheme. In Figure 6.10(a), the system's transient analysis is presented for $\mathcal{A}_s = 0.7$. Due to the spatially-dependent channel allocation proposed in (6.17), the time availability, $\mathcal{A}_t{}^a(t)$ (reliability $\mathcal{R}_t{}^a(t)$) for a request originating from the (θ,ξ)-available region should be higher than the time availability $\mathcal{A}_t{}^n(t)$ (reliability $R^n(t)$) of the (θ,ξ)-non available region. This is explained due to the larger number of channels that can be utilized for the spatially available region. It is noticeable that at $t = 0$, all channels are available, thus, leading to time availability and reliability equal to one. Additionally, numerical results confirm that the time reliability is upper bounded by time availability, as well as that such a bound is time-dependent since it loosens over time till a maximum performance gap is reached which is then fixed as time evolves.

In Figure 6.10(b), the steady state analysis is illustrated for varying values of the arrival to service ratio $\mathcal{Y}$. As $\mathcal{Y}$ increases, the steady state temporal availability decreases; this occurs due to the fact that the available channels are less in such regimes. Also, as explained earlier, as a result of the adopted channel access scheme, larger values of spatial availability lead to higher time availability. It is, therefore, concluded that $\mathcal{Y}$ is a fundamental performance limitation factor, as for extremely large values of it, even a 100% spatial availability is unable to be translated to high time availability.

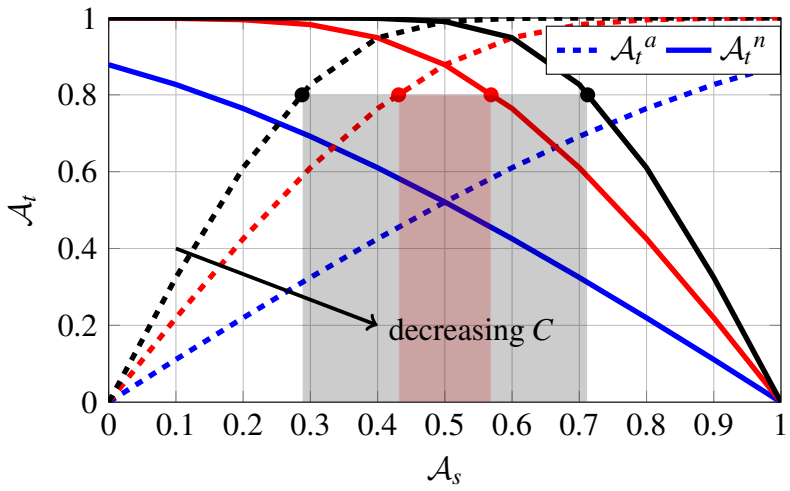


Figure 6.11: Steady state time availability as a function of spatial availability for varying numbers of channels (C = (10, 20, 30)).

6.3.5 Spatiotemporal Joint Analysis

Finally, in Figure 6.11, the relationship between the steady-state time availability and the spatial availability is presented for different total channel numbers, C. First, we observe a symmetric time availability performance for a fixed $\mathcal{R}_t$ between the (θ,ξ)-available and non-available regions, due to the proposed channel allocation scheme. For $\mathcal{A}_s = 0.5$, the number of channels allocated to each region will be the same, hence, leading to an identical time availability performance, as requests arrive uniformly in space. Additionally, fixing the value of $\mathcal{A}_s$, time availability increases together with C. This result intuitively emphasizes the role of redundancy and provisioning in wireless systems. Equivalently, through our proposed space-time analysis, the minimum total number of channels needed to achieve a targeted temporal availability level can be identified. To further highlight this, a steady state time availability requirement of 0.8 is marked for $NumberULchannels = 20$ and $NumberULchannels = 30$ curves. As expected, the range of $\mathcal{A}_s$ meeting the imposed requirement is larger in the latter case. This means that, a sufficient amount of resources can guarantee the time availability performance of multiple service classes.

6.4 Conclusion

Focusing on successful task execution and the availability of a given service within wireless-based systems, in this chapter, novel definitions of dependability attributes for communication and computation services are provided. First, a spatiotemporal framework is proposed to characterize the network-wide task completion from a dependability perspective considering that devices employ a coverage-based offloading criterion. Modeling tools are utilized to derive closed form expressions of the offloading success probability and a number of novel defined task execution dependability-relevant metrics, such as communication resources availability, task execution retainability and task execution capacity. To yield the framework practical, models of VMs failures and repairing are considered. Numerical results showcase regimes where the system transitions from the offloading-dominant to the local execution-

dominant regime. Different system parameters such as task arrival rate, device spatial density and VM computation capabilities, are presented to obtain a complete understanding of system behavior. Finally, we show that assuming a given system parameterization, there exists an optimal number of VMs, which, when deployed, maximizes the task execution capacity.

Additionally, we present a unified framework characterizing the temporal and spatial availability for a service-agnostic wireless-based system. A novel, service-relevant definition of spatial availability is introduced taking into account the probability to achieve a targeted SIR threshold with a given confidence level. Temporal availability is investigated considering a novel, space availability-driven channel access scheme based on the concept of channel provisioning, bringing up the coupled relation between spatial and temporal availability and reliability. The study is supported by numerical evaluation results which underline the impact of different system parameter values on space/ time availability and time reliability, as well as the coupled nature of these metrics.

Chapter 7

Conclusion and Future Work

This chapter summarizes the technical contents presented throughout the previous chapters and discusses future research directions. In Section 7.1, the main contributions and key observations and conclusions will be highlighted. Possible future research directions that are building on this thesis will be presented to finalize this dissertation in Section 7.2 .

7.1 Conclusion

Throughout this thesis, large scale IoT networks were modeled and studied for different use cases and network deployment variants. The thesis was based on two main pillars: namely, the communication and computation pillars. The motivation, background and a tutorial-style preview of the mathematical tools and key enablers that are utilized throughout this thesis were discussed in **Chapters 1** and **2**, respectively. To provide a contained summary of the technical chapters, our findings can be summarized as follows:

- **Communication pillar**: Utilizing stochastic geometry and queueing theory to characterize the macroscopic (i.e., network-wide mutual interference) and microscopic (i.e., device traffic dynamics) scales of the network, we presented different spatiotemporal frameworks that aim to analyze large scale uplink IoT networks. In details, **Chapter 3** investigated the co-existence of prioritized multi-stream traffic generated at the device side via a tractable and scalable vacation-based model. The developed spatiotemporal model was used afterwards to assess and compare three priority aware channel allocation strategies; namely dedicated-equal allocation, dedicated weighted allocation and shared allocation strategy. A major key result of this chapter was the reported superiority of the shared channel allocation strategy over the dedicated ones, as the former offers higher pool of channels, enabling interference diversification. In addition, various KPIs, such as transmission success probability, average queue length, average-delay, delay distribution, and PAoI were derived and discussed for each priority class. Leveraging the presented framework, interesting insights regarding heterogeneous traffic co-existence, QoS requirements and cost of traffic prioritization were presented.

  Focusing on timely status updates and information freshness within large scale uplink IoT networks, we focused in **chapter 4** on the impact of time-triggered and event-triggered traffic from the PAoI perspective. In contrast to the spatially averaged performance, we leveraged tools from stochastic

geometry to analyze the location-dependent performance, taking into consideration the impact of the aforementioned traffic models. In essence, we showed that both the time-triggered and event-triggered traffic models can be captured via a unified queueing analysis, However, in large scale networks, both traffic patterns lead to different network-wide mutual interactions between the coexisting IoT devices, which was captured via our proposed framework. The presented results unveiled the counter-intuitive superiority of event-triggered traffic over the time-triggered traffic, when it comes to the PAoI. Such a result was attributed to the higher temporal interference correlations of the time-triggered traffic. By virtue of the spatiotemporal perspective, the Pareto frontiers that characterize stable operation of the devices within the network were derived and discussed.

- **Edge-computing pillar**: Leveraging MEC deployment within the large scale IoT networks, advents of MEC gains, such as reduced experienced latency within heterogeneous and vehicular networks, dependable task execution and more computation power, were the highlight of our research within this pillar. In **Chapter 5**, the problem of cell association was studied for heterogeneous networks, that entailed radio and computation disparity between the different tiers. It was shown that, when compared to state-of-the-art association criterion, the proposed rule offers up till 60% reduction in the experienced one-way latency. Our main finding in this chapter was the need to have an adaptive association criterion that takes not only the radio aspect, but also the computation perspective, when designing advanced cell association rules. In addition, improving the timeliness of collective road awareness via MEC deployment, concentrating on the vehicular VRU use case was investigated. It was shown that the proposed overlaid deployment of MEC hosts offers up to 80% average latency reduction, as compared to the conventional network architecture. Furthermore, for a number of network parametrization, the network-wide PAoI of the conventional system architecture can be reduced by nearly 61% when a MEC-enabled architecture is deployed.

  In addition, dependable and ubiquitous computing services within wireless networks were considered in **Chapter 6**, to understand the feasibility of wireless-based networks to meet the futuristic requirements of future services. A novel spatiotemporal framework was presented that utilizes stochastic geometry and continuous time Markov chains to jointly characterize the communication and computation performance of MEC-enabled wireless systems, while considering the influence of various system parameters on dependability metrics such as (i) computation resources availability, (ii) task execution retainability, and (iii) task execution capacity. We showed that there exists an optimal number of virtual machines for parallel computing at the MEC host to maximize the task execution capacity. Additionally, since wireless links are characterized by fluctuating quality, we provided a first attempt to quantify the spatial and temporal availability of a service and its reliability via utilizing tools from dependability theorem. In the space domain, we characterized spatially available areas consisting of all locations that meet a performance requirement with given confidence. In the time domain, we proposed a channel allocation scheme accounting for the spatial availability of a given cell. With the aim to reveal the incurred space-time performance trade-offs, numerical results were presented. In addition, the effect of different system parameters on the achievable service availability and reliability were discussed and highlighted.

7.2 Future work

Owing to the conducted analysis throughout this thesis, different future research directions are envisioned. From the spatiotemporal modeling, future and current deployment scenarios create complex topological structures, especially from the devices point of view, that cannot be captured by the adopted PPP. Accordingly, there is an utter need to develop spatiotemporal models that utilize other point processes. In addition, since equal resource allocation was considered throughout this thesis, advanced resource allocation schemes, such as, Markov decision processes, can be utilized in order to provide smarter and efficient resource utilization among the active devices with different QoS requirements. Further considerations of different queueing disciplines, rather than the FCFS, might be interesting for specific applications (e.g., last come first serve in sensors measurement reporting).

Moreover, the trend of technology evolution is towards more intelligent services and applications within the network, which will require a more reliable, efficient, resilient, and secure connectivity. When the connected objects are more intelligent it becomes difficult to deal with their complexity by using the communication network in a static, simplistic and rigid manner. To fully address this foreseeable shift, the employment of machine learning and artificial intelligence, leveraging the computation power provided by MEC deployment, is attractive from cost, enhanced QoS and scalability perspectives. Research directions regarding the impact of network deployments and network architecture on the effectiveness of distributed learning is gaining attention. Under the same scope, federated learning refers to the notion of multiple network nodes training a shared model in a distributed fashion. The main idea is that user devices in a network will collaboratively learn a shared prediction model without the raw training data leaving the device. This is motivated from several perspectives; limitations in uplink bandwidth, limitations in network coverage, and restrictions in transferring privacy-sensitive data across the network.

Finally, several vertical segments (e.g., V2X and industrial automation) are characterized by stringent requirements that require novel methods to assure the determinism of operation. Further research on mechanisms to achieve high dependability, reliability, availability and liability, is necessary. As a crucial prerequisite for such service-critical segments, it is important to include an in-depth analysis of the potentials and limitations of cross-layer technologies such as communication-control co-design and more involved spatiotemporal network design. The dependability-based framework presented in this thesis serves as a starting point to more in-depth research possibilities in this timely and technologically challenging domain.

Bibliography

[1] "Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2016-2021," tech. rep., http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.pdf, 2017.

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

[3] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on Multi-Access Edge Computing for Internet of Things Realization," *IEEE Communications Surveys Tutorials*, vol. 20, no. 4, pp. 2961–2991, 2018.

[4] 3GPP, "TR 36.746 Study on Further Enhancements to LTE Device to Device, User Equipment to Network Relays for Internet of Things and Wearables," *3rd Generation Partnership Project (3GPP), v15.1.1*, 2018.

[5] M. Bergés and C. Samaras, "A Path Forward for Smart Cities and IoT Devices," *IEEE Internet of Things Magazine*, vol. 2, no. 2, pp. 2–4, 2019.

[6] A. Al-Fuqaha, M. Guizani, M. Mohammadi, M. Aledhari, and M. Ayyash, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, pp. 2347–2376, Fourth quarter 2015.

[7] P. Mach and Z. Becvar, "Mobile Edge Computing: A Survey on Architecture and Computation Offloading," *IEEE Communications Surveys Tutorials*, vol. 19, pp. 1628–1656, third quarter 2017.

[8] M. R. Palattella, M. Dohler, A. Grieco, G. Rizzo, J. Torsner, T. Engel, and L. Ladid, "Internet of Things in the 5G Era: Enablers, Architecture, and Business Models," *IEEE Journal on Selected Areas in Communications*, vol. 34, pp. 510–527, March 2016.

[9] Cisco, "Cisco Annual Internet Report (2018–2023)," *White Paper*, 2020.

[10] Q. Pham, F. Fang, V. N. Ha, M. J. Piran, M. Le, L. B. Le, W. Hwang, and Z. Ding, "A Survey of Multi-Access Edge Computing in 5G and Beyond: Fundamentals, Technology Integration, and State-of-the-Art," *IEEE Access*, vol. 8, pp. 116974–117017, 2020.

[11] R. Li, "Towards a New Internet for the Year 2030 and Beyond," *ITU IMT-2020/5G Workshop*, July 2018.

[12] 3GPP, "System Architecture for the 5G System (5GS)," Technical Specification (TS) 23.501 v15.9.0, 3rd Generation Partnership Project (3GPP), Mar. 2020.

[13] NGMNA, "Recommendations for NGMN KPIs and Requirements for 5G," *Next Generation Mobile Networks Alliance*, 2016.

[14] 3GPP, "NR and NG-RAN Overall description; Stage-2," Technical Specification (TS) 38.300 v16.2.0, 3rd Generation Partnership Project (3GPP), Jul. 2020.

[15] B. P. Crow, I. Widjaja, J. G. Kim, and P. T. Sakai, "IEEE 802.11 Wireless Local Area Networks," *IEEE Communications Magazine*, vol. 35, no. 9, pp. 116–126, 1997.

[16] G. Fettweis and S. Alamouti, "5G: Personal mobile internet beyond what cellular did to telephony," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 140–145, 2014.

[17] A. D. Wyner, "Shannon-theoretic Approach to a Gaussian Cellular Multiple-access Channel," *IEEE Transactions on Information Theory*, vol. 40, no. 6, pp. 1713–1727, 1994.

[18] K. S. Gilhousen, I. M. Jacobs, R. Padovani, A. J. Viterbi, L. A. Weaver, and C. E. Wheatley, "On the Capacity of a Cellular CDMA System," *IEEE Transactions on Vehicular Technology*, vol. 40, no. 2, pp. 303–312, 1991.

[19] M. Alouini and A. J. Goldsmith, "Area Spectral Efficiency of Cellular Mobile Radio Systems," *IEEE Transactions on Vehicular Technology*, vol. 48, no. 4, pp. 1047–1066, 1999.

[20] V. H. M. Donald, "Advanced Mobile Phone Service: The Cellular Concept," *Bell System Technical Journal*, vol. 58, no. 1, pp. 15–41, 1979.

[21] J. G. Andrews, F. Baccelli, and R. K. Ganti, "A Tractable Approach to Coverage and Rate in Cellular Networks," *IEEE Transactions on Communications*, vol. 59, pp. 3122–3134, Nov. 2011.

[22] M. Haenggi, *Stochastic Geometry for Wireless Networks*. New York, NY, USA: Cambridge University Press, 2012.

[23] H. ElSawy, A. Sultan-Salem, M. Alouini, and M. Z. Win, "Modeling and Analysis of Cellular Networks Using Stochastic Geometry: A Tutorial," *IEEE Communications Surveys Tutorials*, vol. 19, pp. 167–203, Firstquarter 2017.

[24] M. Haenggi, J. G. Andrews, F. Baccelli, O. Dousse, and M. Franceschetti, "Stochastic Geometry and Random Graphs for the Analysis and Design of Wireless Networks," *IEEE Journal on Selected Areas in Communications*, vol. 27, no. 7, pp. 1029–1046, 2009.

[25] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume I - Theory*, vol. 1 of *Foundations and Trends in Networking Vol. 3: No 3-4, pp 249-449*. NoW Publishers, 2009. Stochastic Geometry and Wireless Networks, Volume II - Applications; see http://hal.inria.fr/inria-00403040.

[26] F. Baccelli and B. Blaszczyszyn, *Stochastic Geometry and Wireless Networks, Volume II - Applications*, vol. 2 of *Foundations and Trends in Networking: Vol. 4: No 1-2, pp 1-312*. NoW Publishers, 2009. Stochastic Geometry and Wireless Networks, Volume I - Theory; see http://hal.inria.fr/inria-00403039.

[27] H. ElSawy, E. Hossain, and M. Haenggi, "Stochastic Geometry for Modeling, Analysis, and Design of Multi-Tier and Cognitive Cellular Wireless Networks: A Survey," *IEEE Communications Surveys Tutorials*, vol. 15, pp. 996–1019, Third 2013.

[28] R. G. Gallager, *Discrete Stochastic Processes*, vol. 321. Springer US, January 1996.

[29] A. S. Alfa, *Applied Discrete-time Queues*. Springer-New York USA, 01 2015.

[30] N. Jiang, Y. Deng, X. Kang, and A. Nallanathan, "Random Access Analysis for Massive IoT Networks Under a New Spatio-Temporal Model: A Stochastic Geometry Approach," *IEEE Transactions on Communications*, vol. 66, pp. 5788–5803, Nov 2018.

[31] G. Latouche and V. Ramaswami, *Introduction to Matrix Analytic Methods in Stochastic Modeling*. Society for Industrial and Applied Mathematics, 1999.

[32] R. R. Rao and A. Ephremides, "On the Stability of Interacting Queues in a Multiple-access System," *IEEE Transactions on Information Theory*, vol. 34, pp. 918–930, Sep. 1988.

[33] Wei Luo and A. Ephremides, "Stability of *N* Interacting Queues in Random-access Systems," *IEEE Transactions on Information Theory*, vol. 45, pp. 1579–1587, July 1999.

[34] Y. Zhong, T. Q. S. Quek, and X. Ge, "Heterogeneous Cellular Networks With Spatio-Temporal Traffic: Delay Analysis and Scheduling," *IEEE Journal on Selected Areas in Communications*, vol. 35, pp. 1373–1386, June 2017.

[35] M. Gharbieh, H. ElSawy, A. Bader, and M. Alouini, "Spatiotemporal Stochastic Modeling of IoT Enabled Cellular Networks: Scalability and Stability Analysis," *IEEE Transactions on Communications*, vol. 65, pp. 3585–3600, Aug 2017.

[36] M. Gharbieh, H. ElSawy, H. Yang, A. Bader, and M. Alouini, "Spatiotemporal Model for Uplink IoT Traffic: Scheduling and Random Access Paradox," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 8357–8372, Dec 2018.

[37] M. Gharbieh, H. ElSawy, M. Emara, H.-C. Yang, and M.-S. Alouini, "Grant-free opportunistic uplink transmission in wireless-powered iot: A spatio-temporal model," *IEEE Transactions on Communications*, vol. 69, no. 2, pp. 991–1006, 2021.

[38] H. H. Yang and T. Q. S. Quek, "Spatiotemporal Analysis for SINR Coverage in Small Cell Networks," *IEEE Transactions on Communications*, pp. 1–1, 2019.

[39] G. Chisci, H. ElSawy, A. Conti, M. Alouini, and M. Z. Win, "Uncoordinated Massive Wireless Networks: Spatiotemporal Models and Multiaccess Strategies," *IEEE/ACM Transactions on Networking*, 2019.

[40] Z. Chen, N. Pappas, M. Kountouris, and V. Angelakis, "Throughput With Delay Constraints in a Shared Access Network With Priorities," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 5885–5899, Sep. 2018.

[41] P. S. Dester, P. Cardieri, P. H. J. Nardelli, and J. M. C. Brito, "Performance Analysis and Optimization of a *N*-Class Bipolar Network," *IEEE Access*, vol. 7, pp. 135118–135132, 2019.

[42] F. Benkhelifa, H. ElSawy, J. A. Mccann, and M. Alouini, "Recycling Cellular Energy for Self-Sustainable IoT Networks: A Spatiotemporal Study," *IEEE Transactions on Wireless Communications*, vol. 19, no. 4, pp. 2699–2712, 2020.

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

[44] W. Saad, M. Bennis, and M. Chen, "A Vision of 6G Wireless Systems: Applications, Trends, Technologies, and Open Research Problems," *IEEE Network*, pp. 1–9, 2019.

[45] K. Kim and P. R. Kumar, "Cyber–Physical Systems: A Perspective at the Centennial," *Proceedings of the IEEE*, vol. 100, pp. 1287–1308, May 2012.

[46] A. A. et al, "Internet of Things: A Survey on Enabling Technologies, Protocols, and Applications," *IEEE Communications Surveys Tutorials*, vol. 17, pp. 2347–2376, Fourthquarter 2015.

[47] E. Soltanmohammadi, K. Ghavami, and M. Naraghi-Pour, "A Survey of Traffic Issues in Machine-to-Machine Communications Over LTE," *IEEE Internet of Things Journal*, vol. 3, pp. 865–884, Dec 2016.

[48] S. Kaul, R. Yates, and M. Gruteser, "Real-time status: How Often Should One Update?," in *2012 Proceedings IEEE INFOCOM*, pp. 2731–2735, March 2012.

[49] R. D. Yates and S. K. Kaul, "The Age of Information: Real-Time Status Updating by Multiple Sources," *IEEE Transactions on Information Theory*, vol. 65, pp. 1807–1827, March 2019.

[50] R. Talak, S. Karaman, and E. Modiano, "Can Determinacy Minimize Age of Information?," *CoRR*, vol. abs/1810.04371, 2018.

[51] E. T. Ceran, D. Gündüz, and A. György, "Average Age of Information with Hybrid ARQ under a Resource Constraint," in *2018 IEEE Wireless Communications and Networking Conference (WCNC)*, April 2018.

[52] L. Huang and E. Modiano, "Optimizing Age-of-Information in a Multi-class Queueing System," in *2015 IEEE International Symposium on Information Theory (ISIT)*, pp. 1681–1685, June 2015.

[53] C. Xu, H. H. Yang, X. Wang, and T. Q. S. Quek, "Optimizing Information Freshness in Computing enabled IoT Networks," *IEEE Internet of Things Journal*, pp. 1–1, 2019.

[54] S. Bagchi, M. Siddiqui, P. Wood, and H. Zhang, "Dependability in Edge Computing," *Communications of the ACM*, vol. 63, no. 1, pp. 58–66, 2020.

[55] Begleitforschung zur zuverlässigen, drahtlosen Kommunikation in der Industrie Fachgruppe 1, *Aspects of Dependability Assessment in ZDKI*, 2016.

[56] A. Birolini, *Reliability Engineering: Theory and Practice*. Springer, 2010.

[57] ETSI, "Multi-access Edge Computing (MEC); Terminology," Group Specification (GS) 001 v2.1.1, European Telecommunications Standards Institute (ETSI), Jan. 2019.

[58] Ericsson, "Edge computing and 5G: Harnessing the distributed cloud for 5G success," white paper, Jun. 2019.

[59] S. Kekki, W. Featherstone, Y. Fang, and P. Kuure, "MEC in 5G networks," *ETSI, White Paper*, 2018.

[60] M. C. Filippou, D. Sabella, M. Emara, S. Prabhakaran, Y. Shi, B. Bian, and A. Rao, "Multi-Access Edge Computing: A Comparative Analysis of 5G System Deployments and Service Consumption Locality Variants," *IEEE Communications Standards Magazine*, vol. 4, no. 2, pp. 32–39, 2020.

[61] ETSI, "Multi-access Edge Computing (MEC); Phase 2: Use Cases and Requirements," Group Specification (GS) 002 v2.1.1, European Telecommunications Standards Institute (ETSI), Oct. 2018.

[62] W. Ayoub, A. E. Samhat, F. Nouvel, M. Mroue, and J. Prévotet, "Internet of Mobile Things: Overview of LoRaWAN, DASH7, and NB-IoT in LPWANs standards and Supported Mobility," *IEEE Communications Surveys Tutorials*, pp. 1–1, 2018.

[63] 3GPP, "TS 23.203 Policy and charging control architecture," *3rd Generation Partnership Project (3GPP), v16.0.0*, 2019.

[64] 5G-ACIA, "5G for Connected Industries and Automation," *5G Alliance for Connected Industries and Automation*, 2018.

[65] IEEE, "802.1Qbv-enhancements for Scheduled Traffic," *Institute of Electrical and Electronics Engineers, Inc*, 2016.

[66] A. Bader, H. ElSawy, M. Gharbieh, M. Alouini, A. Adinoyi, and F. Alshaalan, "First Mile Challenges for Large-Scale IoT," *IEEE Communications Magazine*, vol. 55, pp. 138–144, March 2017.

[67] H. Takagi and Y. Takahashi, "Priority Queues with Batch Poisson Arrivals," *Operations Research Letters*, vol. 10, no. 4, pp. 225 – 232, 1991.

[68] B. T. Doshi, "Queueing Systems with Vacations - A Survey," *Queueing Systems*, vol. 1, no. 1, pp. 29 – 66, 1986.

[69] F. Machihara, "A PReemptive Priority Queue as a Model with Server Vacations," *Journal of the Operations Research Society of Japan*, vol. 39, no. 1, pp. 118–131, 1996.

[70] M. Harchol-Balter, T. Osogami, A. Scheller-Wolf, and A. Wierman, "Multi-Server Queueing Systems with Multiple Priority Classes," *Queueing Systems*, vol. 51, pp. 331–360, Dec 2005.

[71] M. V. Vuuren and I. Adan, "Approximate Analysis of General Priority Queues," in *Analysis of manufacturing systems*, 2007.

[72] A. Sleptchenko, J. Selen, I. Adan, and G.-J. van Houtum, "Joint Queue Length Distribution of Multi-class, Single-server Queues with Preemptive Priorities," *Queueing Systems*, vol. 81, pp. 379–395, Dec 2015.

[73] H. ElSawy and E. Hossain, "On Stochastic Geometry Modeling of Cellular Uplink Transmission With Truncated Channel Inversion Power Control," *IEEE Transactions on Wireless Communications*, vol. 13, pp. 4454–4469, Aug 2014.

[74] V. G. Kulkarni, "Introduction to Matrix Analytic Methods in Stochastic Modeling," *Journal of Applied Mathematics and Stochastic Analysis*, vol. 12, 01 1999.

[75] R. M. Loynes, "The Stability of a Queue with Non-independent Inter-arrival and Service Times," *Mathematical Proceedings of the Cambridge Philosophical Society*, vol. 58, no. 3, p. 497–520, 1962.

[76] I. Krikidis, T. Charalambous, and J. S. Thompson, "Buffer-Aided Relay Selection for Cooperative Diversity Systems without Delay Constraints," *IEEE Transactions on Wireless Communications*, vol. 11, pp. 1957–1967, May 2012.

[77] H. Y. Lee, Y. J. Sang, and K. S. Kim, "On the Uplink SIR Distributions in Heterogeneous Cellular Networks," *IEEE Communications Letters*, vol. 18, pp. 2145–2148, Dec 2014.

[78] M. Haenggi, "The Meta Distribution of the SIR in Poisson Bipolar and Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 2577–2589, April 2016.

[79] H. ElSawy and M. Alouini, "On the Meta Distribution of Coverage Probability in Uplink Cellular Networks," *IEEE Communications Letters*, vol. 21, pp. 1625–1628, July 2017.

[80] Y. Wang, M. Haenggi, and Z. Tan, "The Meta Distribution of the SIR for Cellular Networks With Power Control," *IEEE Transactions on Communications*, vol. 66, pp. 1745–1757, April 2018.

[81] Y. Zhou and W. Zhuang, "Performance Analysis of Cooperative Communication in Decentralized Wireless Networks With Unsaturated Traffic," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 3518–3530, May 2016.

[82] S. K. Kaul and R. D. Yates, "Age of Information: Updates with Priority," in *2018 IEEE International Symposium on Information Theory (ISIT)*, pp. 2644–2648, June 2018.

[83] "A 5G Traffic Model for Industrial Use Cases," tech. rep., 5G Alliance for Connected Industries and Automation, 2019.

[84] F. Metzger, T. Hoßfeld, A. Bauer, S. Kounev, and P. E. Heegaard, "Modeling of Aggregated IoT Traffic and Its Application to an IoT Cloud," *Proceedings of the IEEE*, vol. 107, pp. 679–694, April 2019.

[85] H. ElSawy, "Characterizing IoT Networks with Asynchronous Time-Sensitive Periodic Traffic," *IEEE Wireless Communications Letters*, pp. 1–1, 2020.

[86] V. Gupta, S. K. Devar, N. H. Kumar, and K. P. Bagadi, "Modelling of IoT Traffic and Its Impact on LoRaWAN," in *GLOBECOM 2017 - 2017 IEEE Global Communications Conference*, Dec 2017.

[87] 3GPP, "TR 22.804 Study on Communication for Automation in Vertical domains," *3rd Generation Partnership Project (3GPP), v16.2.0*, 2018.

[88] C. Kam, S. Kompella, and A. Ephremides, "Age of Information under Random Updates," in *2013 IEEE International Symposium on Information Theory*, pp. 66–70, July 2013.

[89] S. K. Kaul, R. D. Yates, and M. Gruteser, "Status Updates Through Queues," in *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6, March 2012.

[90] N. Pappas, J. Gunnarsson, L. Kratz, M. Kountouris, and V. Angelakis, "Age of Information of Multiple Sources with Queue Management," in *2015 IEEE International Conference on Communications (ICC)*, pp. 5935–5940, June 2015.

[91] B. Zhou and W. Saad, "Joint Status Sampling and Updating for Minimizing Age of Information in the Internet of Things," *IEEE Transactions on Communications*, vol. 67, pp. 7468–7482, Nov 2019.

[92] M. Costa, M. Codreanu, and A. Ephremides, "On the Age of Information in Status Update Systems With Packet Management," *IEEE Transactions on Information Theory*, vol. 62, pp. 1897–1910, April 2016.

[93] Q. He, D. Yuan, and A. Ephremides, "On Optimal Link Scheduling with Min-max Peak Age of Information in Wireless Systems," in *2016 IEEE International Conference on Communications (ICC)*, May 2016.

[94] M. A. Abd-Elmagid and H. S. Dhillon, "Average Peak Age-of-Information Minimization in UAV-Assisted IoT Networks," *IEEE Transactions on Vehicular Technology*, vol. 68, Feb 2019.

[95] Y. Hu, Y. Zhong, and W. Zhang, "Age of Information in Poisson Networks," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*, pp. 1–6, Oct 2018.

[96] Y. H. Yang, A. Arafa, T. Q. S. Quek, and H. V. Poor, "Locally Adaptive Scheduling Policy for Optimizing Information Freshness in Wireless Networks," in *2019 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, Dec 2019.

[97] Y. Wang, M. Haenggi, and Z. Tan, "The Meta Distribution of the SIR for Cellular Networks With Power Control," *IEEE Transactions on Communications*, vol. 66, no. 4, pp. 1745–1757, 2018.

[98] X. Y., Q. C., Y. W., N. L., X. T., and M. V., "Stochastic Geometry Based Analysis for Heterogeneous Networks: A Perspective on Meta Distribution," *Sciece China. Information Sciences*, 12 2020.

[99] S. Singh, X. Zhang, and J. G. Andrews, "Joint Rate and SINR Coverage Analysis for Decoupled Uplink-Downlink Biased Cell Associations in HetNets," *IEEE Transactions on Wireless Communications*, vol. 14, pp. 5360–5373, Oct 2015.

[100] F. J. Martin-Vega, G. Gomez, M. C. Aguayo-Torres, and M. Di Renzo, "Analytical Modeling of Interference Aware Power Control for the Uplink of Heterogeneous Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 15, pp. 6742–6757, Oct 2016.

[101] M. Di Renzo and P. Guan, "Stochastic Geometry Modeling and System-Level Analysis of Uplink Heterogeneous Cellular Networks With Multi-Antenna Base Stations," *IEEE Transactions on Communications*, vol. 64, pp. 2453–2476, June 2016.

[102] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1. Wiley, January 1968.

[103] P. Reinecke and G. Horváth, "Phase-Type Distributions for Realistic Modelling in Discrete-Event Simulation," in *Proceedings of the 5th International ICST Conference on Simulation Tools and Techniques*, SIMUTOOLS '12, (Brussels, BEL), p. 283–290, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), 2012.

[104] H. H. Yang, Y. Wang, and T. Q. S. Quek, "Delay Analysis of Random Scheduling and Round Robin in Small Cell Networks," *IEEE Wireless Communications Letters*, vol. 7, pp. 978–981, Dec 2018.

[105] R. H. Khan and J. Y. Khan, "A comprehensive Review of the Application Characteristics and Traffic Requirements of a Smart Grid Communications Network," *Computer Networks*, vol. 57, no. 3, pp. 825 – 845, 2013.

[106] N. Nikaein, M. Laner, K. Zhou, P. Svoboda, D. Drajic, M. Popovic, and S. Krco, "Simple Traffic Modeling Framework for Machine Type Communication," in *ISWCS 2013; The Tenth International Symposium on Wireless Communication Systems*, pp. 1–5, Aug 2013.

[107] C. Park and J. Lee, "Successful Edge Computing Probability Analysis in Heterogeneous Networks," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2018.

[108] H. S. Jo, Y. J. Sang, P. Xia, and J. G. Andrews, "Heterogeneous Cellular Networks with Flexible Cell Association: A Comprehensive Downlink SINR Analysis," *IEEE Transactions on Wireless Communications*, vol. 11, pp. 3484–3495, Oct. 2012.

[109] K. Sato and T. Fujii, "Radio Environment Aware Computation Offloading with Multiple Mobile Edge Computing Servers," in *2017 IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, pp. 1–5, Mar. 2017.

[110] H. Q. Le, H. Al-Shatri, and A. Klein, "Efficient Resource Allocation in Mobile-edge Computation Offloading: Completion Time Minimization," in *2017 IEEE International Symposium on Information Theory (ISIT)*, pp. 2513–2517, Jun. 2017.

[111] A. Hekmati, P. Teymoori, T. D. Todd, D. Zhao, and G. Karakostas, "Optimal Mobile Computation Offloading With Hard Deadline Constraints," *IEEE Transactions on Mobile Computing*, pp. 1–1, 2019.

[112] Y. Mao, J. Zhang, and K. B. Letaief, "Joint Task Offloading Scheduling and Transmit Power Allocation for Mobile-Edge Computing Systems," in *2017 IEEE Wireless Communications and Networking Conference (WCNC)*, pp. 1–6, Mar. 2017.

[113] T. Li, C. S. Magurawalage, K. Wang, K. Xu, K. Yang, and H. Wang, "On Efficient Offloading Control in Cloud Radio Access Network with Mobile Edge Computing," in *2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS)*, pp. 2258–2263, Jun. 2017.

[114] T. D. Novlan, H. S. Dhillon, and J. G. Andrews, "Analytical Modeling of Uplink Cellular Networks," *IEEE Transactions on Wireless Communications*, vol. 12, pp. 2669–2679, Jun. 2013.

[115] X. Ge, S. Tu, G. Mao, C. X. Wang, and T. Han, "5G Ultra-Dense Cellular Networks," *IEEE Wireless Communications*, vol. 23, pp. 72–79, Feb. 2016.

[116] A. Rajanna and M. Haenggi, "Enhanced Cellular Coverage and Throughput Using Rateless Codes," *IEEE Transactions on Communications*, vol. 65, pp. 1899–1912, May 2017.

[117] Y. Lin, W. Bao, W. Yu, and B. Liang, "Optimizing User Association and Spectrum Allocation in HetNets: A Utility Perspective," *IEEE Journal on Selected Areas in Communications*, vol. 33, pp. 1025–1039, Jun. 2015.

[118] F. Boccardi, J. Andrews, H. Elshaer, M. Dohler, S. Parkvall, P. Popovski, and S. Singh, "Why to Decouple the Uplink and Downlink in Cellular Networks and How to Do it," *IEEE Communications Magazine*, vol. 54, pp. 110–117, Mar. 2016.

[119] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and Uplink Decoupling: A disruptive architectural design for 5G networks," in *2014 IEEE Global Communications Conference*, pp. 1798–1803, Dec. 2014.

[120] J. G. Andrews, S. Singh, Q. Ye, X. Lin, and H. S. Dhillon, "An Overview of Load Balancing in HetNets: Old Myths and Open Problems," *IEEE Wireless Communications*, vol. 21, no. 2, pp. 18–25, 2014.

[121] H. Holma and A. Toskala, *LTE Advanced: 3GPP Solution for IMT-Advanced*. Wiley Publishing, 1st ed., 2012.

[122] A. H. Sakr and E. Hossain, "Analysis of Multi-tier Uplink Cellular Networks with Energy Harvesting and Flexible Cell Association," in *2014 IEEE Global Communications Conference*, pp. 4525–4530, Dec. 2014.

[123] 5GAA, "C-V2X Use Cases: Methodology, Examples and Service Level Requirements," *White Paper*, 2019.

[124] 5GAA, "The Case for Cellular V2X for Safety and Cooperative Driving," *White Paper*, 2016.

[125] "Leading the world to 5G: Cellular Vehicle-to-Everything (C-V2X) technologies," tech. rep., Qualcomm [Online]. Available: https://www.qualcomm.com/media/documents/files/cellular-vehicle-to-everything-c-v2x-technologies.pdf, 2016.

[126] H. j. Gunther, O. Trauer, and L. Wolf, "The Potential of Collective Perception in Vehicular Ad-hoc Networks," in *2015 14th International Conference on ITS Telecommunications (ITST)*, pp. 1–5, Dec. 2015.

[127] K. Abboud, H. A. Omar, and W. Zhuang, "Interworking of DSRC and Cellular Network Technologies for V2X Communications: A Survey," *IEEE Transactions on Vehicular Technology*, vol. 65, pp. 9457–9470, Dec. 2016.

[128] I. Safiulin, S. Schwarz, T. Philosof, and M. Rupp, "Latency and Resource Utilization Analysis for V2X Communication over LTE MBSFN Transmission," in *WSA 2016; 20th International ITG Workshop on Smart Antennas*, pp. 1–6, Mar. 2016.

[129] A. F. Cattoni, D. Chandramouli, C. Sartori, R. Stademann, and P. Zanier, "Mobile Low Latency Services in 5G," in *2015 IEEE 81st Vehicular Technology Conference (VTC Spring)*, pp. 1–6, May 2015.

[130] K. Lee, J. Kim, Y. Park, H. Wang, and D. Hong, "Latency of Cellular-Based V2X: Perspectives on TTI-Proportional Latency and TTI-Independent Latency," *IEEE Access*, vol. 5, pp. 15800–15809, 2017.

[131] H. Cao, S. Gangakhedkar, A. R. Ali, M. Gharba, and J. Eichinger, "A Testbed for Experimenting 5G-V2X Requiring Ultra Reliability and Low-Latency," in *WSA 2017; 21th International ITG Workshop on Smart Antennas*, pp. 1–4, Mar. 2017.

[132] F. Giust, V. Sciancalepore, D. Sabella, M. C. Filippou, S. Mangiante, W. Featherstone, and D. Munaretto, "Multi-access Edge Computing: The driver behind the wheel of 5G-connected cars," *CoRR*, vol. abs/1803.07009, 2018.

[133] 5GAA, "Toward Fully Connected Vehicles: Edge Computing for Advanced Automotive Communications," *White Paper*, 2017.

[134] R. Kawasaki, H. Onishi, and T. Murase, "Performance Evaluation on V2X Communication with PC5-based and Uu-based LTE in Crash Warning Application," in *2017 IEEE 6th Global Conference on Consumer Electronics (GCCE)*, pp. 1–2, Oct. 2017.

[135] P. Luoto et al., "Vehicle Clustering for Improving Enhanced LTE-V2X Network Performance," in *2017 European Conference on Networks and Communications (EuCNC)*, pp. 1–5, Jun. 2017.

[136] "WINNER II Channel Model, D1.1.2 V1.0," *White Paper*, 2007.

[137] D. Sabella, N. Nikaein, A. Huang, J. Xhembulla, G. Malnati, and S. Scarpina, "A Hierarchical MEC Architecture: Experimenting the RAVEN Use-Case," in *2018 IEEE 87th Vehicular Technology Conference (VTC Spring)*, pp. 1–5, 2018.

[138] Z. Jiang, B. Krishnamachari, X. Zheng, S. Zhou, and Z. Niu, "Timely Status Update in Wireless Uplinks: Analytical Solutions With Asymptotic Optimality," *IEEE Internet of Things Journal*, vol. 6, pp. 3885–3898, April 2019.

[139] M. Bennis, M. Debbah, and H. V. Poor, "Ultra-reliable and Low-Latency Wireless Communication: Tail, Risk, and Scale," *Proceedings of the IEEE*, vol. 106, pp. 1834–1853, Oct 2018.

[140] C. Colman-Meixner, C. Develder, M. Tornatore, and B. Mukherjee, "A Survey on Resiliency Techniques in Cloud Computing Infrastructures and Applications," *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 2244–2281, 2016.

[141] R. Birke, I. Giurgiu, L. Y. Chen, D. Wiesmann, and T. Engbersen, "Failure Analysis of Virtual and Physical Machines: Patterns, Causes and Characteristics," in *2014 44th Annual IEEE International Conference on Dependable Systems and Networks*, 2014.

[142] S. Ko, K. Han, and K. Huang, "Wireless Networks for Mobile Edge Computing: Spatial Modeling and Latency Analysis," *IEEE Transactions on Wireless Communications*, vol. 17, pp. 5225–5240, Aug 2018.

[143] H. Lee and J. Lee, "Task Offloading in Heterogeneous Mobile Cloud Computing: Modeling, Analysis, and Cloudlet Deployment," *IEEE Access*, vol. 6, pp. 14908–14925, 2018.

[144] H. Ko, J. Lee, and S. Pack, "Spatial and Temporal Computation Offloading Decision Algorithm in Edge Cloud-Enabled Heterogeneous Networks," *IEEE Access*, vol. 6, pp. 18920–18932, 2018.

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

[146] R. Arshad, L. H. Afify, H. ElSawy, T. Y. Al-Naffouri, and M. Alouini, "On the Effect of Uplink Power Control on Temporal Retransmission Diversity," *IEEE Wireless Communications Letters*, vol. 8, no. 1, pp. 309–312, 2019.

[147] D. Bruneo, "A Stochastic Model to Investigate Data Center Performance and QoS in IaaS Cloud Computing Systems," *IEEE Transactions on Parallel and Distributed Systems*, vol. 25, no. 3, pp. 560–569, 2014.

[148] S. Fu, "Failure-Aware Construction and Reconfiguration of Distributed Virtual Machines for High Availability Computing," in *9th IEEE/ACM International Symposium on Cluster Computing*, pp. 372–379, 2009.

[149] I. A. M. Balapuwaduge and F. Y. Li, "A Joint Time-Space Domain Analysis for Ultra-Reliable Communication in 5G Networks," in *2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, 2018.

[150] 3GPP, "TS 22.261 Service requirements for next generation new services and markets," *3rd Generation Partnership Project (3GPP), v16.8.0*, 2019.

[151] H. V. K. Mendis and F. Y. Li, "Achieving Ultra Reliable Communication in 5G Networks: A Dependability Perspective Availability Analysis in the Space Domain," *IEEE Communications Letters*, vol. 21, pp. 2057–2060, Sep. 2017.

[152] I. A. M. Balapuwaduge, F. Y. Li, and V. Pla, "Dynamic Spectrum Reservation for CR Networks in the Presence of Channel Failures: Channel Allocation and Reliability Analysis," *IEEE Transactions on Wireless Communications*, vol. 17, no. 2, pp. 882–898, 2018.

[153] T. Hößler, L. Scheuvens, N. Franchi, M. Simsek, and G. P. Fettweis, "Applying Reliability Theory for Future Wireless Communication Networks," in *2017 IEEE 28th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, pp. 1–7, Oct. 2017.

[154] R. J. Boucherie and E. A. van Doom, "Uniformization for λ-positive Markov Chains," *Communications in Statistics. Stochastic Models*, vol. 14, no. 1-2, pp. 171–186, 1998.

Appendix A

A.1 Proof of Proposition 1

In order to fully characterize the vacation period for a given priority class, the aggregate busy periods of higher priority queues need to be characterized. For the highest priority class (i.e., $i = 1$), its transition matrix $\mathbf{P}_1$ is that of a simple birth-death process. Consequently, its busy period, denoted as $\mathbf{V}_1$ is represented via the following absorbing Markov chain

$$\mathbf{V}_1 = \begin{bmatrix} \bar{\alpha}_1 \bar{P_{\mathrm{s},1}} + \alpha_1 P_{\mathrm{s},1} & \alpha_1 \bar{P_{\mathrm{s},1}} & & \\ \bar{\alpha}_1 P_{\mathrm{s},1} & \bar{\alpha}_1 \bar{P_{\mathrm{s},1}} + \alpha_1 p & \alpha_1 \bar{P_{\mathrm{s},1}} & \\ & & \ddots & \ddots \\ & & \bar{\alpha}_1 P_{\mathrm{s},1} & \bar{\alpha}_1 \bar{P_{\mathrm{s},1}} + \alpha_1 \end{bmatrix}. \tag{A.1}$$

Let $\tilde{\mathbf{v}}_1 = [\bar{\alpha}_1 P_{\mathrm{s},1}\ 0_{q_1}] \in \mathbb{R}^{q_1 - 1 \times 1}$ denotes the absorption vector. Through $\mathbf{V}_1$ and $\tilde{\mathbf{v}}_1$, one can fully characterize the transitions when a first priority packet arrives as well as its successful departure (i.e., absorption). The second priority queue can be modeled as Geo/PH/1 queue, where the PH type distribution models the busy period of the first priority queue. Consider then the case of serving second priority packets, if a first priority packet arrives, an initialization vector $\mathbf{v}_1 = [1\ \mathbf{0}_{q_1}]$ is required to characterize the states distribution and the probability of their occurrence χ_1. Since the queue is initialized as empty, $\chi_1 = \alpha_1$. The analysis for a generic i-th priority class is extended and with some mathematical adaptations, the lemma is finalized.

A.2 Proof of Theorem 1

For the dedicated access scheme, a packet belonging to the i-th priority queue will only experience aggregate interference from packets belonging to the same priority class.This packet will be granted transmission only if all the higher priority queues are empty. The portion of interfering device for the i-th queue at the BS is $\mu \mathcal{J}_i$, where $\mathcal{J}_i = \sum_{z_i=1}^{q_i} \mathbb{P}\{(0,0,\cdots,0,z_i)\}$ is the joint probability of having idle i-1 priority queues and non-idle i-th priority queue. Additionally, the adopted grant-free transmission scheme among the devices imposes a differentiation between the experienced interference into intra-cell

and inter-cell interference, thus (3.15) is written as

$$P_{\mathrm{s},i} = \exp\Big\{ - \frac{\sigma^2 \theta}{\rho} \Big\} \mathcal{L}_{I_{\mathrm{out},i}}\Big(\frac{\theta}{\rho}\Big) \mathcal{L}_{I_{\mathrm{in}}}\Big(\frac{\theta}{\rho}\Big). \tag{A.2}$$

Since full channel inversion power control with threshold ρ is employed, two main results hold: (i) received power from the devices at a given BS equals ρ (ii) interference power from the neighboring devices is strictly lower than ρ. Following [73, Theorem 1], the LT of the aggregate intra-cell interference at the serving BS for an i-th priority packet is

$$\mathcal{L}_{I_{\mathrm{out},i}}(s) \approx \exp\Big(- 2\pi\mu \mathcal{J}_i s^{\frac{2}{\eta}} \mathbb{E}_P \Big[P^{\frac{2}{\eta}} \Big] \int_{(s\rho)^{\frac{-1}{\eta}}}^{\infty} \frac{y}{y^{\eta}+1} dy \Big), \tag{A.3}$$

where the approximation is due to the assumed independent transmission powers of the devices (Approximation 1(i)). The LT of the inter-cell interference can be evaluated as [35, Lemma 1]

$$\mathcal{L}_{I_{\mathrm{in}}}(s) \approx \mathbb{P}\{\mathcal{N}=0\} + \sum_{n=1}^{\infty} \frac{\mu^n (\lambda c)^c \Gamma(n+c)}{(1+s\rho)^n \mu + \lambda c)^{n+c} \Gamma(n+1)\Gamma(c)}, \tag{A.4}$$

where $\Gamma(\cdot)$ is the gamma function, $\mathcal{N}$ is a random variable representing the number of neighbors and $c = 3.575$ is a constant defined to approximate Voronoi cell's PDF in $\mathbb{R}^2$. Plugging (A.3) and (A.4) into (A.2) and following [35, Lemma 1], the theorem is derived.

A.3 Proof of Lemma 3

The b-th moment of the TSP can be derived from eq.(4.13) as

$$M_b = \mathbb{E}^{!}_{r_i,P_i,r_o} \left\{ \prod_{\omega_i \in \tilde{\Phi}_T} \left(\frac{1}{1 + \frac{\theta r_o^{\eta(1-\varepsilon)}}{\rho \omega_i}} \right)^b \middle| \hat{\Phi}, \Psi \right\}, \tag{A.5}$$

where, the uplink transmission power P_i of the i-th device is a random variable due to the employed fractional path-loss power control [23]. In (A.5), the average is first conditioned on r_o then evaluated via the probability generating functional of the PPP with the intensity function $\tilde{\lambda}_T(\omega)$. The distribution of r_o is given by $f_{r_o}(r) = 2\pi\lambda r e^{-\pi\lambda r^2}$. With some mathematical operations following [79], the lemma is proved.

A.4 Proof of Lemma 6

Based on [74],[29], $\mathbf{R}_n$ is the minimal non-negative solution to the quadratic equation $\mathbf{R}_n = \mathbf{A}_{0,n} + \mathbf{R}_n \mathbf{A}_{1,n} + \mathbf{R}_n^2 \mathbf{A}_{2,n}$. Let $\mathbf{x}_{0,n}$ and $\mathbf{x}_{1,n}$ be the solution to

$$\begin{bmatrix} \mathbf{x}_{0,n} & \mathbf{x}_{1,n} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_{i,n} & \mathbf{x}_{i,n} \end{bmatrix} \begin{bmatrix} \mathbf{B} & \mathbf{C} \\ \mathbf{A}_{2,n} & \mathbf{A}_{1,n} + \mathbf{R}_n \mathbf{A}_{2,n} \end{bmatrix}. \tag{A.6}$$

Since $\mathbf{A}_{0,n}$ is rank 1, $\mathbf{R}_n$ can be rewritten as

$$\mathbf{R}_n = \mathbf{A}_{0,n}(\mathbf{I}_T - \mathbf{A}_{1,n} - \mathbf{A}_{2,n}\mathbf{G}_n)^{-1}, \tag{A.7}$$

where $\mathbf{G}_n$ is the minimal non-negative solution to $\mathbf{G}_n = \mathbf{A}_{2,n} + \mathbf{A}_{1,n}\mathbf{G}_n + \mathbf{A}_{0,n}\mathbf{G}_n^2$. Following [29][Chapter 5.9], the lemma can be proved.