Project number: 826276

CPS4EU

Cyber Physical Systems for Europe

D2.3 - Propositions for 5G, including URLLC evolution - v1

Reviewer (name – company): Antoine Dupret (CEA)

Dissemination level: Public

Version	Date	Author (name – company)	Comments
V1.0		D. Demmer (CEA) N. di Pietro (CEA) J.B. Doré (CEA) V. Savin (CEA) A. Falempin (CEA) S. Bicaïs (CEA) M.-T. Thi (CEA) S. Ben Hadj Said (CEA) M. Boc (CEA) J. Schmitt (VSORA) G. Vivier (SEQ)	

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

EXECUTIVE SUMMARY

In recent years, Cyber Physical System (CPS) technologies have become a game changer in strategic sectors where Europe is a world leader, such as Automotive, Energy and Industry Automation. CPS4EU is a European project that aims at developing enabling technologies for CPSs, encompassing various vertical use cases. CPS4EU's main overall goal is to strengthen the CPS value chain by creating world-class European Small and Medium Enterprises (SME) and by providing CPS technologies that in turn will sustain the leadership of the large European groups in key economy sectors. In this way, the project is stimulating the development of innovative products to support the massive digitalization increasingly integrated in our everyday environment.

Work Package (WP) 2 of CPS4EU deals with the communication aspects of CPSs. It investigates how a CPS can be connected efficiently and securely to a network, in the context of emerging technologies, as Internet of Things (IoT) and 5G. It also targets the design and development of solutions that guarantee specific performance and quality of service for various CPS applications. Part of the communication modules developed in WP2 will be integrated into the communication platform of WP6.

The numerous emerging CPS application scenarios and use cases impose constraints and require features that cannot be met by nowadays technologies. Therefore, Task 2.2 of WP2 has the role of investigating novel communication solutions that surpass the state of the art and are especially adapted to the requirements coming from other WPs, analysed during the work of Task 2.1 and reported in deliverables D2.1 [D2.1] and D2.2 [D2.2].

This report is the first deliverable produced by Task 2.2 and describes the activities of this task during the first half of the project's lifetime. These activities were mainly devoted to the study of novel telecommunication solutions at different network layers and for different application scenarios, though all with the common goal of serving CPSs. A particular attention has been paid so far to enablers of Ultra-Reliable Low-Latency Communications (URLCC) and Time-Sensitive Networking (TSN), identified by the project as key features of an effective CPS-supporting communication infrastructure.

Besides briefly recalling in Section 2 the main outcomes of Task 2.1, we present in Section 3 of this document the results of five prospective studies. Although being in line with the actual and foreseen standardization directions for modern communication networks, our work also makes a further step "beyond 5G" to address the new challenges posed by CPSs. These studies span through several domains: physical-layer enablers for URLLC networks at the service of CPSs, optimization of collaborative communications in industrial environments, algorithms for energy efficiency in scenarios where CPSs leverage edge computing under latency and reliability constraints, TSN-dedicated network architectures, and solutions for the improvement of the quality of wireless communications in mobility scenarios. The techniques and algorithms presented in Section 3.2 and 3.5 have been selected for further development and evaluation in Task 2.3, as it will be reported in full detail in deliverables D2.5 [D2.5] and D2.6 [D2.6].

A second version of this report will be produced at the end of the project's lifetime and will include possible updates on the activities carried out so far and a detailed description of Task 2.2's work in the next 17 months.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

Table of content

Tables

Figures

1. INTRODUCTION

1.1. Purpose

This deliverable is the first report of Task 2.2 of WP2, which aims at investigating novel solutions for efficient 5G communication modules for Cyber Physical Systems (CPSs). In particular, since CPSs are envisaged to be deployed in new telecommunication scenarios with different and more stringent performance requirements compared to what nowadays mobile technologies can guarantee, this deliverable reports on CPS4EU's innovative studies in the domain of Ultra-Reliable and Low-Latency Communications (URLLC) and Time-Sensitive Networking (TSN), in line with the 5G development in 3GPP.

The research work carried out by CPS4EU's WP2 is made of both a theoretical analysis and a numerical validation of the proposed solutions. The studies presented in this document are grounded on the requirements, use cases, and motivations defined in Task 2.1 and cover several aspects of URLLC and CPS applications, such as network architectural enablers, innovations at the physical layer, resource orchestration solutions for URLLC applied to edge computing scenarios, mobility aspects. Part of the ideas and algorithms presented in this document will be further experimented with within the activities of Task 2.3.

1.2. Definition, acronyms, and abbreviations

Acronym / abbreviation	Description
3GPP	Third Generation Partnership Project
3G	3rd Generation of technology standard for broadband cellular networks
4G	4th Generation of technology standard for broadband cellular networks
5G	5th Generation of technology standard for broadband cellular networks
5GS	5G System
AF	Application Function
ALU	Arithmetic Logic Unit
AP	Access Point
API	Application Programming Interface
AS	IEEE's standard n. 802.1AS
ASIC	Application-Specific Integrated Circuit
BCH code	Bose, Ray-Chaudhuri and Hocquenghem code
BER	Bit Error Rate
CF	Correction Field
CFO	Carrier Frequency Offset
CNC	Centralized Network Configuration
CoMP	Coordinated Multi-Point
CPS	Cyber-Physical System
CPS4EU	Cyber-Physical Systems for Europe
CPU	Central Processing Unit
DMA	Direct Memory Access
D-MEC	Discontinuous Mobile Edge Computing
D-RAM	Dynamic Random-Access Memory
DSP	Digital Signal Processing
DS-TT	Device-Side TSN Translator
DTX	Discontinuous Transmission
ED	Energy Detection
ES	Edge Server
ETSI	European Telecommunications Standards Institute
FDD	Frequency-Division Duplex
FEC	Forward Error Correcting
FPGA	Field-Programmable Gate Array
GM	Grand Master
GPU	Graphics Processing Unit

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

HW	Hardware
ICT	Information and Communications Technology
IEEE	Institute of Electrical and Electronics Engineers
IoT	Internet of Things
IT	Information Technology
KPI	Key Parameter Indicator
LAN	Local Area Networking
LDPC	Low-Density Parity-Check
LOS	Line Of Sight
LTE	Long Term Evolution
MAC	Medium Access Control
MCS	Modulation and Coding Scheme
MEC	Multi-access Edge Computing
MHCP	Matérn Hard-Core Process
MIMO	Multiple-input Multiple-output
ML	Maximum Likelihood
MLD-GA	Maximum Likelihood Detector with Gaussian Approximation
mmWave	Millimetre Wave
NB-IoT	Narrow-Band Internet of Things
NLOS	Non Line Of Sight
NND	Neural-Network-based Detector
NW-TT	Network-Side TSN Translator
OOK	On-Off Keying
PER	Packet Error Rate
PHY	Physical layer
PLC	Programmable Logic Controller
PPP	Poisson Point Process
PTP	Precision Time Protocol
QAM	Quadrature Amplitude Modulation
QoS	Quality of Service
ReLU	Rectified Linear Unit
RF	Radio Frequency
RTL	Register-Transfer Level
Rx	Reception
SDN	Software-Defined Networking
SME	Small and Medium Enterprises
SNR	Signal-to-Noise Ratio
SW	Software
TCM	Tightly-Coupled Memory
TD	Threshold Detection
TDD	Time-Division Duplexing
TLM	Transaction-Level Modelling
TRP	Transmit-Receive Point
TSN	Time Sensitive Networking
Tx	Transmission
UE	User Equipment
ULA	Uniform Linear Array
u.r.g.	Uniformly Randomly Generated
URLLC	Ultra-Reliable and Low-Latency Communications
V2X	Vehicle-to-everything
WP	Work Package

2. REQUIREMENTS AND MOTIVATION

Task 2.1 of CPS4EU's WP2 had the objective to collect requirements from the vertical WPs with respect to the communication needs. The outcome of the task was captured initially in deliverable D2.1 [D2.1] and completed in deliverable D2.2 [D2.2].

CPS4EU covers a large variety of vertical use cases, from energy, factory automation, automotive and we realized that every specific use case has finally its own set of requirements for communication. First, most of use cases mentioned communication as a "must have". Very few consider it as just "nice to have": for instance, automotive driving relies on local sensors and processing for emergency situations, but leverages communication for a better service (e.g. for map upgrade, traffic information retrieval, or simply secure software upgrade). This variety of use cases yield a variety of requirements with respect to communication: either very low and infrequent data exchanges (e.g. for sensors reporting few measurements per day like metering use cases) or more frequent (e.g. for asset tracking), and in many cases including very stringent requirements either in terms of data rate or in terms of latency or reliability (e.g. for remote control of moving devices, virtual or augmented reality, digital twin, coordinated lifting for cranes, robots coordination in factory 4.0, etc.).

Task 2.1 conducted a quick survey on existing technologies and we concluded that there is no single wireless communication technique that could meet all the requirements. However, we also realized that the cellular framework offers a set of solutions that could cover most of the requirements. Indeed, cat-M and NB-IoT is an efficient solution for battery powered low-end devices while LTE could meet higher-end use cases, thanks to its various categories (from category 1 allowing throughput of few tens of Mbps to category 18, up to around 1 Gbps). So, 4G is already a satisfactory candidate as a universal communication system for CPS.

However, some specific use case cannot be met with 4G mostly because of three characteristics: the need of low (and sometimes deterministic) latency; the need of higher throughput, and the foreseen explosion of the number of connected low-end devices (massive machine-type communications). 5G then becomes a natural candidate to fulfil such CPS requirements, especially for the significant flexibility it will provide, at least from the network perspective and deployment possibilities. Hence, with a single network solution, an operator could offer various "slices" with different QoS profiles. In addition, 5G extends the possibility to offer private network deployments: a network could be deployed specifically for a vertical, locally, e.g. over its factory, independently of consumer networks, naturally fulfilling requirements in terms of security and data protection (e.g. the network servers being hosted locally at the factory).

As a result, 5G opens the door for a new set of IoT, often referred to as "critical" or "industrial" IoT, encompassing the complete set of communication requirements of CPSs, as gathered from WP6, WP7, WP8 and WP9 of CPS4EU.

However, 5G was initially defined for the consumer market and the requirements of verticals were captured but not prioritized. The first 5G version, release 15, was mostly serving higher data rates, with little work on critical communications. Release 16 compensated a bit this gap but still important aspects are not fully defined, such as the notion of URLLC profile (what is exactly an URLLC communication?), how to integrate TSN within a 5G network, or how we could use new frequency bands, especially the millimetre-wave (mmWave) one, to support harsh communications scenarios.

In the perspective of 5G development and evolution, WP2 has also the mission to propose innovative solutions, in line with standard work, and to provide missing building blocks (it has to be noted that a standard ensures interoperability between device and network equipment, but does not define internal algorithms to support this interoperability).

As a result, in WP2, Task 2.2 has the objective to work mid- to long-term solutions to enable the few CPS use cases that cannot be served with nowadays and emerging wireless communications systems. The next section presents the results obtained by Task 2.2 in the first half of CPS4EU's lifetime.

3. INNOVATIVE SOLUTIONS FOR CPS COMMUNICATION MODULES AND NETWORKING

3.1. Context and relation between the project goals and the proposed studies

As introduced above, CPSs represent key drivers for the innovation capacity of European industries in sectors like automotive, energy and industry automation. CPSs combine intensive connectivity, embedded computing and local intelligence to create a link between physical and digital worlds and to enable cooperation among systems. The importance of CPSs is increasing with massive digitalization and their development is opening new market opportunities and requires non-straightforward technological innovations. Therefore, one of CPS4EU's main objective is to foster the creation of world-class CPS technology providers. In line with this goal, WP2 focuses on the design and development of CPS HW/SW communication modules to be used across different industrial sectors, leveraging requirements coming directly from the significant number of CPS4EU industrial partners, thus reinforcing their leadership in a very competitive market. The goal is to guarantee dynamic management of applicative flows (ex: reliability, latency, and data rate) and continuity of service in order to ensure the control, localization of vehicles and CPS devices, distribution and collection of contents and critical or non-critical data.

Connectivity is critical for CPSs. Indeed, it must bear communications of hundreds, or even thousands, of dynamically connected objects that are also connected to external operators while ensuring reliability. Such connectivity, whether object-to-object or object-to-network infrastructure, often needs to be wireless (or to integrate complementary wireless and wired features) to offer more flexibility, adaptability to changing environment and much lower cost of operation compared to a wired communication approach. It must also be able to auto-adapt dynamically in time and space and to support the differentiation of operations or services. Thus, CPS requirements in terms of connectivity introduce the need for a very limited latency of communication between the device and the infrastructure. This is particularly the case of remote control of heavy machinery in hazardous places or used for monitoring and controlling smart grids. The challenge is to provide extremely fast and reliable connectivity at the radio level, as well as for end-to-end performance of the system. This requires also a more distributed infrastructure design that include mobile edge computing (MEC) capabilities.

In general, to achieve the performance required by URLLC applications for CPSs, it is necessary to combine in the architecture of 5G networks a series of enablers and technological innovations. In particular, it will be necessary to implement at all layers solutions that make the network more robust and able to guarantee both low latencies and low transmission error probabilities. These solutions include multi-connectivity enablers (network reliability through spatial and frequency redundancy or diversity), high coordination and synchronization of network nodes (for instance via TSN) and new coding, modulation, and scheduling schemes for the PHY, MAC, and transport layers. One of the goal of CPS4EU's WP2 is to investigate these URLLC enablers and propose novel technological and theoretical solutions for 5G networks and beyond, adapted to CPS-specific use cases and applications. These solutions will be defined on top or be complementary to the solutions defined by the standardisation bodies addressing these challenges, such as 3GPP or IEEE.

A common feature of such solutions from a topological and infrastructural point of view, will be the densification of the network and the presence of more access points and more cooperation between them, typically in "mesh" topologies. To guarantee URLLC services, it will therefore be necessary to make communication between access points and backhauling extremely efficient, even more so than they already are. Therefore, we present in Section 3.2 our innovative results on mmWave MIMO xHauling. The demand for wireless mmWave xHaul connection is increasing especially for the rapid deployment of private networks. mmWave xHauls improve the data rate, latency, and power consumption of modern networks and are thus particularly important for URLLC networks. They provide greater deployment "agility" in deployment, very high capacities thanks to high-frequency communications and reliability through the possibility of multi-connectivity of the "mesh" topology. The solution that we propose has also the advantage of a relatively low implementation cost, because it is based on a radio architecture that is simpler and cheaper than others. In spite of these advantages, the performance of mmWave radio-frequency systems is severely degraded by a typically strong oscillator phase noise. Therefore, in Section 3.2 we investigate the use of MIMO systems with energy detection receivers to achieve high-rate communications robust to phase noise. First, the design of the receiver detection algorithm is addressed. In particular, we propose to use an original and efficient detector based on neural networks. The communication performance is assessed through numerical simulations for uncoded and coded systems. Our results, further detailed in [BFD+20] , demonstrate that spatial multiplexing with non-coherent mmWave transceivers can be realized on strongly correlated line-of-sight channels using the proposed detection schemes. Thereby, we highlight that high-rate mmWave RF systems can be implemented with low-complexity and low-power architectures using MIMO systems with energy detection receivers.

In Section 3.3, we focus on cooperative reliable communications for CPSs in industrial environments. Namely, we propose a solution for an industry automation scenario where a central controller broadcasts critical messages to the wireless devices (e.g., sensors/actuators). We devise a stochastic geometry framework where the rate coverage probability of devices is modelled by taking into account the density of roaming blockages over the factory floor. To alleviate the loss in the coverage, we adopt a two-phase transmission policy, where in the broadcast phase, the central controller broadcasts the messages intended for the devices in the network area. The devices in coverage during the broadcast phase act as decode-and-forward relays in the relay phase, so as to reinforce the signal strength at the devices in outage. The total downlink transmission time is therefore partitioned into two phases by a tuneable factor. Finally, we study the optimal value of the partitioning factor with varying device densities, blockage densities, and file sizes and we highlight that a longer transmission time should be allotted to the broadcast phase in the case of larger file sizes or lower transmit power of the controller.

Section 3.4 is dedicated to scenarios where CPSs are assisted by an edge computing server that can rapidly and efficiently run computational tasks on their behalf. In particular, we describe a novel strategy for energy-efficient dynamic computation offloading aimed at minimizing the energy consumption of the overall system, comprising user devices and network elements, under URLLC constraints on the end-to-end service delay and the packet error rate over the wireless interface. To reduce the energy consumption, we exploit low-power sleep operation modes for the user devices, the access point and the edge server, shifting the edge computing paradigm from an always on to an always available architecture, capable of guaranteeing an on-demand target service quality with the minimum energy consumption. To this aim, we introduce Discontinuous Mobile Edge Computing (D-MEC): an online algorithm that dynamically and optimally orchestrates the sleep mode operations and the duty cycles of the network's elements. In such a dynamic framework, end-to-end delay constraints translate into constraints on overall queueing delays, including both the communication and the computation phases of the offloading service. D-MEC hinges on stochastic Lyapunov optimization, does not require any prior knowledge on the statistics of the offloading traffic or the radio channels, and satisfies the long-term performance constraints imposed by users. Numerical results illustrate the advantages of the proposed method.

Moreover, in Section 3.5, we discuss TSN, which is a set of standards that provide deterministic communication to standard Ethernet and are being nowadays extended to wireless. An important objective of TSN is to guarantee data delivery with deterministic delay and jitter for real-time applications, which can be a crucial feature for URLLC service and CPSs. Section 3.5 recalls the main features of TSN and briefly describes the testbed of Ethernet TSN implemented by CEA, whose evaluated performance on time synchronization is reported in [TBB20].

Finally, in Section 3.6, we present our initial ideas for a novel channel estimation technique in scenarios with non-static users, that we will further develop in the second half of CPS4EU's lifetime. This solution will leverage cooperation between network access points and will be exploitable in industrial environments.

3.2. *Design and analysis of MIMO systems using energy detectors for mmWave applications*

3.2.1. Motivation and related work

Multi-connectivity schemes (based on frequency, time, or spatial diversity) are a way to increase the reliability of a wireless transmission. This feature is critical for CPS application scenarios. From the infrastructure side, reliability can be supported via the deployment of a mesh network system. This requires though extreme xHaul capabilities (bandwidth, latency requirements). These requirements make fibre solutions desirable, but sometimes complicated by local installation constraints to efficiently deploy networks. The wireless infrastructure is foreseen as a complement to the optical fibre deployment as it offers more agility, shorter installation times, and, in the case of a mesh architecture, strong reliability. It may also provide connectivity to mobile, removable or even flying access points. Finally yet importantly, these solutions will be valuable if and only if their cost efficiency is demonstrated. From a spectrum perspective, this paradigm will require large bandwidth and therefore wireless mmWave links in V-, E-, W-, D-bands, namely 60 GHz to 81 GHz, 90 GHz and 140 GHz, are investigated by 3GPP's standardization groups as a complement to 5G mmWave bands (26-28 and 39 GHz). These facts support the search for novel architectural and hardware solutions for cost-effective and high-performance wireless infrastructures.

To achieve high-data-rate mmWave communications, additional research is required to design efficient and new physical layer algorithms. Traditional techniques cannot be directly transposed to mmWave bands (especially for spectrum above 60GHz) as they do not consider the specific features of RF impairments of mmWave systems. In

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

particular, they suffer from strong phase impairments due to the poor performance of high-frequency oscillators [VPZ13]. State-of-the-art approaches [JDC20] investigate the use of coherent systems together with channel bonding. This type of architecture needs to be further combined with signal processing optimizations to mitigate the effects of phase impairments leading to complex practical implementations. Conversely, we consider non-coherent detection for its inherent robustness to phase noise and simple implementation. As an example, fully integrated 260GHz on-off keying (OOK) transceiver is demonstrated in [PKT+12]. Transceivers based on energy detection (ED) have been extensively studied for systems with a single transmit antenna and multiple receive antennas, see [JDPM16] and references therein. Nevertheless, for non-coherent sub-THz systems, the main challenge is to increase the spectral efficiency while maintaining a low complexity. With regard to this objective, the work in [PW09] is relevant as it shows that MIMO systems with amplitude detection receivers may exploit spatial multiplexing to increase their spectral efficiency. Therefore, we investigate the design of MIMO systems with ED receivers to achieve high rate mmWave communications.

In our scenario, the channels are strongly correlated, and moreover, the resulting interference is nonlinear due to the ED at the receiver. The strong and nonlinear interference between channels represents a significant challenge to achieve spatial multiplexing with non-coherent transceivers in sub-THz frequencies. The contributions of this work are the following. First, we derive the model for MIMO systems using ED receivers. Second, the design of the receiver detection algorithm is investigated. We derive the joint maximum likelihood (ML) detector corresponding to the studied nonlinear MIMO channel using a Gaussian approximation approach. In addition, we propose an original and efficient detector based on neural networks, which does not need any knowledge or assumption of the channel. We also detail the differences between state-of-the-art detection methods and the two proposed detectors. Third, the system performance is evaluated through numerical simulations. We introduce a realistic scenario modelling a fixed indoor wireless link in the D band at 145 GHz. Our results show that low-error rate communications can be achieved with strong spatial interference between channels using the proposed demodulation algorithms. Fourth, we consider the integration of a forward error correcting code (FEC) scheme in order to achieve channel coding gains. With regard to targeted high-rate low-complexity applications, we propose the use of a Bose, Ray-Chaudhuri and Hocquenghem (BCH) code with a short packet length that can be implemented with a low-latency low-complexity decoder. The results of numerical simulations confirm that the integration of the FEC scheme leads to significant performance gains in terms of achievable data rate. Finally, we highlight that this work lead to the submission of patent proposal n. FR2005670, 29/05/2020 [BDF20].

3.2.2. System model

We consider a MIMO communication system with N_t transmit antennas and N_r receive antennas with $N_t \leq N_r$, as illustrated in Figure 3.2-1. The propagation channel is described by two $N_r \times N_t$ matrices: $H = \left(h_{k,n}\right)_{k,n}$ and $\Phi = \left(\phi_{k,n}\right)_{k,n}$, where $h_{k,n}$ and $\phi_{k,n}$ denote respectively the propagation gain and phase shift of the channel for signals transmitted on the n-th Tx RF chain and received on the k-th Rx RF chain.

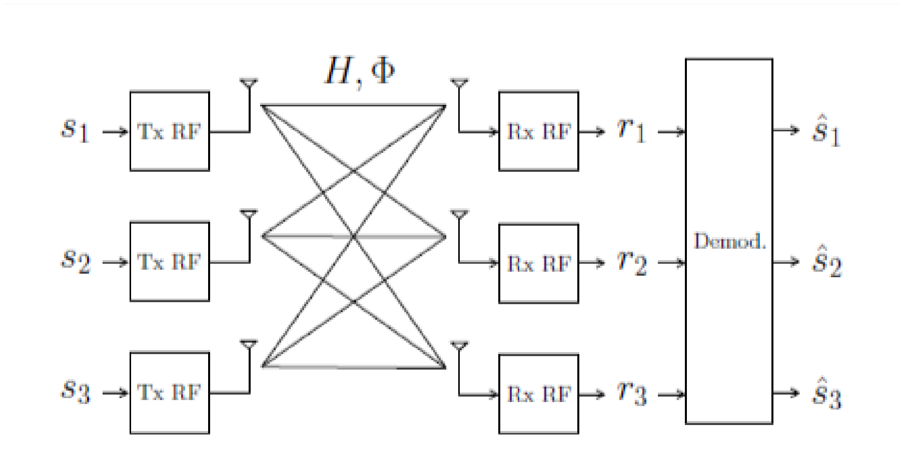

Figure 3.2-1 - Block diagram of a 3 x 3 MIMO transceiver.

The transmitter implements envelope modulation and the architecture of one of its RF chains is depicted in Figure 3.2-2. Using envelope modulation allows a simple implementation and an efficient use of power amplifiers. In this case, OOK appears to be a simple and efficient modulation scheme considering a non-coherent demodulation.

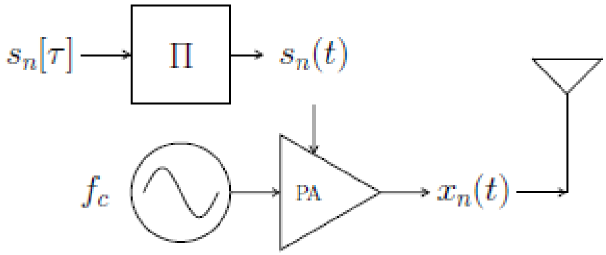

Figure 3.2-2 - Block diagram of one Tx RF chain.

The recent measurement campaigns have shown that mmWave propagation channels are largely dominated by a single path, often the line-of-sight (LOS) direct path, which provides most of the energy contribution. It is due to the stronger channel sparsity at those frequencies, in particular in open or urban environment, and to the usage of highly directive antennas, sometimes at both transceiver sides.

The receiver RF chains whose architecture is depicted in Figure 3.2-3 is simply a non-coherent detector.

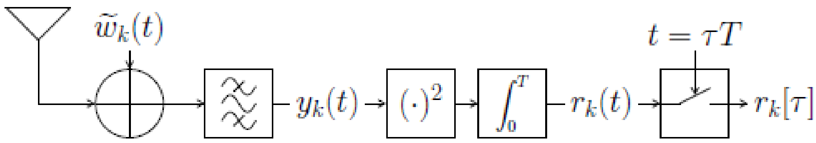

Figure 3.2-3 - Block diagram of one Rx RF chain.

3.2.3. Detection algorithm

Three detectors have been studied, namely Threshold Detection (TD), Maximum likelihood detector with Gaussian approximation (MLD-GA) and Neural-Network-based Detector (NND). While the two first ones are well known and straightforward to derive, the NND will be deeply described.

The motivation of using neural networks is twofold. First, as discussed previously, the considered system communicates through a nonlinear MIMO and neural networks are efficient to solve multi-variables non-linear problems. Moreover, in the case of the MLD-GA algorithm, the impact of noise on received symbols is assumed Gaussian, assumption we do not make using the NND. In addition, the NND does not explicitly need the propagation matrices H and Φ to estimate symbols as it learns the channel features during the training phase.

The architecture of the NND is illustrated in Figure 3.2-4. It is worth noticing that the NND uses one neural network by receive antenna. Each of the N_t neural networks estimates a single transmitted symbol. This neural network uses fully connected hidden layers N_{hl}, each one composed of N_n neurons using rectified linear unit (ReLU) as activation function. The number of hidden layers is function of the number of transmit antennas.[1] The input layer has N_r units, each one representing a received symbol r_k. Finally, a prediction s_k – homologous to the probability $Prob(s_n = 0 \mid r)$ – is produced at the output layer with a sigmoid unit, i.e a non linear function defined by $f(x) = \frac{1}{1+e^{-x}}$. Thus, we build a multi-layer perceptron classifier in order to estimate thee transmitted symbols s_n. It should be mentioned that this architecture is replicated N_t times. Besides, it is worth mentioning that optimizing jointly the N_t neural networks is more complex, and we have not observed any performance improvement with respect to the unitary optimization. Therefore, we consider the unitary optimization of the

---

[1] It shall be noted that the choice of these parameters is empirical and we have noticed some relations between the NND parameters and the system parameters such as $N_{hl} = N_t/2$ and $N_n = N_t^2$.

neural networks in this paper. Eventually, the use of the NND is relevant for high-rate applications requiring parallel processing since sent symbols are estimated independently.

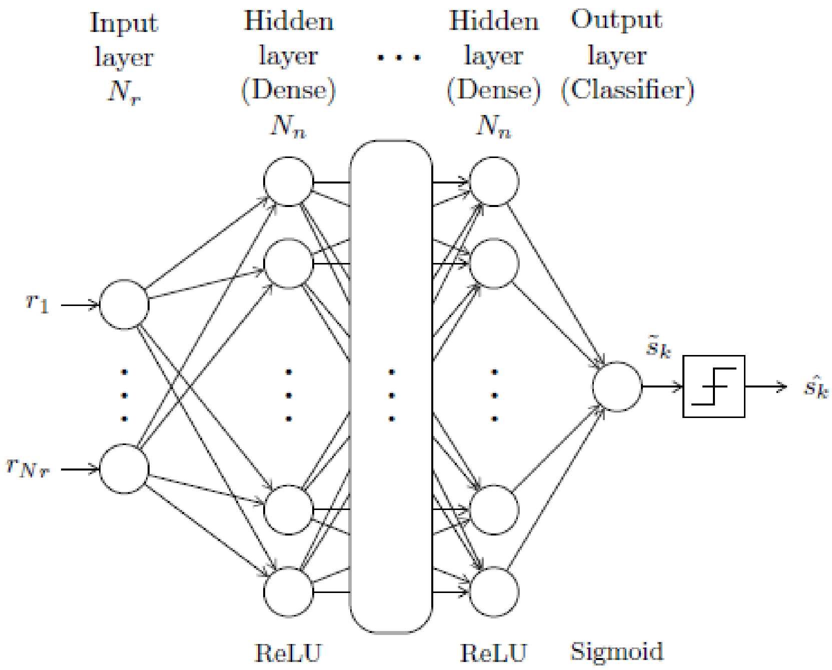

Figure 3.2-4 - Architecture of the n-th neural network of the NND.

The training of the NND is realized by transmitting some reference symbols known by the receiver. The set of weights is optimized during the training phase to maximize the detection performance of the NND. Each neural network, one per transmit antenna, is trained independently using an *Adam* optimizer, an algorithm for first-order gradient-based optimization of stochastic objective functions, based on adaptive estimates of lower-order moments [KB17]. The NND presents the advantage of having one loss function to optimize per transmitted symbol.

3.2.4. Main results

To evaluate the performance of the detectors, we first investigate MIMO systems without spatial interference. The channels are perfectly spatially multiplexed, i.e. H is diagonal. To implement the TD decision rule in practical systems, the threshold λ_{opt} has to be evaluated efficiently. Therefore, we use the expression of λ_{opt} proposed in [PAU14] using a polynomial approximation. Figure 3.2-5 presents the results of numerical simulations for systems using OOK with $N_t = N_r$ and no spatial interference. The communication performance is expressed in terms of bit error rate (BER) as a function of the SNR. We can remark that the TD and the NND present a performance gain in comparison to the MLD-GA decision rule. This performance loss of the MLD-GA may result from the Gaussian approximation of the channel. Furthermore, for transceivers without spatial interference, it can be shown that the NND achieves the optimal detection performance given by the TD.

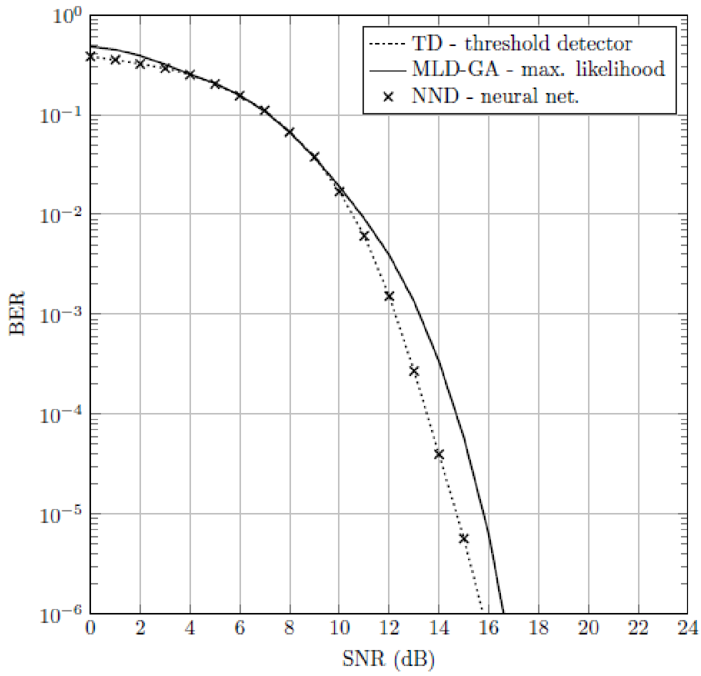

Figure 3.2-5 - Detection performance without spatial interference.

Then, we introduce a realistic mmWave scenario. The targeted application is a fixed indoor wireless link in the D-band. Table 3.2-1 outlines the main simulation parameters for this scenario. The considered system uses a uniform linear array (ULA) of antennas with $N_t = N_r = N$. The disposition of the antennas is depicted in Figure 3.2-6. The specification of the antenna is extracted from [MCG20], which describes the design of a high-gain antenna for the D-band based on transmit arrays.

Table 3.2-1 - Simulation parameters.

Parameters	**Notation**	**Values**
Carrier frequency	f_c	145 GHz
Symbol rate	$1/T$	1 GHz
Bandwidth	$B = 2/T$	2 GHz
Thermal noise	N_0	-174 dBm/Hz
Noise figure	N_f	10 dB
Antenna gain	g_0	32 dBi
Beam width	θ	3 °
Side lobe level	ϵ	-20 dB
Distance Tx - Rx	d_0	10 m

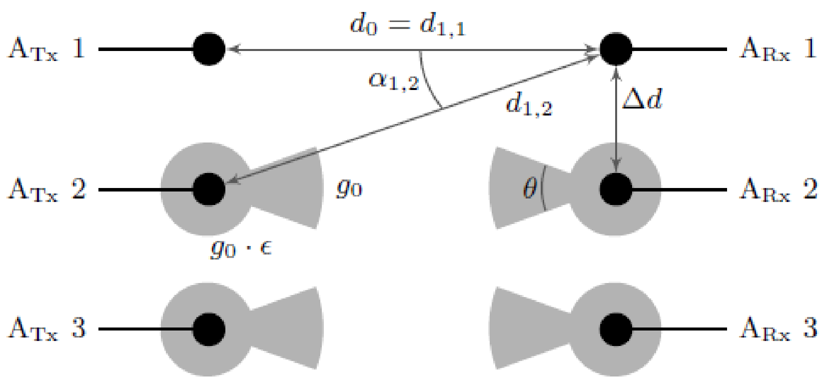

Figure 3.2-6 - Disposition of the antennas in the scenario.

Table 3.2-2 presents the parameters of the NND architecture for the different system configurations. Here the tuning of the NND parameters is empirical, yet we can admit that increasing the number of antennas induces an increase of the number of hidden layers and neurons. The results of numerical simulations are depicted in Figure 3.2-7 and Figure 3.2-8. The BER performance is assessed for $N = 4$, $N = 8$, and different levels of spatial interference. First, we can notice that, for $\kappa = 1$[2] (low interference level) with any number of antennas, the performance equals the one without interference. Second, for $N = 4$ with spatial interference, it can be observed that the MLD-GA and the NND present similar detection performance. Though, the NND demonstrates slight performance gains. Nonetheless, it should be noted that if the spatial interference is too strong, e.g. $\kappa = 5$ for $N = 4$ or $\kappa = 9$ for $N = 8$, the BER reaches an error floor. Third, for $N = 8$, we notice that the system performance of the NND is close to the one of the MLD-GA if the spatial interference is not too strong. With strong interference, $\kappa = 9$ and $N = 8$, using the NND leads to a significant performance loss. It should be mentioned that increasing the number of neurons in that case does not induce performance gain meaning that with this architecture we may not expect better performance for $\kappa = 9$ and $N = 8$.

Table 3.2-2 - NND parameters.

Parameters	**MIMO no interference**	4×4 **MIMO**	8×8 **MIMO**
Hidden layers (HL)	1	2	4
Neurons/HL	8	16	64
Batch size	32	32	256
Epochs	50	150	150
Optimizer	Adam		
Learning rate	0.001		
Loss function	Binary cross-entropy Eq. (16)		
Activation functions	ReLU and Sigmoid		

[2] Parameter κ is not described here. For the sake of this document, it is enough to know it indicates the level of spatial interference (growing level for k from 1 to 9). For a formal definition of κ, the reader is referred to [BFD+20] .

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

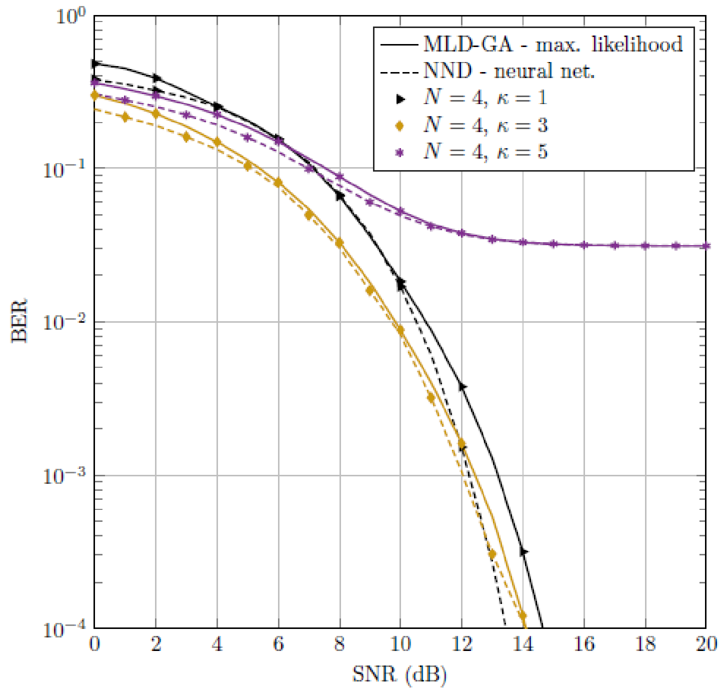

Figure 3.2-7 - Performance of the MLD-GA and the NND for N = 4.

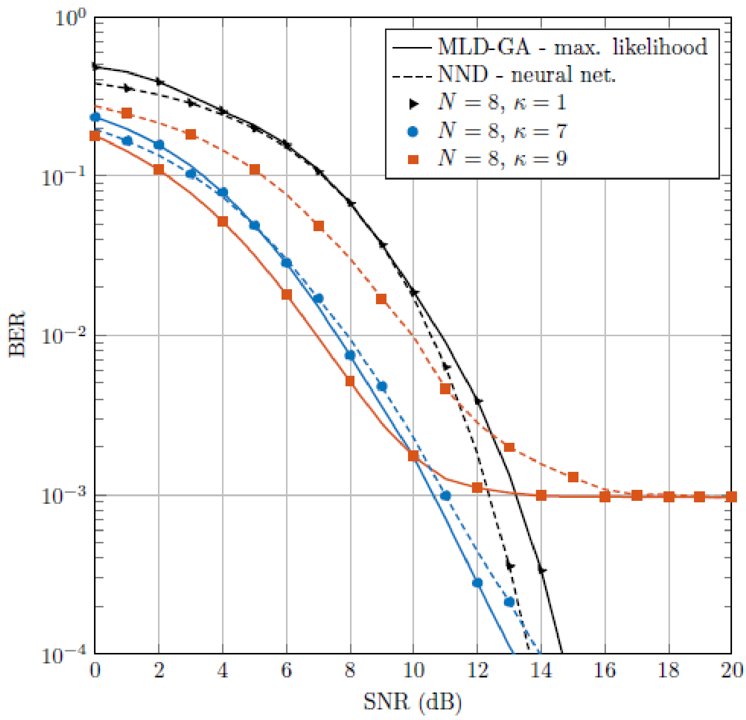

Figure 3.2-8 - Performance of the MLD-GA and the NND for N = 8.

First, and similarly to linear MIMO systems, the considered system benefits from the spatial diversity. Sent symbols of a transmit-antenna may be received on several receive-antennas. This property thus improves the robustness to thermal noise. It justifies the performance gains achieved by systems with strong interference in the moderate SNR regime in comparison to the multiplexed case. In addition to diversity, the studied system is also subject to ambiguity. Since the communication channel is a nonlinear MIMO channel, different transmitted MIMO symbols may lead to similar received observations. For this reason, the BER of system configurations with strong spatial interference reaches an error floor. Subsequently, we claim that the level of spatial interference, and hence the inter-antenna spacing, is directly related to a trade-off between diversity and ambiguity.

Second, it is also worth mentioning that significant differences in channel qualities exist. The different spatial streams – corresponding to a pair of aligned transmit and receive antennas – are not subject to the same interference. The numbers of symbols strongly interfering is larger for the receive antennas in the middle of the ULA than for the antennas on the extremities. Subsequently, the interference is stronger for the antennas in the middle of the ULA than the ones on the extremities. We will see in the next section that this property can be exploited by the channel coding scheme to enhance the system performance.

Eventually, our results show that low-error rate communications can be achieved with strong spatial interference between channels. Still, the detection algorithm must use jointly the different received symbols to demodulate transmitted symbols, similarly to the NND or the MLD-GA. Our results demonstrate that spatial multiplexing with non-coherent mmWave transceivers can be realized on strongly correlated LOS channels.

It is interesting to consider the integration of a FEC scheme to achieve channel coding gain and low error rates. However, implementing the FEC, and in particular its decoder, may entail a significant complexity and power consumption. To achieve a low-complexity low-power transceiver, we propose here to use a Bose, Ray-Chaudhuri and Hocquenghem (BCH) code. The considered FEC scheme is a BCH code with a packet size of 63 bits, a coding rate ranging from 0.4 to 0.9 and a decoder based on hard decisions. It should be mentioned that the key features of this code are a low-complexity implementation and a low-power consumption [MHZ11], opposed to LDPC or polar codes, for instance, for which high-data-rate architectures are very costly. In addition, with regard to the short packet size, this code has a low-latency decoder. These features appear to be highly relevant for the investigated scenario presented in Section 3.2.1. The considered transceiver architecture with the integration of a channel coding is presented in Figure 3.2-9. Multiple FEC schemes are used and the coding rate can be adapted to the channel quality of the receive antenna. The channel presents significant differences in terms of quality depending on the receive antenna. Receive antennas in the middle of the ULA are subject to stronger interference than the ones on the extremities. For this reason, adapting the coding rate to the receive antenna enables us to capitalize on the latter property to further enhance demodulation performance. The system architecture in Figure 3.2-9 is particularly interesting as it also maintains a high degree of parallelism. Figure 3.2-10 presents the achievable rates as function of E_b/N_0 for systems with a BCH code such that the BER is below 10^{-6}. The BCH code is implemented with a coding rate ranging from 0.4 to 1 and a channel decoder based on the hard decisions produced by the MLD-GA. Numerical results have been obtained through the Monte Carlo simulations with the real antenna gain diagram. First, it should be remarked that integrating a FEC scheme enables to achieve a significant channel coding gain. Second, it can be remarked that the adaption of the coding rate to the antenna leads to performance gains in comparison to setting a fixed coding rate for all antennas. In particular, we can see that for $N = 4$ and $N = 6$ the performance gains are larger than 2 dB.

To present the results of Figure 3.2-10 *differently, we propose in*

Table 3.2-3 a synthesis of the system performance and parameters for different number of antennas. For all system configurations, the system bandwidth is $B = 2$ GHz, the distance Tx-Rx is $d_0 = 10$ m, and the coding rate of the BCH is 0.9. Though the system performance is evaluated with the MLD-GA, similar results are expected using a demodulation based on the NND. Further, a channel bonding scheme, aggregating several sub-bands, could increase the throughput and allow to benefit from the large available free spectrum offered in sub-THz bands. It can be concluded that MIMO systems using ED receivers may achieve high rate communications in sub-THz bands with low-power.

Performance of coded systems could be further improved by considering longer packet length, soft-decision channel decoding, or capacity-achieving codes – e.g. a polar code – yet at the detriment of a complexity increase.

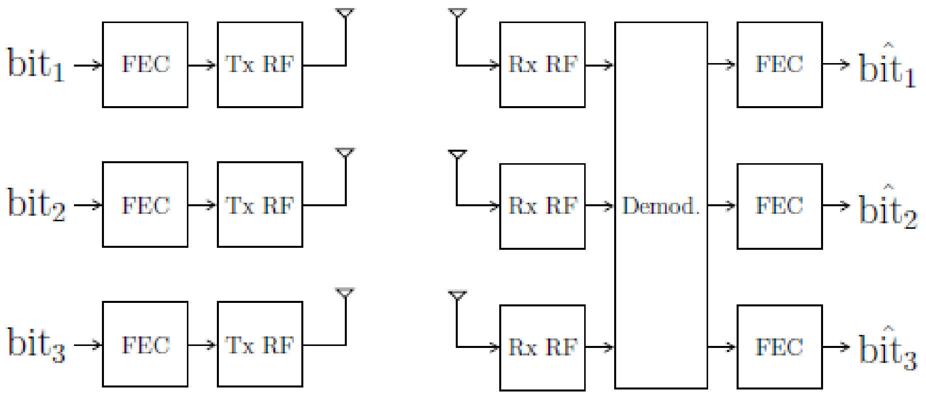

Figure 3.2-9 - System architecture integrating a FEC scheme.

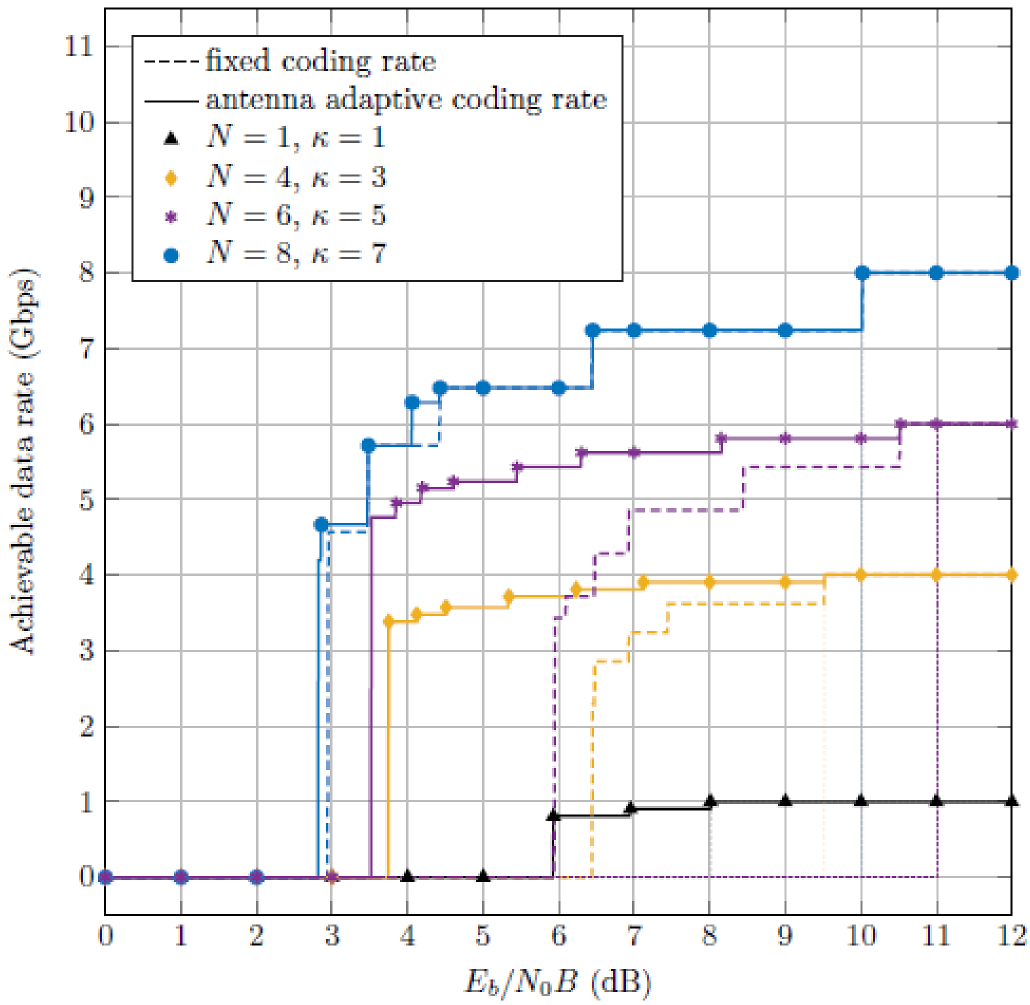

Figure 3.2-10 - Achievable data rate with a BCH code and the real antenna gain.

Table 3.2-3 - Synthesis of the main parameters and key performance indicators.

Carrier frequency	f_c	145 GHz			
Bandwidth	B	2 GHz			
Propagation distance	d_0	10 m			
Antenna gain	g_0	32 dBi			
Number of antennas	N	1	4	6	8
Throughput	$N/T \cdot 0.9$	0.9 Gbps	3.6 Gbps	5.4 Gbps	7.2 Gbps
Power by antenna	$P_{\mathrm{A_{Tx}}}$	-31.8 dBm	-31.2 dBm	-30.4 dBm	-32.3 dBm
Transceiver width	ℓ	5 cm	44 cm	50 cm	55 cm
Inter-antenna distance	Δd	$\varnothing$	13 cm	9 cm	7 cm

3.2.5. Discussions

This section points out a quick overview of the practical aspect of MLD-GA and NND techniques to understand the benefits and drawbacks of each technique.

The first difference between the MLD-GA and NND is that to perform its decision rules, the MLD-GA algorithm requires an explicit knowledge of H, of the Gaussian noise variance, and Φ. Hence, also the design of a channel estimation algorithm. In contrast, the NND is able to learn the channel features, and implicitly the channel matrices, during the training phase in order to demodulate symbols. The transmission of reference symbols is required for both detection algorithms which results in a spectral efficiency loss.

It is delicate to discuss the complexity of MLD-GA and NND algorithms since they lie in different paradigms. The MLD-GA algorithm is a common detection method. Though the complexity of the MLD-GA is $O(|C|N)$ and increases exponentially with the number of antennas, the decision rule only requires the evaluation of simple weighted Euclidean distances. Moreover, the complexity of the MLD-GA depends on the order of the modulation scheme. Using an OOK, the resulting complexity is $O(2N)$, which is reasonable for practical implementation. The implementation of the MLD-GA in practical systems would likely use on a common digital signal processor. In contrast, the complexity of the NND depends on the size of the used neural networks. It means that increasing the number of layers would drastically increase the complexity. Thereby, in our case, increasing the number of antennas will also increase the complexity of the NND. The complexity of both training and inference stages should be differentiated. Indeed, the training stage involves many data and consuming resources whereas the inference stage could be quite simple. It can be envisaged pre-training or offline training to reduce the cost and the complexity of the receivers. Besides, the implementation of neural networks is currently a largely investigated research topic. Thus, implementing neural networks on dedicated hardware such as GPU, FPGA and ASIC can significantly decrease computation time regarding the chosen neural network architecture.

Eventually, it is interesting to mention that RF components of the transceiver might present additional nonlinearities, e.g. quantization, RF power amplifier. In particular, envelope detectors based on diodes may present non-ideal square law response. It is expected that the NND might learn these channel nonlinearities and still demodulates symbol efficiently. The MLD-GA does not consider any other nonlinearities, such that it might be sensitive to these impairments and might require additional modeling. Nevertheless, it is of practical interest to characterize the detector robustness to imperfect RF components.

3.2.6. Implementation of the proposed neural network MIMO detector on a real target

The numerical results proposed so far were obtained via Matlab simulations. Matlab is an environment with its own programming language for directly expressing mathematics in arrays and matrices. Its API is user friendly but the drawback is the relatively long processing time: in particular, this application could not be embedded in a real-time system as it is. In the context of the activity of Task 3 of WP2, CEA and VSORA plan to collaborate and leverage VSORA's cutting-edge development tools and expertise to implement the neural network detector described above onto a real DSP target,. The details of the implementation and the consequent evaluation results will be described in the deliverables of Task 3 (D2.5 [D2.5] and D2.6 [D2.6]), but we give here a first presentation of the deployment constraint and the development flow.

In general, porting a numerically validated algorithm to a real DSP target introduces new constraints:

- **Real time constraint**

  The application is a stream that must be processed periodically. If the processing time is higher than the sampling period, the system will lose packets. The experimental study will explore this constraint and will determine the processing needs (basically the number of ALUs of the system). VSORA's DSP is scalable: we can increase the number of processing units to fit the requirements.

  Another axis of research is the coding style: the DSP has some features (such as the combination of several instructions) which improve the processing capacity. This is theoretically recognized and optimized by the compiler released with the development kit, but in some cases, optimization patterns are not recognized. A fine study of the embedded code could improve the coding style and finally gain performances.

- **Memory constraint**

  The DSP contains several memories: On one hand, a tightly-coupled memory (TCM) located in the core of the DSP is a memory directly connected to the processing units. This memory is fast but has a low density. On the other hand, D-RAM, used as peripheral, can store large amount of data. This memory is dense (small footprint) but it is slow. The connection between these two types of memory is done via direct memory accesses (DMAs). By targeting a real DSP, we will study the needs for both types of memory, the number of DMAs to be implemented and the limitations introduced by them.

- **Quantization constraint**

  The quantization is the operation which cuts numbers to be stored in memory (precision limitation or saturation). This is done on all computer systems, including software like Matlab: for instance the result of $\cos(\pi/2)$ executed by Matlab is not equal to 0, but it gives a very small number. The IEEE consortium defined a standard (IEEE754) to represent floating point values. The common data type used is "double": in this case data is represented over 64 bits: the exponent coded over 11 bits, the mantissa coded over 52 bits plus the sign.

  The type "double" introduces a very small error of quantization, which can be neglected during the algorithm study phase. However, this quantization is too large: we do not need such precision on a real target and the memory footprint is too large.

  VSORA's DSP encodes data following the IEEE754 norm, but the field (exponent, mantissa) width is a static configuration (decided by the customer before the silicon design). The development tools allows to finely study the quantization impact on the system.

As depicted in Figure 3.2-11, the development kit on VSORA's DSP is organized in platforms and each platform has a dedicated goal. They all share the same source code. This allows to "navigate" between the simulation platforms very easily: if a problem is detected on the last simulation platform (FPGA), we can quickly modify the algorithm if needed.

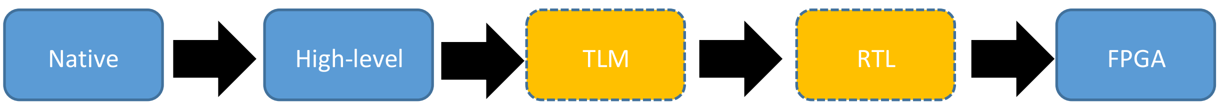

Figure 3.2-11 - VSORA's development platforms.

The **native** platform describes the system in C++. It is released with a mathematical library (vslib) developed by VSORA: it defines containers (matrix, vector,…) and associated operations (linear operations, multiplication, inversion,…). This library follows as far as possible the coding style of Matlab: the porting of code from the Matlab's environment to VSORA's environment should be easy and quick. This platform is used to develop the algorithm.

The **high-level** platform introduces the DSP: the code developed during the native phase is compiled with special options to be mapped on the target: the binary generated is the same for all lower level platforms (up to silicon). Simulations on the high-level platform generate reports about the processing cycles and the memory needs (rough estimation). This platform is also used to study the quantization since the DSP's model has the exact chosen precision.

The **TLM** (transaction-level modelling) and **RTL** (register-transfer level) platforms introduce more accurate models of DSP: These platforms are useful to develop the final silicon, we will not use them in the context of CPS4EU.

The **FPGA** platform is the final one: it runs application on "true" silicon. Only the core of the DSP is mapped over the FPGA. We use remote FPGA through the Amazon Web Service: this allows to use FPGA while optimizing the costs. This platform is used to validate the whole system and to run long simulations.

The final goal of CPSs is to be embedded in autonomous devices. New features and new approach in communication systems will require to be implemented on efficient DSP. However, this implementation could have an impact on the algorithm itself, since the real time aspect, the memory issues and the quantization effect are hard constraints. VSORA's development flow allows to manage these constraints globally and at early stage of development. Finally, simulations on a remote FPGA authorize long duration patterns. This will be investigated in "*the simulation tools and experimental platform"* deliverable [D2.5].

3.3. Stochastic geometry framework for ultra-reliable cooperative communications with random blockages

3.3.1. Motivation and related work

Traditionally, wireless applications in an industrial setup such as automated manufacturing, packaging, and on-field process monitoring employ wired communications, e.g., Ethernet-based solutions. These solutions can be expensive to deploy and cumbersome to maintain. Additionally, several of these applications require a highly flexible and dynamic communication infrastructure to support mobility. As a result, there is an increased interest to replace wired communication systems with wireless alternatives to reduce bulk as well as installation and maintenance costs.

Future industrial wireless networks will consist of a massive number of sensing, computing, and actuating devices and CPSs that communicate with each other with ultra-high reliability [LUSS18]. Furthermore, for several applications such as critical control of these devices, the delay constraints on the control loop latency are of the order of a few milliseconds [JPY+18]. As a result, the communication link latency must be of the order of a millisecond, while, simultaneously, maintaining ultra-high reliability. This comes under the purview of URLLC mode of operation in the 5G ecosystem.

On the downside, the presence of physical blockages or unfavourable channel conditions in the factory environment can severely degrade the received signal strength [Che16]. This may be detrimental, especially for URLLC applications. To address this issue, multi-hop transmission with cooperative relaying is proposed in [Swa+15], where devices with strong channel cooperate to provide a reliable communication link to devices in poor channel conditions. In this study, we extend the previous work by assessing this protocol in a realistic industrial propagation environment, considering the effect of random blockages on the system performance.

In order to enhance reliability, researchers have investigated time and frequency diversity, spatial diversity and multi-device cooperation. Particularly, Swamy *et al.* [Swa+15] have shown that with multi-device cooperation, ultrahigh reliability can be achieved even in low or moderate signal-to-noise ratio (SNR) regime. Furthermore, they show that the target error rate, and consequently the SNR necessary for achieving high reliability, cannot be achieved only by frequency diversity under realistic channel conditions. However, in their work, the authors do not take into account the instantaneous variation in channel conditions, and thus, their study results in a conservative choice of transmission rate set by the worst device's conditions.

In this regard, Jurdi *et al.* [JKV18] have proposed an adaptive rate selection scheme based on the instantaneous channel conditions. In their proposal, first, the channel states between the controller and the different nodes in the network are estimated. Then, the controller adapts the transmission rate for each node in the network to its instantaneous channel condition. In [BYW+15], the capacity and the coverage aspects of a factory automation setup is studied, by considering the effect of small and large scale fading. Interesting system level insights into the design of a wireless automation control network are presented, focusing on the antenna configuration. It is shown that equipping the devices with two antennas can dramatically reduce the required SNR for decoding and increase the capacity and the coverage of the network.

To the best of our knowledge, the effect of roaming blockages in a wireless factory automation setup has not been investigated in the literature. In a typical factory environment, the presence of metal surfaces and high density of industrial machinery determines the multi-path fading. Moreover, static or moving devices can cause an obstruction and strongly attenuate signal strength. The goal of our work is to assess device cooperation for reliable industrial wireless control, taking into account the effect of channel blockages. We model the location of the devices and the blockages as random processes and use tools from stochastic geometry to analytically characterize the probability of a typical node to be under coverage with and without cooperation. Specifically, pertaining to the blockage process, Matérn [Mat60] has studied two variants of repulsive point processes in his work. We use the Matérn hard-core process (MHCP) type II to model the locations of the centres of the blocking objects.

In the rest of the section, we will:

- Characterize the blockage statistics of the communication links in an industrial wireless network by modelling the position of the blockages as points of a MHCP. First, we use tools from stochastic geometry to characterize the SNR and the rate coverage probability of an arbitrarily located device in the network without multi-device cooperation.
- Adopt the two-phase communication protocol proposed in [Swa+15], to enable reliable industry automation services. In particular, we characterize analytically the rate coverage performance of a system using this cooperative protocol in an environment with random blockages. We also study the trends of the optimal value of the resource partitioning factor with respect to the rate coverage probability with varying blockage densities, device densities, and file sizes. More specifically, we show that, as the device density increases, more time should be allotted to the broadcast phase for maximizing the rate coverage if the data size depends on the number of devices. On the contrary, for a fixed data size broadcast, an increase in the number of devices necessitates an increase in the resources of the relay phase to limit the outage.

Notice that in this study we are focusing on downlink communications, for which the broadcast and relay paradigm arises more naturally. Nonetheless, similar techniques can be devised for uplink communications as well.

3.3.2. System model

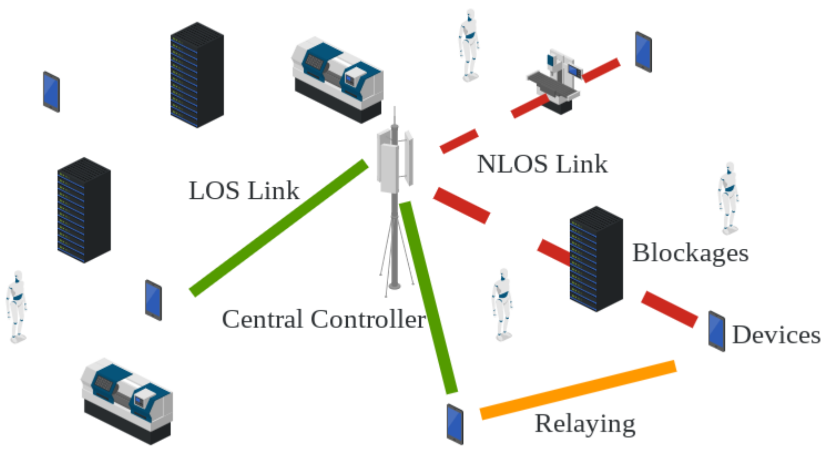

Figure 3.3-1 - Illustration of the considered system model.

We consider the factory area to be a closed bounded subset S of the two-dimensional Euclidean plane. For simplicity of the analysis, we assume S to be a disk of radius R. As depicted in Figure 3.3-1, we consider a multi-device wireless network where a controller located at the centre of S broadcasts data to multiple devices

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

distributed in the network area according to a homogeneous Poisson point process (PPP) ϕ with intensity λ. We assume that the wireless links can be blocked by physical objects circularly shaped with a radius R_B (see Figure 3.3-2), and whose locations are distributed in the network area according to a MHCP. This model, in contrast with classic PPP, ensures that there is a minimum distance between two nearby blockages. In the following, we carry out our analysis from the perspective of the test device in the network located at distance r_0 from the central controller.

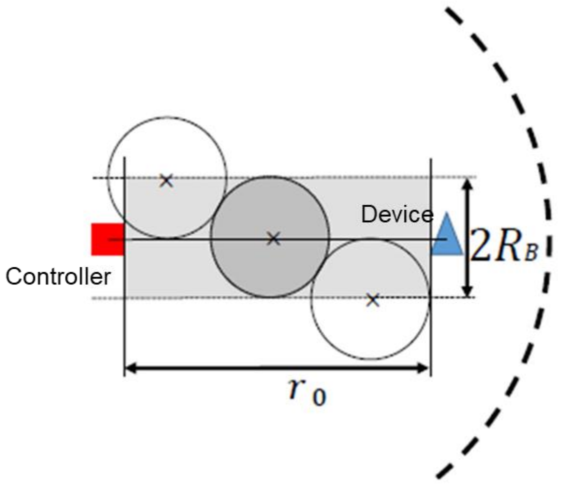

Figure 3.3-2 - Link-blockage illustration. The blue triangle is a device and the red square is the controller. The dotted line is the cell boundary.

From the perspective of the test device, the link to the controller is blocked by a circular blockage of radius R_B, only if any point of the circular blockage falls over the link connecting the test device and the controller. This is illustrated in Figure 3.3-2, where a blockage occurs if the centre of the circular blockage falls inside the shaded rectangular area (of size $r_0 \times 2\, R_B$).

Here we give a brief overview of the derivation of the MHCP, and then we present the intensity of such a process. An MHCP is formed from a dependent thinning of a parent PPP ϕ_P of intensity λ_P [CSKM13]. First, each point $x_i \in \phi_P$ is marked independently with a random mark $\mathcal{M}_i \in (0,1)$. Then, a point $x_i \in \phi_P$ is retained in the MHCP if and only if the ball of radius R_B and centred in x_i, $B(x_i, R_B)$ does not contain any point of ϕ_P with marks smaller than $\mathcal{M}_i$. Mathematically, the MHCP is described as:

$$\phi_B = \{x_i \in \phi_P : \mathcal{M}_i < \mathcal{M}_j, \forall x_j \in \phi_P \cap\ B(x_i, R_B)\backslash x_i\}.$$

It can be shown [GKD21] that the intensity of the blockage process is given by

$$\lambda_B = \frac{1 - \exp(-\lambda_P \pi R_B^2)}{\pi R_B^2}.$$

As depicted in Figure 3.3-2, the link between the controller and a node located at a distance r from the controller is blocked if a point of ϕ_B falls in the shaded area. As the thinning of ϕ_P preserves the Poisson nature of ϕ_B, the probability that the center of at least one blocking object falls in the shaded area can be calculated using the void probability:

$$\mathcal{P}_B(r) = 1 - \exp(-2\lambda_B r R_B).$$

We deploy a two-hop transmission policy, inspired by [Swa+15], where the total transmission time T is divided into two time-orthogonal phases and the overall bandwidth W is used in both the phases. Therefore, we consider that the two phases are split by a factor $0 \leq \beta \leq 1$, where the duration of phases I and II are respectively βT and $(1-\beta)T$. Furthermore, we assume that the downlink messages intended for the factory devices are concatenated together into a message of size b bits. Depending of the type of data, b can have a fixed value or be a function of the number of devices in the network i.e., $b = b_0 N_0$, where $N_0 = \lambda \pi R^2$ is the average number of devices in the network and b_0 denotes the message size in bits for each device.

During phase I, the controller broadcasts the concatenated message with rate $R_1 = b/\beta T$. Due to the presence of blockages and the impact of the path loss, the devices located farther from the controller are expected to have a higher probability of outage during phase I. To improve the coverage performance, we assume that the devices

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

under coverage in phase I, act as decode-and-forward relays to forward the message to the rest of the devices in phase II. Let the point process representing the locations of the devices under coverage after phase I (or the relays) be denoted by ϕ_1. In the second phase, a device in outage in phase I attempts to receive the message from the nearest relay from it. The nearest relay transmits the decoded message with a rate $R_2 = b/(1-\beta)T$. During phase II, each device under outage in phase I is assumed to receive the data from its closest device that was under coverage in phase I. We adopt such single closest relay assumption to make our stochastic geometry analysis tractable. Incidentally, the closest relay is statistically also the best relay due to the smaller chance of having its link to the device blocked by a random blockage and for a smaller path loss impact from the shorter distance.

We assume that the communication links experience a Rayleigh distributed fast fading with variance equal to 1. Furthermore, in case the transmission links are blocked by objects, the path-loss exponent is assumed to be α_N; otherwise for links in LOS, the path-loss exponent is assumed to be α_L. As a result, the received power at a device located at a distance r_j from transmitter j is given as $P_r = P_j h_j K r_j^{-\alpha_i}$, where K is the path-loss coefficient, P_j is the transmit power from transmitter j, and $j \in \{L, N\}$ depending on if the communication link is in LOS or non-line-of-sight (NLOS). Furthermore, h_j denotes the channel gain over the link from transmitter j and the receiver. We assume long packet communication, where a device is said to be in outage if the transmission rate exceeds the instantaneous channel capacity, and it is considered successful otherwise. In other words, a device is in outage if the received SNR is lower that a threshold imposed by the channel capacity. We denote the SNR thresholds γ_1 and γ_2 as functions of the time partitioning parameter β:

$$\gamma_1(\beta) = 2^{\frac{R_1}{W}} - 1 = 2^{\frac{b}{\beta T W}} - 1.$$

$$\gamma_2(\beta) = 2^{\frac{R_2}{W}} - 1 = 2^{\frac{b}{(1-\beta) T W}} - 1.$$

3.3.3. Coverage probability and frame design

Based on the SNR thresholds γ_1 and γ_2 and the corresponding outage definition, via the mathematical derivations described in all detail in [GKD21], it is possible to show that the SNR coverage probability of a device located at a distance r_0 from the controller in phase I is

$$\mathcal{P}_{C1}(r_0, \gamma_1) = \left[\exp\left(\frac{-\gamma_1 \sigma^2}{P_0 K\left(\mathcal{P}_B(r_0) r_0^{-\alpha_N} + \left(1 - \mathcal{P}_B(r_0)\right) r_0^{-\alpha_L}\right)}\right)\right],$$

where σ^2 is the Gaussian noise variance and P_0 is the transmit power at the central controller. Then, the point process ϕ_1 of the locations of the devices under coverage after phase I has intensity of $\lambda \mathcal{P}_{C1}(r, \gamma_1)$ at distance r from the controller.

Analogously, we derived in [GKD21] that the coverage probability of a device in the second phase is

$$\mathcal{P}_{C2}(r_0, \gamma_2) = \mathbb{E}_{r_2}\left[\exp\left(\frac{-\gamma_2 \sigma^2}{P_2 K\left(\mathcal{P}_B(r_2) r_2^{-\alpha_N} + \left(1 - \mathcal{P}_B(r_2)\right) r_2^{-\alpha_L}\right)}\right)\right],$$

where r_2 is the (random) distance of the nearest point of ϕ_1 from the device. The explicit probability density function of r_2 (that is a function of r_0) is characterized in [GKD21]. Using the theorem of total probability, and due to the independence of the random events of failure in phase I and II, the overall coverage probability after the two phases can be written as follows:

$$\mathcal{P}_C(r_0, \gamma_1, \gamma_2) = \mathcal{P}_{C1}(r_0, \gamma_1) + \mathcal{P}_{C2}(r_0, \gamma_2) - \mathcal{P}_{C1}(r_0, \gamma_1)\mathcal{P}_{C2}(r_0, \gamma_2).$$

To optimize the coverage, it is important to study the optimal time division between the two phases. When a small share of the total time T is allotted to phase I, only few devices that experience a high quality link are able to correctly receive the transmitted data, which limits the number of available relays in phase II. Whereas, in case a large amount of time is allotted to phase I, this leads to a high number of devices successfully decoding the transmissions in phase I but it requires to use a very high data rate in phase II, to relay the intended message. Therefore, we analyse the optimal time partitioning parameter $\beta \in [0,1]$ that minimizes the total probability of outage in cycle time T. Then, the problem of designing the optimal β to optimize the system performance can be mathematically stated as

$$\beta^* = \arg\min_{\beta \in [0,1]} \mathcal{P}_{out}(\beta),$$

Where the outage probability is defined as

$$\mathcal{P}_{out}(\beta) = 1 - \int_0^{2\pi}\int_0^R \mathcal{P}_C\big(r, \gamma_1(\beta), \gamma_2(\beta)\big)\, f_{r,\theta}(r,\theta) dr d\theta,$$

Where $f_{r,\theta}(r,\theta)$ is the uniform distribution of the location of the test-device, in which r refers to the distance of the device from the central controller and θ denotes the angle between the x-axis and the line joining the controller and the device.

3.3.4. Numerical results

Now, we present our most meaningful numerical results on the cooperative communication protocol analysed so far, with respect to the number of devices, density of the random blockages, transmit power, and the time splitting factor β. Other results are available in [GKD21].

Throughout simulations, we tune the blockage density using parameter λ_P, i.e., the density of the parent PPP from which the MHCP is derived (notice that one can easily verify that $\lambda_P \approx \lambda_B$ in the scenarios that we are investigating, therefore λ_P s an accurate estimation of the actual blockage density in the network). We assume a circular deployment area with dimension $R = 50$ m, a system bandwidth $W = 1$ MHz, carrier frequency equal to 3.5 GHz, $\alpha_L = 2$, and $\alpha_N = 4$. The blockages are considered to be of radius $R_B = 0.5$ m. Unless otherwise specified, we assume $P_0 = 23$ dBm and $P_2 = 20$ dBm.

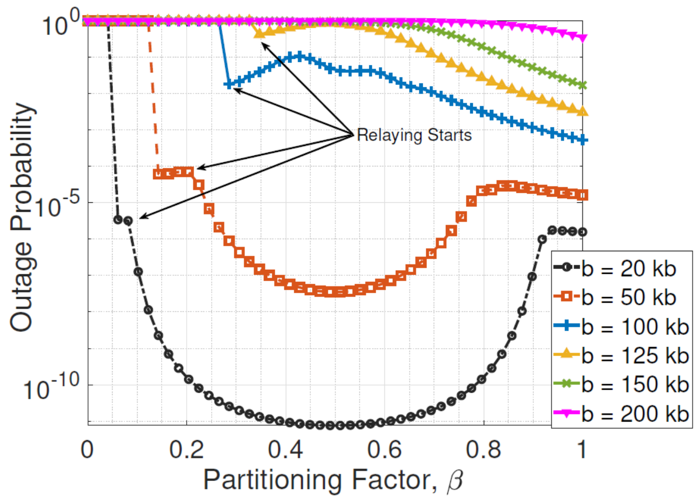

Figure 3.3-3 - Outage probability vs β for different values of the data size b. $\lambda = \lambda_P = 0.01/m^2$.

In Figure 3.3-3, we plot the outage probability with respect to the resource partitioning factor for different values of the data size b. Evidently, based on the size of the data packet, the trend of the outage probability with respect to β follows one of the following. First, for small packet sizes (e.g., $b = 20$ kb or 50 kb), the two-phase strategy with relaying provides a better outage probability than single-phase broadcast strategy. Therefore, optimizing β has a meaningful impact on the outage probability, which reaches an optimal value near $\beta = 0.5$ with equal partition of resources. Second, for large packet sizes (e.g., $b = 100$ kb and more), the gain of the two-phase strategy is marginalized, resulting in optimal partitioning of $\beta = 1$. In particular, we see that for $b = 150$ kb and $b = 200$ kb, the outage monotonically decreases with β, indicating that it is directly dependent on the resources allotted to the phase I. It should be noted that for intermediate data sizes, e.g., 100 kb and 125 kb, we observe an initial decrease in the outage probability with the onset of relaying. This is followed by an increase in the outage. Indeed, for these data sizes, although an increased number of devices get coverage in the broadcast phase with an increase in β, a reduced amount of resources are available for successful relaying of a file with large data size. After a threshold on β, the outage reaches the regime where the outage decreases with increasing the length of phase I.

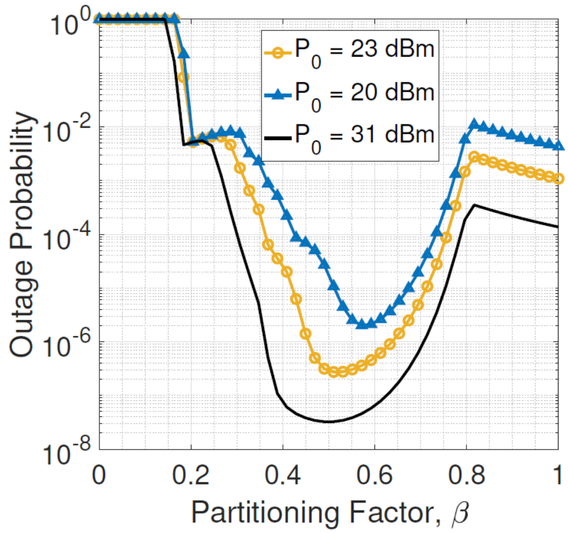

Figure 3.3-4 - Outage probability vs β for different values of controller power. $\lambda_P = 0.1/m^2$ and $\lambda = 0.05/m^2$.

In Figure 3.3-4, we plot the outage probability with respect to β for different values of controller transmit power in a dense blocking environment ($\lambda_P = 0.1/\mathrm{m}^2$). We observe that the controller power has an effect on the optimal value of β. In particular, as the transmit power of the controller decreases, the optimal value of β shifts towards 1. This indicates that for a lower transmit power of the controller, a larger amount of resources need to be allotted to the broadcast phase so as to minimize the outage. Naturally, for lower P_0, we observe a higher value of outage.

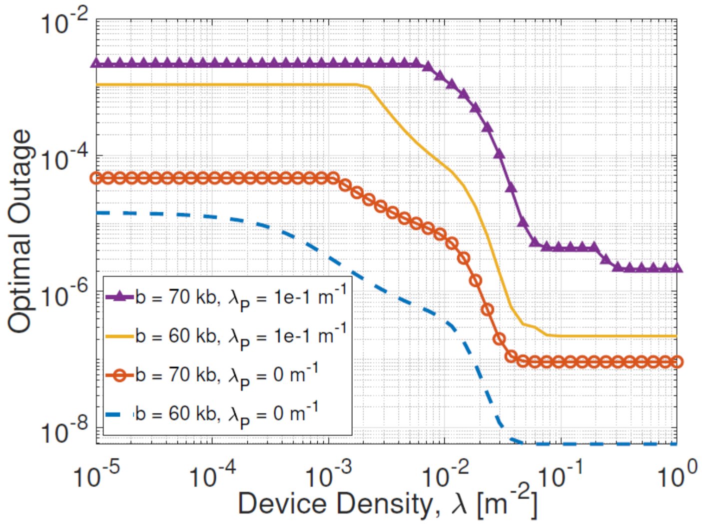

Figure 3.3-5 - Optimal outage probability with fixed data size.

The optimal outage values achievable in our scenario are shown Figure 3.3-5. In the "broadcast regime" (i.e., when the device density is still too small to have efficient relaying), the optimal outage is constant up till a certain λ for a given b and λ_P, for which a cooperative transmission is beneficial. For a large value of device density (e.g., $\lambda \geq 0.5/\mathrm{m}^2$), increasing λ does not bring the nearest device significantly closer to the test device. Hence, relaying cannot increase its effectiveness and the optimal outage saturates.

3.3.5. Conclusions

We have investigated an industry automation scenario where a central node communicates with wireless devices on the factory floor and the communication can be impaired by the presence of physical blockages. Then, to mitigate the service outage, we proposed and characterized a resource partitioning scheme in which the communication resources are divided into broadcast and relay phases so as to facilitate successful transmission of the intended data. Our study highlights that when the data size depends on the number of users, the optimal resource partitioning factor is tightly coupled with the device density. In that case, the optimal solution is to increase the length of the phase I with the device density. Whereas, when the data size is fixed, the optimal resource partitioning factor shows a non-trivial behavior due to the contending effects of increasing β, which increases the relay density, but also decreases the resources available to these relays.

3.4. D-MEC: discontinuous mobile edge computing

3.4.1. Motivation and related work

With the advent of 5G (and beyond) networks, we live at the edge of a revolution of mobile communication systems, which are evolving from a pure communication framework to service enablers, building on the tight integration of communication, computation, caching, and control. Indeed, future networks will enable a plethora of new services, not only to mobile end users, but also for whole different sectors (verticals), such as Industry 4.0, Internet of Things (IoT), autonomous driving, remote surgery, etc. These new services have very different requirements and they generally involve massive data processing within low end-to-end delays (in the order of ms). Among several technology enablers at different layers (e.g., artificial intelligence, network function virtualization, millimeter-wave communications), a prominent role will be played by Edge Computing, whose aim is to move cloud functionalities (e.g., computing and storage resources) at the edge of the network to avoid the relatively long delays necessary to reach central clouds. Edge computing is also the object of an ETSI Industry Specification Group, called Multi-Access Edge Computing (MEC) [MEC]. MEC is foreseen to enable several novel applications and use cases, relying on the enhanced performance of new 5G technologies, due to the massive volume of data to be transferred within low-latency and/or extremely high reliability constraints [PTSD18].

In the context of CPS4EU and MEC applications for cyber-physical systems, we focus on computation offloading, a way to transfer the execution of applications from mobile devices (or sensors in IoT environments) to a nearby edge server (ES) [MB17]. Computation offloading helps in reducing the User Equipment's (UE) energy consumption and/or the overall delay of the service. When an application is offloaded, the overall service time is composed of the transmission time of input data, the computation time at the ES, and the time needed to send the result back to the UE [SSB15]. In edge-computing-aided networks, a critical aspect for real-life implementations is the limited energy made available by the battery at the mobile device, the need for frequent battery recharge, and the high energy consumption of network elements, due to the dense deployment of Access Points (APs) and ESs necessary to enable the described ecosystem. In traditional mobile networks, a large portion of the power is consumed at the AP site [Aue+11], [Tom+15]. With the deployment of ESs, the power consumption will certainly increase, so that new methods are essential to reduce the impact of the ICT industry on the global carbon footprint. In such a context, the main target of this contribution is energy efficiency.

In the context of mobile networks, several works focus on novel strategies to reduce system power consumption. In general, it is well-known that a large portion of the power is consumed by the AP only for being in active state (RF chains, power amplifiers, cooling, etc.) [Tom+15]. Thus, most of the works in the literature propose strategies based on possible ON/OFF behaviour of the APs, known as Discontinuous Transmission (DTX), by which some components of the AP are put in low-power sleep states when possible, e.g., in case of low traffic. In the context of edge computing and computation offloading, there exists a wide literature on energy efficiency. In particular, [MZSL17] proposes a dynamic computation offloading strategy based on stochastic Lyapunov optimization to reduce a weighted sum of UE and ES power consumption. [LBDP19] extends the strategy to a multi-server multi-cell scenario, introducing average delay and reliability constraint on the queue lengths. In [MDBF20], the authors propose a joint dynamic computation offloading strategy with reliability guarantees, incorporating URLLC and energy harvesting devices. All these works mainly focus on power consumption at the UE and ignore the network. [HCF20] proposes a dynamic strategy aimed at minimizing the average power consumption of mobile devices, under a latency constraint and a constraint on the ES average power consumption, without considering the AP. Recent contributions consider the energy consumption of both radio access and MEC network. In particular, in [LGJG19], a scheduling strategy is proposed to find a trade-off between task completion ratio and throughput, hinging on Lyapunov optimization. [CZX17] aims at minimizing the long-term average delay under a long-term average power consumption constraint. In [WZYW19], the long-term average energy consumption of a MEC network is minimized under a delay constraint, using a MEC sleep control. Also, in [CM18] the problem is formulated as the minimization of the energy consumption under a mean service delay constraint, optimizing the number of active base stations and the computation resource allocation at the ES, while considering a sleep mode for both APs and ESs. In [Nan+17], Lyapunov optimization is used to reduce the energy consumption of a fog network while guaranteeing an average response time. All these works do not consider a holistic management of AP sleep control, radio resource allocation, ES sleep control and CPU scheduling, UE sleep control and energy consumption, and end-to-end delay constraints.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

In our study, we extend and improve our preliminary results of [MdD+19]. We propose a dynamic computation offloading strategy based on stochastic Lyapunov optimization that minimizes the weighted sum of UEs', AP's, and ES's long-term average energy consumption, under an average end-to-end delay constraint and a reliability constraint. The latter is defined as the probability that the end-to-end delay exceeds a prescribed threshold. These constraints are handled through the definition of an uplink queue of data to be offloaded by each UE, a computation queue at the ES, and a downlink queue of results at the AP. These constraints translate into a constraint on the average length of the sum of the three queues and a probabilistic bound on the maximum total queue length, as in [MdD+19]. However, differently from [MdD+19], we introduce the sleep mode operation at the UE's side and an adaptive algorithm to translate the probabilistic constraint on the queue lengths into a reliability constraint on the actual end-to-end delay. Our strategy does not require any a priori knowledge of the statistics of the radio channels or of the data arrivals. In particular, starting from a non-convex non-differentiable long-term average optimization problem, we devise an algorithm that solves a deterministic problem on a per-slot basis. The proposed optimal solution of each deterministic problem has very low computational complexity and can be found via fast iterative algorithms. Several numerical results show the performance of our strategy, also compared with other possible methods. We highlight that the novelty of our work lies in the joint optimization of the radio resources, computing resources, and duty cycles of all the network elements, taking into account performance constraints and mathematically showing that the optimal solution is reached.

3.4.2. System model

To capture the dynamic aspect of the problem, we consider a transmission of packets organized in time slots $t = 1, 2, 3, \ldots$ of equal duration τ_l. In the following, we present: the UEs', AP's, and ES's energy consumption model; the queueing model used to handle the delay constraints; the reliability performance over the radio interface in terms of Packet Error Rate (PER).

Computation offloading generally entails three phases: an uplink phase, where the UE sends data to the AP, a computation phase at the ES, and a downlink phase for the transmission of the results to the UE. In our dynamic scenario, the overall slot duration τ_l is divided into two portions: a period of τ_s seconds dedicated to control signalling and a period of τ seconds for the actual three phases of computation offloading Here, we assume that control signalling happens before computation offloading due to the need of exchanging the state variables necessary to run the optimization algorithm to allocate radio and computation resources. Then, the total duration of the time slot is $\tau_l = \tau_s + \tau$. We assume that the AP, the ES, and the UEs can be put into a low-power sleep state for energy saving purposes during the slot fraction reserved to offloading (not for the whole slot duration due to the need for control signalling). When in sleep state, the AP and the UEs cannot receive nor transmit and the ES cannot process data, thus consuming less power. Our goal is to optimize the long-term fraction of time that the entities spend in sleep state with the aim of minimizing the overall system energy consumption, guaranteeing a targeted Quality of Service (QoS) measured by the overall delay of the computation offloading strategy.

Nowadays, around 80% of the total power consumption of the wireless networks is consumed at the AP [DDL15], which consumes a considerable fraction of its total power only for being in active state [Aue+11], [DDL15]. Among the different models available in the literature, we exploit that of [DDL15], which provides a tool, available online, to model the power consumption of base stations of different kinds, with details on the specific components (power amplifiers, supply power, etc.). Let us denote p_a^{on} the overall power consumption of the AP for being in active state. This parameter includes the consumption of power amplifiers, power supply, analogue frontend, digital baseband, and digital control. While in active state, the AP can transmit and/or receive. We denote $p^d(t)$ the overall downlink transmit power. According to the model, the AP can enter a low-power sleep mode to save energy whenever possible, without compromising the QoS. Here, we assume that the choice of when being active or sleeping is performed online. To control the active and sleep state of the AP, we introduce the binary variable $I_a(t) \in \{0, 1\}$, which equals 1 if and only if the AP is in active state at time slot t. In each time slot, the AP is forced to be active for the first τ_s seconds to perform channel state information acquisition and control signalling. For simplicity, we neglect the transmit power necessary for this reduced exchange of information, thus taking into account only the active state power p_a^{on} during the signalling period. Then, the AP energy consumption at time slot t is

$$E_a(t) = \tau\big(I_a(t)p_a^{on} + p^d(t)I_a(t) + \big(1 - I_a(t)\big)p_a^s\big) + \tau_s p_a^{on},$$

where p_a^s represents the (low) power consumed by the AP in sleep mode. The power consumed in the receiver chain is neglected, as it is typically much smaller than the other contributions.

As pointed out in [GSS15], the power management of a CPU is all about efficiently (and dynamically) controlling both current and voltage in order to minimize power while providing a desired performance. Power-saving techniques can be divided into two main categories: turn it off and turn it down. The first consists in switching off some components of the CPU, which are then put into low-power sleep state. In modern processors, there exist several possible idle states, called C-states [GSS15], which allow the processor to enter more or less deep sleep modes. Obviously, a deeper sleep mode provides higher energy savings, but requires more time to wake up. This defines a trade-off between energy consumption and latency. In this work, we adopt C-states operating on a specific core, dedicated to treat the offloaded tasks of all the UEs of our system. In particular, we consider two states: the C0-state, in which the CPU core is active and executing some thread, and the C1-state, in which the CPU clock frequency is driven to zero. Then, in our model, the CPU core consumes a power p_m^{on} just for being in active state (C0-state) and a power p_m^s in sleep state (C1-state). Moreover, when the ES is active, the dynamic power spent for computation is $p_m^c(t) = \kappa f_c^3(t)$, where $f_c(t)$ is the CPU cycle frequency at time slot t and κ is the effective switched capacitance of the processor. We suppose that it is possible to use dynamic voltage frequency scaling to scale down the frequency, thus reducing the dynamic power consumption. In particular, we assume that f_c can be selected from a finite set $\mathcal{F} = \{0, \dots, f_{max}\}$ and we introduce the binary variable $I_m(t) \in \{0, 1\}$, which equals 1 if and only if the ES is in active state. Then, recalling that $\tau_l = \tau_s + \tau$, the energy consumption in each time slot is given by

$$E_m(t) = \tau\left(I_m(t)p_m^{on} + \big(1 - I_m(t)\big)p_m^s + p_m^c(t)\right) + \tau_s p_m^{on},$$

Where $p_m^c(t) = 0$ whenever $f_c(t) = 0$, because $p_m^c(t) = \kappa f_c^3(t)$.

We also assume that all K UEs can switch their radio equipment to a low-power sleep mode whenever possible. In particular, we exploit the empirical model of [LNSM14], in which it is shown that a UE's overall power consumption is mainly determined by its cellular subsystem component, which is about $p_k^{on} = 0.9$ W, and it is also affected by transmit powers above 10 mW, consuming an additional $0.6 - 1.5$ W. We will denote by $p_k^u(t)$ the overall power consumption consumed for communication by UE k (for $k = 1, \dots, K$) at time slot t. According to [LNSM14], $p_k^u(t)$ is a monotone increasing function of the transmit power $p_k^{tx}(t)$ and it is almost independent from the uplink and downlink data rates, thus the sleep operation mode is the most promising solution to achieve energy efficiency. For the sleep operation, similarly to the AP case, two different states are defined [LNSM14]: a light sleep mode, with power consumption $p_k^s = 346$ mW and sub-millisecond transition time, and a deep sleep mode, with $p_k^s = 20.3$ mW and much longer transition time (around 10 ms). In this paper, we exploit the light sleep operation. Recalling that the UE is always active at the beginning of the slot for control signalling, the power consumption model of the UE reads as follows:

$$E_k(t) = \tau\big(I_k(t)p_k^{on} + p_k^u(t)I_k(t) + \big(1 - I_k(t)\big)p_k^s\big) + \tau_s p_k^{on},$$

where $I_k(t) \in \{0, 1\}$ is the binary variable that indicates the active or sleep mode of UE k. Hence, the total energy spent by the K UEs is simply given by

$$E_u(t) = \sum_{k=1}^{K} E_k(t).$$

Then, the total system energy consumption at time slot t is

$$E_{tot}(t) = E_u(t) + E_m(t) + E_a(t).$$

Now, we consider a dynamic scenario, in which new data units are continuously generated from an application at the UE's side and have to be offloaded and processed at the ES. To model the system dynamics, we use a simple queueing model, taking into account the three phases of computation offloading. This model allows us to characterize the total delay experienced by a data unit from its generation at the mobile side until the reception of its corresponding result, sent by the AP to the UE. In particular, the considered queueing model is depicted in Figure 3.4-1.

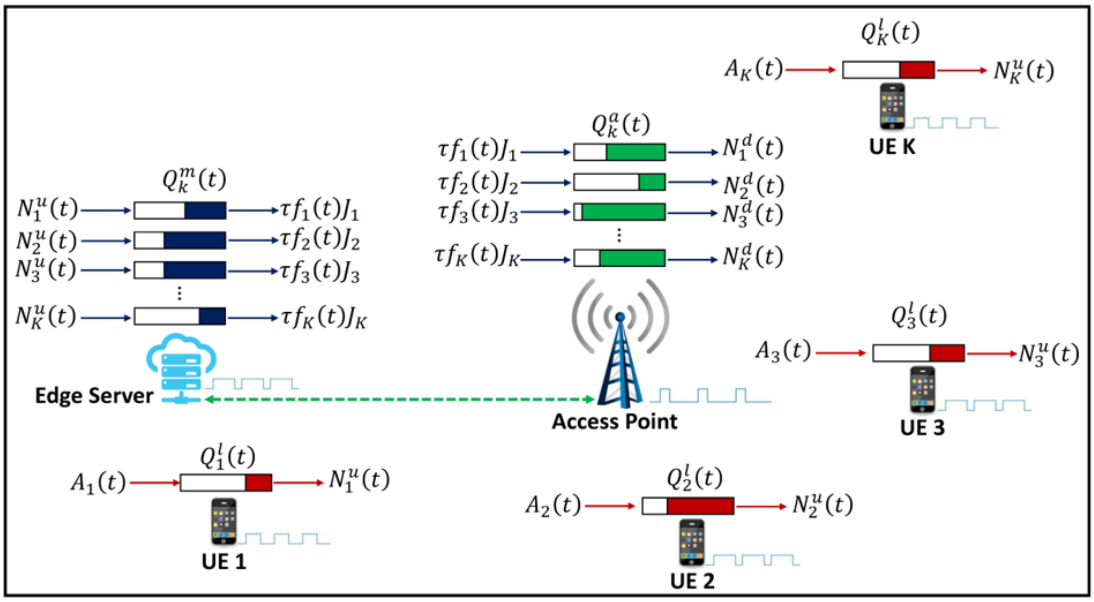

Figure 3.4-1 - Network model.

Accordingly, each data unit experiences three different delays: a communication delay, including buffering at the UE; a computation delay, including buffering at the ES; a communication delay, including buffering at the AP. As we will show later, we take into account these three sources of delay jointly. For the multiple access over the radio channel, we consider a simple Frequency Division Multiple Access, both for the uplink and the downlink.

In uplink, allocating bandwidth B_k^u to UE k, the symbol duration is $T_k^{s,u} = 1/B_k^u$. Since the time for data transmission is τ, under the assumption of a perfect pulse shaping, UE k can transmit the following number of symbols at time t:

$$N_k^{s,u}(t) = \left\lfloor \frac{\tau}{T_k^{s,u}} \right\rfloor = \lfloor \tau B_k^u \rfloor.$$

Assuming that bits are encoded against radio channel noise into packets of fixed length N_b bits, employing an M-QAM modulation, the number of packets transmittable at time t is given by:

$$N_k^{p,u}(t) = \left\lfloor \frac{N_k^{s,u}(t)\log_2\left(M_k^u(t)\right)R_k^{c,u}(t)}{N_b} \right\rfloor,$$

where $M_k^u(t)$ is the modulation order and $R_k^{c,u}(t)$ is the channel coding rate. In particular, we assume that the uplink Modulation and Coding Scheme (MCS) pair $m_k^u = \left(M_k^u(t), R_k^{c,u}(t)\right)$ is chosen from a discrete finite set $\mathcal{MCS}_k^u$. Also, we assume that a data unit has to be transferred in one time slot, i.e., it cannot be split and partially transmitted over different time slots. Thus, the number of data units that UE k can send at time slot t over the radio interface is

$$N_k^u(t) = \left\lfloor \frac{N_k^{p,u}(t)N_b}{S_k^i} \right\rfloor,$$

where S_k^i is the size in bits of an input data unit. Then, the local queue of data to be transmitted evolves as

$$Q_k^l(t+1) = \max\left(0, Q_k^l(t) - N_k^u(t)\right) + A_k(t),$$

where $A_k(t)$ is the number of newly arrived data units generated by the application running at UE k; $A_k(t)$ is modeled as a random process whose statistics are not known a priori.

We assume that the number of input data units processed by the ES to serve UE k is proportional to the number of CPU cycles allocated for this task. Given the computation rate $f_k(t)$ assigned to user k, measured in CPU cycles per second, and defining the coefficient J_k as the ratio between the number of processed data and the number of CPU cycles, the data queue waiting for being processed by the ES for UE k evolves as

$$Q_k^m(t+1) = \max(0, Q_k^m(t) - \lfloor \tau f_k(t)J_k \rfloor) + \min\left(Q_k^l(t), N_k^u(t)\right).$$

Finally, we define K queues at the AP, containing the computation results to be sent back to the UEs. We assume that every processed input data unit produces one output data unit, with size S_k^o possibly different from S_k^i. The queue evolves as

$$Q_k^a(t+1) = \max\left(0, Q_k^a(t) - N_k^d(t)\right) + \min(Q_k^m(t), \lfloor \tau f_k(t) J_k \rfloor),$$

where $N_k^d(t)$ is the number of data units sent back to user k in downlink:

$$N_k^d(t) = \left\lfloor \frac{N_k^{p,d}(t) N_b}{S_k^o} \right\rfloor,$$

where $N_k^{p,d}(t)$ is the number of packets sent in downlink, i.e.,

$$N_k^{p,d}(t) = \left\lfloor \frac{\lfloor \tau B_k^d(t) \rfloor \log_2\left(M_k^d(t)\right) R_k^{c,d}(t)}{N_b} \right\rfloor,$$

where $B_k^d(t)$ is the bandwidth assigned to UE k for downlink communication at time t, $M_k^d(t)$ is the downlink M-QAM modulation order, and $R_k^{c,d}(t)$ is the channel coding rate for downlink. As for the uplink, the pair $m_k^d = \left(M_k^d(t), R_k^{c,d}(t)\right)$ belongs to a discrete finite set $\mathcal{MCS}_k^d$.

As already mentioned, the overall delay experienced by a data unit is the time elapsed from its generation at the mobile side, to the moment the user receives back the result associated with it. Due to Little's law, the average overall service delay is proportional to the average queue length. Then, the overall delay is directly related to the sum of the uplink and downlink communication queues and the computation queue:

$$Q_k^{tot}(t) = Q_k^l(t) + Q_k^m(t) + Q_k^a(t).$$

In particular, given a data unit arrival rate $\overline{A_k} = \mathbb{E}\left\{\frac{A_k(t)}{\tau_l}\right\}$, the long-term average end-to-end delay experienced by a data unit generated by UE k is

$$\overline{D_k^{\infty}} = \lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\left\{\frac{Q_k^{tot}(t)}{\overline{A_k}}\right\};$$

the expectation is taken with respect to the random radio channel and data arrival realizations.

Our first aim is to guarantee a constraint on the long-term average delay D_k^{avg}, formulated as:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\{Q_k^{tot}(t)\} \le Q_k^{avg} = D_k^{avg}\overline{A_k}, \qquad \forall k.$$

Note that $\overline{A_k}$ is not known a priori, but we can estimate it online with a moving average (possibly with a forgetting factor, in the case of non-stationary scenarios). As a second objective, we want to ensure a long-term probabilistic constraint on the overall delay experienced by data units:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} Prob\{D_k(t) > D_k^{max}\} \le \epsilon_k, \qquad \forall k,$$

where D_k^{max} is a predefined threshold, $0 < \epsilon_k < 1$, and $D_k(t)$ represents the overall delay experienced by a generic data unit whose result is received back by UE k at time t. The aim of this constraint is to reduce the variance of the delay (i.e., the jitter). As mentioned before, there is a direct dependence between the overall delay and the overall queue length, therefore we can translate the previous inequality into the following probabilistic constraint on the sum of the queues:

$$\lim_{T\to\infty} \frac{1}{T} \sum_{t=1}^{T} Prob\{Q_k^{tot}(t) > \delta_k Q_k^{avg}\} \le \epsilon_k, \qquad \forall k,$$

with $\delta_k > 1$ conveniently chosen to convert the delay threshold into a queue-length threshold. In principle, there is no direct analytical relation between $\delta_k Q_k^{avg}$ and D_k^{max}, but we propose in [MdD+20] an online method to appropriately select and adapt δ_k. Finally, note that the previous inequality can be equivalently re-written as the expectation of a Bernoulli random variable as follows:

$$\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{u(Q_k^{tot}(t)-\delta_k Q_k^{avg})\}\leq \epsilon_k, \qquad \forall k,$$

where $u(\cdot)$ is the unitary step function. In the sequel, the event $\{Q_k^{tot}(t)>\delta_k Q_k^{avg}\}$ will be termed as "out-of-service", and ϵ_k will be the required out-of-service probability.

For robust radio communications, we suppose that both the UEs and the AP comply with the 5G New Radio standard recommendations for channel coding. Namely, the input data units transmitted in uplink and the output data units transmitted in downlink are both encoded to LDPC packets of size 1500 bytes, with MCS belonging to $\mathcal{MCS}_k^u$ and $\mathcal{MCS}_k^d$. To satisfy a target performance in terms of packet loss, we want to guarantee that the uplink and downlink PER, denoted respectively PER_k^u and PER_k^d for UE k, do not exceed some targeted thresholds θ_k^u and θ_k^d. In this sense, given the radio channel state at time t, recalling that the transmit power $p_k^{tx}(t)$ used by UE k is a function of the chosen MCS $m_k^u\in\mathcal{MCS}_k^u$, we define $p_k^{tx,min}(m_k^u,t)=\min\{p_k^{tx}(t): PER_k^u\leq\theta_k^u\}$. Analogously, the minimum power required for downlink communications towards UE k is $p_k^{d,min}(m_k^d,t)=\min\{p_k^d(t): PER_k^d\leq\theta_k^d\}$. A minimum target PER translates into a minimum target SNR $\overline{\gamma_k}$. Thus, the minimum transmit power is $p_k^{tx,min}=\overline{\gamma_k}N_0B_k^u/h_k^u$, where N_0 is the noise power spectral density at the receiver, and h_k^u is the uplink channel power gain. The same discussion is valid for the downlink transmission.

3.4.3. Problem formulation

Here we formulate our optimization problem, aimed at minimizing the long-term average weighted sum of the UE's, AP's and ES's energy consumption, whose value at time slot t is given by the convex combination:

$$E_{tot}^w(t)=\alpha_1E_u(t)+\alpha_2E_a(t)+\alpha_3E_m(t),$$

where $\alpha_i>0$ for $i=1,2,3$, and $\sum_{i=1}^3\alpha_i=1$, with the coefficients α_i chosen in order to explore alternative priority mechanisms assigned to different energy consumption sources (as explained in the section with numerical results). The long-term optimization problem is then:

$$\min_{\Psi(t)}\lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{E_{tot}^w(t)\}, \qquad \text{subject to:}$$

$$(a)\quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{Q_k^{tot}(t)\}\leq Q_k^{avg}, \qquad \forall k;$$

$$(b)\quad \lim_{T\to\infty}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\{u(Q_k^{tot}(t)-\delta_kQ_k^{avg})\}\leq\epsilon_k, \qquad \forall k;$$

$$(c)\quad m_k^u(t)\in\mathcal{MCS}_k^u, \qquad m_k^d(t)\in\mathcal{MCS}_k^d, \qquad \forall k,t;$$

$$(d)\quad p_k^{tx,min}(m_k^u,t)I_k(t)\leq p_k^{tx}(t)\leq p_k^{tx,max}I_k(t), \qquad p_k^{d,min}(m_k^d,t)I_a(t)\leq p_k^d(t)\leq\frac{p^{d,max}I_a(t)}{K}, \qquad \forall k,t;$$

$$(e)\quad I_k(t)\in\{0,1\}, \qquad \forall k,t, \qquad I_a(t)\in\{0,1\}, \qquad \forall t;$$

$$(f)\quad f_c(t)\in\mathcal{F}, \qquad I_m(t)=\mathcal{I}\{f_c(t)\}, \qquad \forall t;$$

$$(g)\quad f_k(t)\geq 0, \qquad \forall k,t, \qquad \sum_{k=1}^{K}f_k(t)\leq f_c(t), \qquad \forall t;$$

where $\Psi(t)=[\{\Psi_k(t)\}_{k=1}^K, f_c(t), I_a(t), I_m(t)]$, with $\Psi_k(t)=[m_k^u(t), m_k^d(t), p_k^{tx}(t), p_k^d(t), f_k(t), I_k(t)]$ and $\mathcal{I}\{\cdot\}$ is the indicator function. The constraints have the following meaning: (a) the average end-to-end delay of each user does not exceed the required value $D_k^{avg}=Q_k^{avg}/\overline{A_k}$; (b) the out-of-service probability is lower than a threshold ϵ_k; (c) the uplink and downlink MCS belong, respectively, to $\mathcal{MCS}_k^u$ and $\mathcal{MCS}_k^d$; (d) the uplink and downlink transmit powers guarantee the PER constraints and are lower than some fixed budget; in uplink, UE k's transmit power budget is denoted by $p_k^{tx,max}$; in downlink the AP's transmit power budget is denoted by $p^{d,max}$ and we suppose for simplicity that the maximum power allocable for each UE is $p^{d,max}/K$; (e) the indicator variables of each UE's and AP's sleep state are binary; (f) the computation frequency of the ES is selected from $\mathcal{F}$ and the indicator variable of the ES is 1 if $f_c(t)>0$; (g) the CPU cycle frequency assigned to UE k is non-negative and the sum of all CPU cycle frequencies assigned to all UEs does not exceed the ES's total computation

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

frequency $f_c(t)$. Since we do not assume any knowledge on the statistics of $A_k(t)$ (data unit arrivals) and on the radio channels, and due to the mixed-integer nature of its variables, this optimization problem is very challenging. Nonetheless, in [MdD+20], we solve it through a novel low-complexity algorithm, which hinges on Lyapunov optimization tools that provide theoretical optimality guarantees. For the sake of conciseness, we do not report here the details of our optimal resource allocation scheme. Instead, we refer the reader to [MdD+20]. Notice that the mathematics and the optimization algorithms detailed in [MdD+20] apply to many kinds of device. The only architectural device-specific assumption is the possibility to put their radio equipment into sleep mode.

3.4.4. Numerical results

We present here some simulation results to assess the performance of our online optimization strategy. All simulations are performed in Matlab with the following fixed settings. We consider a picocell placed at the center of a square area of side 150 m. We assume an FDD system, with total available bandwidth $B = 10$ MHz equally split between uplink and downlink. In the case of such a picocell, the AP active power is $p_a^{on} = 2.2$ W, while the power consumption in sleep mode $p_a^s = 278$ mW, and the minimum sleep time is 1 ms. The maximum transmit power of the AP is set to 251 mW, so that the maximum transmit power of each user is $251/K$ mW. For the channel power gain, we use the pathloss and shadowing model Alpha-Beta-Gamma for Urban Micro scenario as in [Sun+16], with a carrier frequency of 28 GHz, adding a Rayleigh fading with unit variance. The noise power spectral density is $N_0 = -174$ dBm/Hz, with an additional noise figure of 5 dB both at UEs and at the AP. The UEs' active and sleep power consumptions are $p_k^{on} = 0.9$ W and $p_k^s = 346$ mW, respectively [LNSM14], while the transmit power is computed as in [LNSM14], with a maximum of $p_k^{tx,max} = 100$ mW per UE. For the MCS, we can choose all M-QAM modulations with $M \in \{4, 16, 64, 256\}$, coupled with coding rates in $\{0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$, both in uplink and in downlink, so that $\mathcal{MCS}_k^u$ and $\mathcal{MCS}_k^d$ have 28 elements. The packet length for wireless transmission is 1500 bytes. The ES has a maximum CPU cycle frequency $f_{max} = 4.5 \times 10^9$ CPU cycle/s and an effective switched processor capacitance $\kappa = 10^{-27}$ Ws³/(CPU cycles)³. The vector of all possible CPU cycle frequencies is $\phi = [0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1] \times f_{max}$. The power consumption in active state is $p_m^{on} = 20$ W, whereas the sleep state power consumption $p_m^s = 10$ W. We consider a total time slot duration $\tau_l = 10$ ms, with $\tau_s = 1$ ms the portion of the slot used for control signaling and optimization, i.e. where all entities are in active state. Then, the slot duration for data transmission and computation is $\tau = 9$ ms.

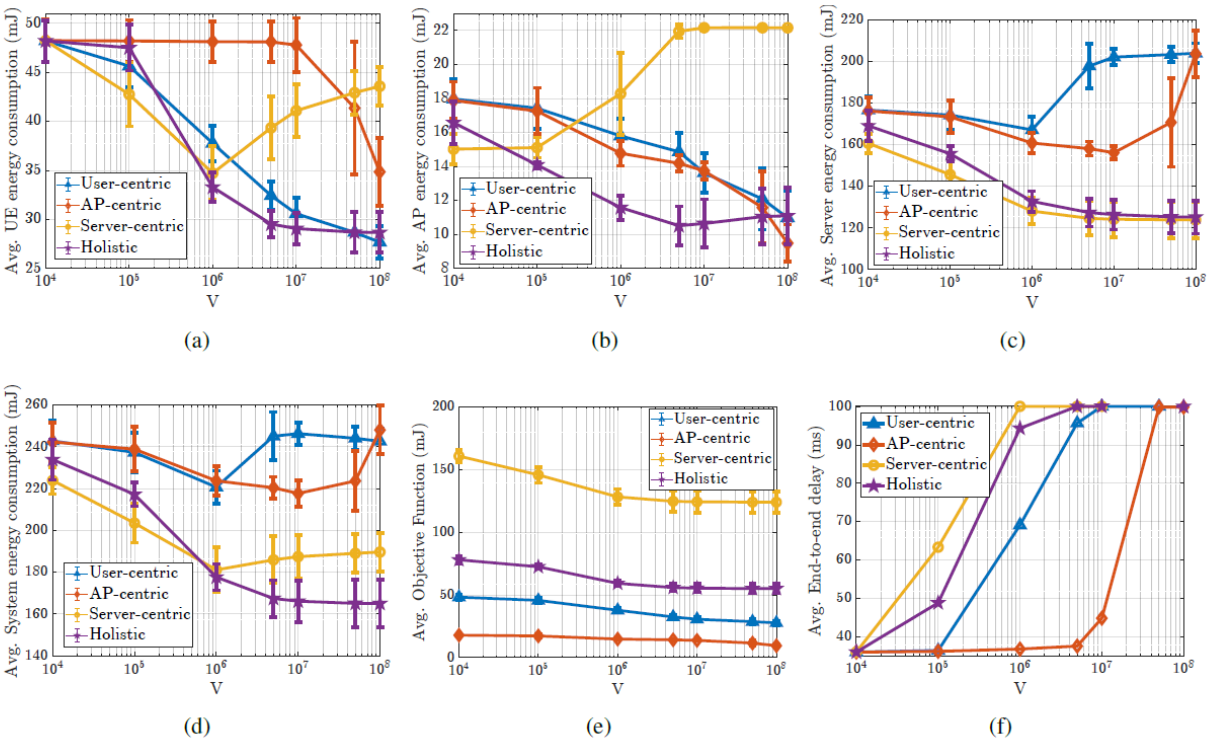

Figure 3.4-2 - Different metrics vs. the Lyapunov parameter V, for different weighting strategies. (a): Average UE energy consumption; (b): Average AP energy consumption; (c): Average ES energy consumption; (d): Average system energy consumption (sum of all energy consumptions); (e): Average objective function of the optimization problem; (f): Worst case average delay.

As a first numerical result, we illustrate the performance of D-MEC in terms of energy-delay trade-off. In particular, in Figure 3.4-2, we show the long-term average energy consumption of all users, the AP, the ES, the overall energy consumption (the sum of the three), the objective function of the optimization problem, and the end-to-end delay, all as a function of the Lyapunov tradeoff parameter V. For the formal definition of V, please see [MdD+20]. For the sake of this document, it is enough to say that the D-MEC resource scheduler achieves optimality when V tends to infinity. Therefore, the most meaningful values shown in Figure 3.4-2 are those obtained for $V = 10^8$. The curves show how the solution converges to the optimal one when we let V grow in our numerical simulations. In Figure 3.4-2, we plot all these quantities for different settings of the weighting parameters α_i, for $i = 1, 2, 3$. We run our simulations with random configurations of the following parameters: the input and output data size $S_k^i = 10^x, S_k^o = 10^y$ bits, with x and y uniformly randomly generated (u.r.g.) in $[2, 3]$ and [1, 3], respectively; we assume Poisson arrivals with A_k^{avg} u.r.g. in $[5, 15]$ data units; finally, $J_k = 10^{-z}$ data/CPU cycle, with z u.r.g. in $[2, 5]$. The simulation has run for $T = 10^5$ slots and it has been repeated over 100 independent realizations of the above random parameters and of $K = 5$ users' positions, uniformly distributed in a square of side 150 m. All UEs have an average delay requirement $D_k^{avg} = 100$ ms and, δ_k is fixed for all k, with $\delta_k = [1.5, 1.6, 1.7, 1.8, 1.9]$. The out-of-service constraint is $\epsilon_k = 10^{-2}$. We assume the bandwidth to be equally shared among all UEs and a target PER of 10^{-4}, both in uplink and downlink. All results are plotted with the standard deviation, except for the delay, whose value is the maximum among all simulation runs, i.e. the worst case.

The results have to be compared for high V, when the objective function is minimized (see Figure 3.4-2(e)). We show the results in four different settings with respect to the α_i:

- User-centric setting (blue curves). In this first setting, obtained with $\alpha_1 = 1, \alpha_2 = \alpha_3 = 0$, the energy consumption of the UE (Figure 3.4-2(a)) reaches its lowest level, while the energy consumption of the ES (Figure 3.4-2(c)) is not optimized. Instead, the AP's energy consumption (Figure 3.4-2(b)) reaches a level very close to its lowest, obtained with the AP-centric setting. This is due to the fact that the AP tends to operate in sleep mode when no UE transmits or requests results back, which happens often, due to the user-centric setting.
- AP-centric setting (orange curves. In this case, obtained with $\alpha_2 = 1, \alpha_1 = \alpha_3 = 0$, the energy consumption of both the AP and the UE approach very low values, for similar reasons as the previous case. This suggests that there exists a strong link between the two energy consumptions, since they must be active at the same time when they need to communicate. We can interpret the user-centric and the AP-centric solutions as "radio-centric".
- Server-centric setting (yellow curves). This solution, obtained with $\alpha_3 = 1, \alpha_1 = \alpha_2 = 0$, yields the lowest possible energy consumption for the ES as expected, but it is detrimental for the radio part, incurring additional energy consumption for the AP and the users.
- Holistic solution (violet curves). This solution aims at minimizing the overall system energy consumption, and is obtained with $\alpha_1 = \alpha_2 = \alpha_3 = 1/3$. This is the most interesting and promising strategy, since it is globally "green" and it reaches very close-to-optimal energy consumptions for each agent (UEs, AP, ES). This suggests that the three sources of energy consumption can be minimized jointly without detrimental effects on the single agents. Practically, the choice of the α_i's is based on the particular needs of the telecom operator, the MEC operator, or the UEs, but could be also based on a global and holistic energy reduction policy. In this paper, we do not tackle the problem of optimizing the α_i for the different needs and leave it for future investigation. Finally, notice that all solutions meet the constraint on the average service delay (Figure 3.4-2(f)), which increases with V until reaching the maximum allowed value of 100 ms. At the same time, the value of the objective function of the optimization problem decreases as V increases for all the policies (Figure 3.4-2(e)).

3.4.5. Conclusions

In this study, we proposed D-MEC, a dynamic resource allocation algorithm for computation offloading that jointly exploits low-power sleep modes of UEs, AP, and ES to reduce the system energy consumption with guaranteed end-to-end average delay and reliability. Via stochastic Lyapunov optimization, we formulated a long-term average optimization problem and solved it using a dynamic algorithm that works on a per-slot basis, without assuming any prior knowledge on the statistics of data arrivals and radio channels. The proposed algorithmic framework is guaranteed to reach a solution as close as wanted to the optimal one. Several numerical results illustrate the performance of our online strategy and how a holistic view of the system can be beneficial for all agents and for the global energy consumption. Future investigations will include optimized scheduling of spectral and time resources, and multi-cell multi-server scenarios, where cooperation among different APs and ESs can help reducing the overall energy consumption.

3.5. End-to-end time synchronization in TSN-5G network architectures

3.5.1. Ethernet time-sensitive network

Ethernet TSN is expected to fulfil various stringent requirements of future real-time networks, especially in Industry 4.0. Being actively developed, Ethernet TSN is standardized and adopted by a number of organizations and associations. IEEE 802.1AS [IEEE17] (hereafter referred to as simply AS) is an important standard, which constitutes one of four main concepts in TSN (the others are guaranteed latency, reliability and resource management [GAS+17], [SMB19]). This standard specifies the clock synchronization/time synchronization for Ethernet-based local area network (LAN), aiming to achieve the highly precise synchronization that is needed in the processes such as timestamped sharing and data fusion. Based on the synchronization protocol known as precision time protocol (PTP), IEEE 802.1AS specifies a new protocol named gPTP. The main procedure of this protocol is showed in Figure 3.5-1. The master node and slave node exchange PTP messages, then the slave node is able to estimate the offset between its clock and the master's clock.

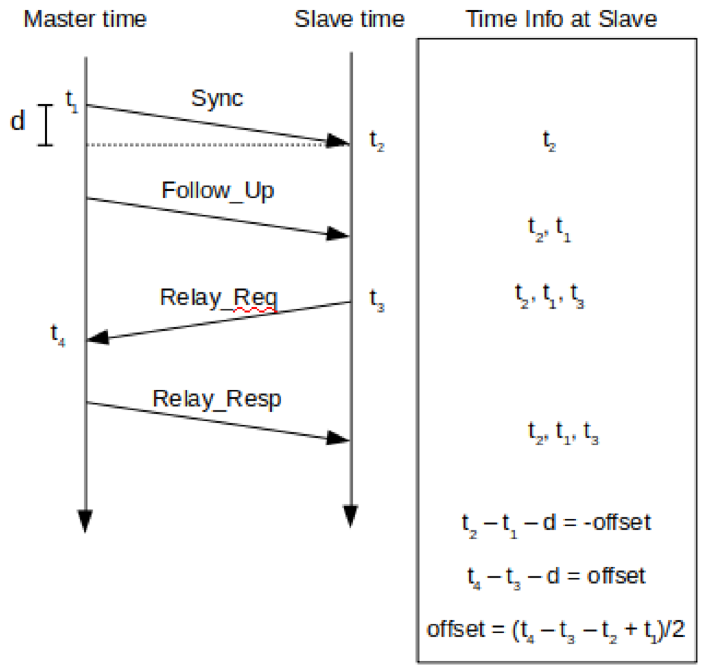

Figure 3.5-1 - PTP message exchange and estimation of time offset between master and slave.

3.5.2. Motivation for TSN-5G interworking

Integration between TSN and 5G is a demanding requirement in future communication systems, especially in the context of Industry 4.0 and CPSs. On one hand, as showed in [FVMS19], four main components in TSN matches the components in 5G URLLC, which are also time synchronization, guaranteed latency, reliability and resource management. On the other hand, in industry 4.0, there is a need for interconnecting wireline Ethernet TSN with other wireless domains. For example, in an industrial CPS, the wireline floor network of devices (e.g., sensors, robots) need to connect to manufacturing execution system (e.g., programmable logic controllers - PLCs) or corporate IT network that are installed at remote site or on cloud (Figure 3.5-2). The connection to this remote site is envisioned to be provided by wireless technologies of 5G and beyond. As a result, integrating TSN into 5G is an important task in 3GPP Release-16 [3GPP19b]. Moreover, another support of TSN in 5G is the use of TSN for 5G fronthaul traffic, which is standardized in IEEE 802.1CM. In this standard, some TSN profiles are defined for fronthaul traffic that requires deterministic transmission.

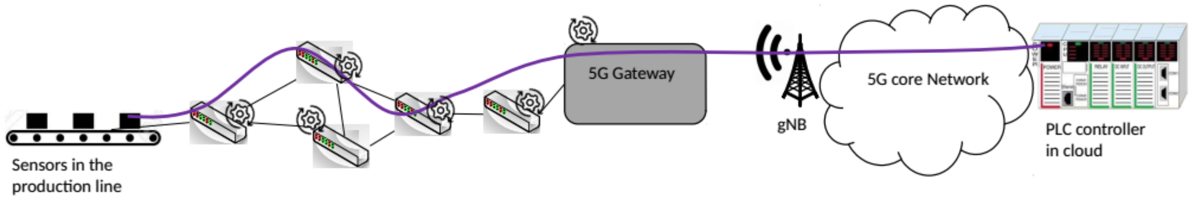

Figure 3.5-2 - Example of industrial CPS.

3.5.3. Time synchronization in TSN-5G networks

In the context of this research on communication system for future CPSs, we have implemented a testbed of Ethernet TSN and evaluated its performance on time synchronization [TBB20]. The evaluation results show that the system provides highly precise synchronization, on the scale of hundred nanoseconds. Based on this result, we propose to develop further the testbed and implement the interworking between wireline TSN and wireless systems. Details on this work will be reported in CPS4EU's deliverable D2.5 [D2.5].

In 3GPP Rel-16, a 5G system (5GS) appears from the rest of the network as a set of virtual TSN bridges. In other words, the Ethernet TSN views the 5G wireless domain as a single virtual gPTP time-aware system. At the edges of this 5GS, we need to implement device-side TSN translator (DS-TT) and network-side TSN translator (NW-TT), in order to translate and pass the gPTP message though the system (Figure 3.5-3). Another important function that these translators need is the Ethernet ports management, e.g., port capabilities.

On receiving *Sync* or *Delay Request* message, the NW-TT adds ingress timestamp into them. When receive these messages, the DS-TT gets the egress timestamp, uses this timestamp to estimate the residence time. Then, the DS-TT modifies the payload by removing the ingress timestamp and adding the residence time to the Correction Field (CF) of the synchronization packets. The message payload is carried within *Sync* message for one-step operation or *Follow up* message for two-step operation. These are two modes of operation in PTP, the latter is showed in Fig.1. Note that the residence time should to be lower than the upper bound required for a time-aware system specified in IEEE 802.1AS.

As showed in [TBB20], the configuration of these translators and the Ethernet devices can be performed based on the TSN standard IEEE 802.1Qcc, which is designed for TSN configuration. The Centralized Network Configuration (CNC) component in 802.1Qcc gets TSN bridge information via the TSN Application Function (AF) (Figure 3.5-3).

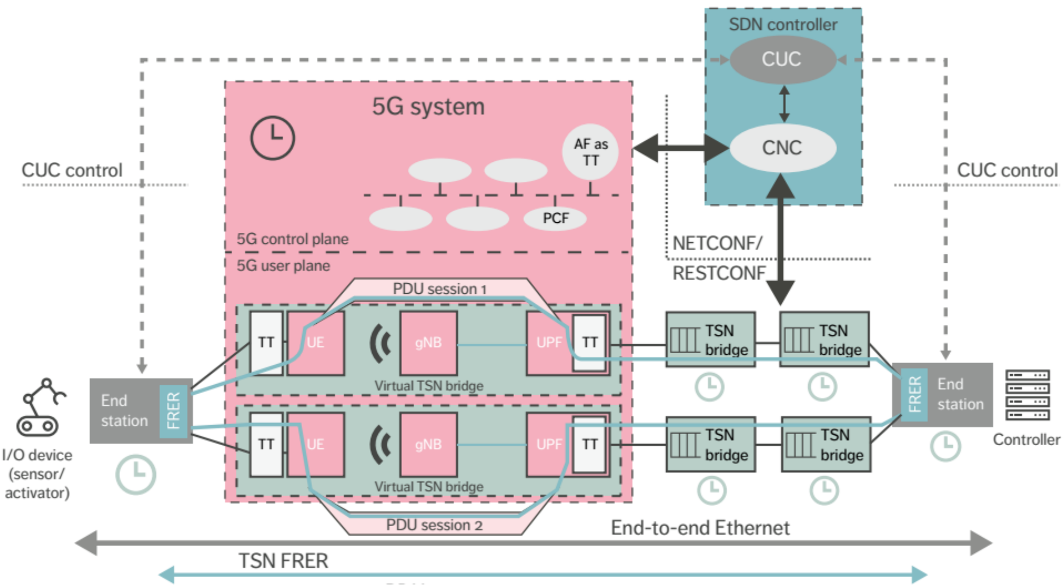

Figure 3.5-3 - Ethernet TSN interworking with 5GS [FVMS19].

The translators allow to compute residence time of gPTP message in the 5GS. As the latency of wireless links is expected to be more asynchronous and uncertain than wireline links, taking into account this residence time allows to estimate the synchronization offset more accurately, as showed in Figure 3.5-4. In the figure, if residence times r_{dow} and r_{up} have significant difference, then (eq. 3) is no longer accurate.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

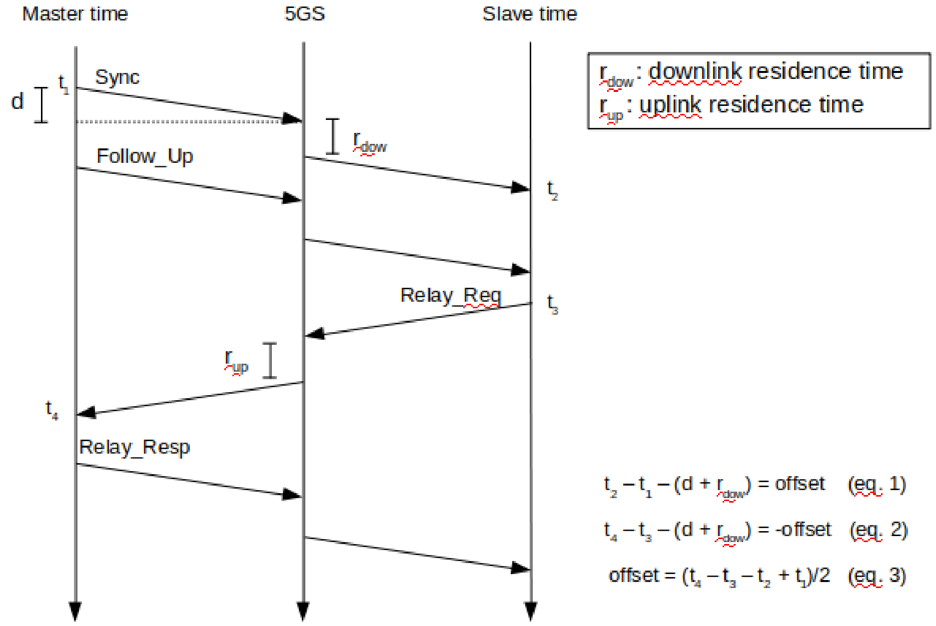

Figure 3.5-4 - Illustration of the importance of residence time in estimating time offset.

3.5.4. End-to-end time synchronization

After implemented the translators, another important task is to manage and configure the AS process of the system. A number of studies have shown the complexity of configuring TSN [SKJ18], [DH16] and AS [BTB19], especially in term of flexibility and reconfigurability. Configuring AS for end-stations/end-nodes and bridges/switches requires working with a high number of parameters. The configuring also has to reflect various network information such as network topology and traffic pattern. In addition, a new revision of IEEE 802.1AS has been recently approved [IEEE19], addressing more industrial use cases. This revision contains new features and new parameters, making the AS configuring more sophisticated. For instance, the new revision supports the configurations that provide additional levels of grand master (GM) and clock path redundancy. One of the solutions for configuring TSN is to employ an SDN-based paradigm. Several studies have presented the relevance of SDN in configuring real-time networks [KRV+15], automotive TSN [TE16], [HMKS19], and industrial TSN [MPA19]. Using an SDN-based paradigm, we implement SDN sound-bound APIs for each bridging devices, then deploy SDN controllers to communicate with these application programming interfaces (APIs) [TBB20]. The synchronization services will run on top of the controllers and set configuration to the bridging devices through the controllers, complying with the TSN standard IEEE 802.1Qcc. In this way, the end nodes (i.e., sensors and PLC controller as in Figure 3.5-5) can synchronize with each other regardless of the intermediate systems.

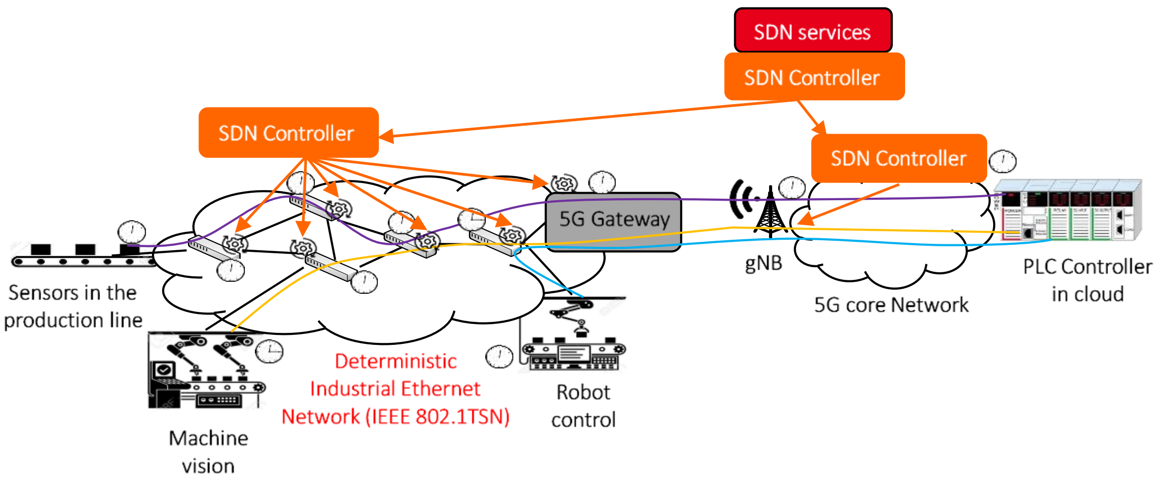

Figure 3.5-5 - End-to-end 5G-TSN synchronization.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

3.5.5. Conclusion

TSN is expected to play an important role in future industrial system, including networks for CPSs and smart factories. Inside TSN, IEEE 802.1AS is a standard that allows to synchronize the time of network nodes. In the context of the research for future CPSs, we implemented 802.1AS on the devices of an Ethernet network, then evaluated the accuracy of their time synchronization. More information about this research can be found in our previous work [TBB20]. Beyond setting up and evaluating TSN on a wireline Ethernet network, we are currently studying and developing the interworking between this wireline TSN and the wireless systems, e.g., 5GS. Surveying through technologies and standards, we are adopting the idea of TSN translators in 3GPP Release 16. TSN translators need to be implemented at the edges of 5GS in order to estimate the residence time of PTP messages inside this 5GS. Given that these translators and the Ethernet devices are configured for 802.1AS, the time synchronization can operate across the wireline and wireless domains.

3.6. Channel estimation for non-static users via base station cooperation

3.6.1. Motivation and related work

Wireless connectivity is making its way into industrial networks. It eases the maintenance and control of devices and enables the coordination of moving devices such as transport vehicles. However, wireless communications are unreliable compared to wired links and poor connectivity can result in delayed transmission or reduced performance which is critical for industrial applications. One of the main challenges for future Industrial IoT is thus how to ensure URLLC wirelessly [KJG+19].

Sub-GHz bands are not suitable to support a dense traffic and their reliability is significantly reduced because of the presence of interfering systems. MmWave spectrum is thus considered as a plausible candidate to replace wires in industrial sectors especially with the licensed band at 28 GHz and the unlicensed band at 60 GHz.

In industrial environments, RF blockages may occur because of moving metallic objects and vehicles as already introduced in Section 3.3. Spatial diversity is therefore appealing to prevent those RF blockages [KJG+19]. In this context, Coordinated Multi-Point (CoMP) is a promising approach. In CoMP networks, user terminals are not connected with a single Transmit and Receive Point (TRP) but simultaneously with several TRPs, which cooperate with each other. The multi-TRP cooperation allows either scheduling coordination (no data is exchanged between the TRPs) or joint-signal processing. It is indeed possible to simultaneously transmit the downlink signal to user terminals from different TRPs (*i.e.* joint-transmission) and evaluating the uplink signal with several TRPs (*i.e.* joint-detection). By doing so, the joint signal processing exploits the inter-cell and inter-user interferences instead of treating them as noise.

With joint-transmission, users receive multiple downlink signals coming from distinct TRPs in the same band to improve the receiver performance. In real systems, channel estimation and synchronization impairments limit the performance of joint-transmission techniques [Man+15], [Jun+14], [MJJB13]. First, channel estimation is imperfect because of the presence of noise and interference. Then, the estimates are quantized in FDD systems and channel reciprocity is generally wrongly assumed in Time-Division Duplex (TDD). In addition to that, there is always a delay between the channel estimation and the computation of the precoding and therefore the estimates can be outdated. This effect is known as channel aging [Man+15]. The effects of channel aging are even worsened with user mobility.

When it comes to TRP synchronization, they typically use precise crystal oscillators and therefore CFO is highly limited [Man+15]. However, the oscillator frequencies of all TRPs must be locked to a common reference. The clock error between two terminals is set to 1 µs by 3GPP for Industrial Internet of Things [3GPP19a]. A common clock signal can be obtained with a network synchronization protocol like Time Sensitive Network (TSN). And when the synchronization provided by the network protocol is not sufficient, extra over-the-air synchronization procedures can be considered [Rog+14]. However, residuals CFOs still remain and it is practically inevitable.

Users thus received multiple downlink signals each with their own frequency offset induced by both the imprecision of the oscillators and by the possible user mobility. A pragmatic solution is hence to estimate those multiple frequency offsets to compensate them at the receiver side. Maximum-likelihood methods are proposed such [ZC08] based on Newton search or [THCY13] based on the design of robust orthogonal training sequences. However, the multi-CFO estimation remain complicated.

3.6.2. Proposed work

In this work, we propose to move the multi frequency offset estimation at the TRP side. Indeed, in industrial scenarios, some devices may have highly limited computation capacity, which hinders the possibility to estimate the frequency offset of the cooperative downlink signals.

The proposed study focuses on the compensation of the frequency offsets at the TRPs side. The purpose is to align the multiple downlink signals to a common frequency in order not to break the precoding and ease the detection and demodulation processes at the device side. Zarikoff and Cavers propose in [ZC10] a method to jointly estimate the CFO of the TRPs and the devices from the frequency offset for each link. They show that the spread of frequency offsets is significantly reduced for realistic systems. However, Doppler shifts are not considered in the aforementioned study. Without taking into consideration the Doppler effects, the frequency shift compensation is incomplete resulting in the accumulation of frequency offsets [ZHYZ19]. In mobility scenarios, a joint estimation of the CFOs (caused by oscillator imperfections) and Doppler effect (induced by device mobility) is thus required. We therefore propose in this work a frequency offset compensation technique for cooperative joint-transmission taking into account Doppler shifts induced by device mobility.

When only one path exists, the Doppler shift and the CFO sum up and it is impossible to separate them. That is why in most studies dealing with Doppler estimation a multipath channel model is considered [ZHYZ19], [ZSF+20]. However, in industrial environment, this assumption is not valid anymore because of the many obstacles. The spatial diversity granted by a distributed network of cooperative TRPs is thus appealing to ensure different paths in our case.

To estimate a Doppler shift, it is important to know the angle of arrival of the signal and the angle of motion of the users. Generally, the angles of arrival are estimated thanks to antenna arrays which allow an angle-domain projection of the received signals. When it comes to the angles of motion, we have an a priori knowledge in high-speed train scenarios because to movement of the trains is limited by the railway [ZHYZ19]. However, in industrial scenarios, the movement of the users are not predictable and therefore the angle of motion must be estimated as well.

In this work, we therefore propose a joint-estimation of the CFO of the users and the TRPs, the maximum Doppler shift and the angle of motion for each user. The proposed technique allows a fine frequency offset compensation at the TRPs ensuring high-capacity downlink links in environments where typically devices’ mobility constitutes an impairment. A patent application is being prepared on this topic and therefore the proposed technique will be fully described and the results presented in CPS4EU’s deliverable D2.4 [D2.4].

4. CONCLUSION

CPSs will play a more and more prominent role in many strategic industrial sectors and European SMEs as well as big groups will benefit from the exploitation of CPS technologies from various points of view: CPSs will contribute to improve companies' productivity, diminish their operational costs, and support the optimization of their logistics. By their very own nature of highly interconnected systems both among themselves and with the surrounding environment, CPSs need novel dedicated communication solutions, suitable for a completely new variety of applications and requirements. These solutions may partially come from a straightforward adaptation or application of already standardized technologies, but in non-negligible part require an innovation effort that goes beyond what 4G and 5G mobile networks currently offer.

In line with the vision of "beyond-5G" networks that standardization bodies are starting to develop, in this report we described the research studies carried out so far by Task 2.2 of CPS4EU. This work is grounded on the motivation and use case requirements highlighted by Task 2.1 and concerns both CPSs' communication modules and their entire supporting network infrastructure. Our opinion is that an efficient integration of CPS technologies into the European industrial scenario must rely on fast and reliable communications. For this reason, our contributions specifically focused on URLLC and TSN.

Our main conclusion is that the conception of CPS-adapted communication networks and technologies will be effective if it happens in the most holistic and inter-layer manner possible: it will have to combine new networking solutions for xHauling with physical-layer innovation and optimized resource allocation and management schemes for all the involved network elements. The system or network models, the algorithms, and the numerical results described and summarized in this report strengthen this vision and provide promising tools for the future development of CPS technologies. We will provide in D2.5 [D2.5] and D2.6 [D2.6] further implementation and evaluation results of a selection of the techniques investigated in these pages. Further results on the activities of Task 2.2 will instead appear in D2.4 [D2.4], the final version of this report due at the end of the project's lifetime.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

5. REFERENCES

[3GPP19a] 3GPP, “TR 38.325 - Technical specification group radio access network; Study on NR industrial Internet of Things (IoT)”, release 16, March 2019.

[3GPP19b] 3GPP, “TS 24.519 - Time-Sensitive Networking (TSN) Application Function (AF) to Device-Side TSN Translator (DS-TT) and Network-Side TSN Translator (NW-TT) protocol aspects”, release 16, 2019.

[Aue+11] G. Auer *et al.*, “How much energy is needed to run a wireless network?” *IEEE Wireless Comm.*, vol. 18, no. 5, pp. 40–49, Oct. 2011.

[BDF20] S. Bicaïs, J.-B. Doré, A. Falempin, “Méthode de démodulation par apprentissage automatique pour récepteurs MIMO à détection d’énergie, ” patent proposal n. FR2005670, submitted on May 29th, 2020.

[BFD+20] S. Bicaïs, A. Falempin, J.-B. Doré, V. Savin, “Design and analysis of MIMO systems using energy detectors for sub-THz applications,” submitted for publication to *IEEE Trans. Wireless Commun.*, 2020.

[BTB19] S. Ben Hadj Said, Q. H. Truong, M. Boc, “SDN-based configuration solution for IEEE 802.1 time sensitive networking (TSN),” *ACM SIGBED Rev.*, vol. 16, no. 1, pp. 27-32, 2019.

[BYW+15] N. Brahmi, O. N. C. Yilmaz, K. Wang Helmersson, S. A. Ashraf, J. Torsner, “Deployment strategies for ultra-reliable and low latency communication in factory automation,” in *Proc. IEEE GC Wkshps*, San Diego, CA, USA, 2015, pp. 1-6.

[Che16] M. Cheffena, “Industrial wireless communications over the millimeter wave spectrum: Opportunities and challenges,” *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 66-72, 2016.

[CM18] P. Chang, G. Miao, “Resource provision for energy-efficient mobile edge computing systems,” in *Proc. IEEE GLOBECOM*, Abu Dhabi, UAE, pp. 1-6, 2018.

[CSKM13] S. N. Chiu, D. Stoyan, W. S. Kendall, J. Mecke, *“Stochastic geometry and its applications”*. John Wiley & Sons, 2013.

[CZX17] L. Chen, S. Zhou, J. Xu, “Energy efficient mobile edge computing in dense cellular networks,” in *Proc. IEEE ICC*, Paris, France, pp. 1-6, 2017.

[D2.1] CPS4EU, Deliverable 2.1, “Specification and architecture of the communication modules”, Dec. 2019.

[D2.2] CPS4EU, Deliverable 2.1, “Specification and architecture of the communication modules – v2”, Nov. 2020.

[D2.4] CPS4EU Deliverable 2.4, “Propositions for 5G, including URLLC evolution – v2”, due date: May 2022.

[D2.5] CPS4EU, Deliverable 2.5, “Simulation tools and experimental platforms – v1”, Dec. 2020.

[D2.6] CPS4EU, Deliverable 2.6, “Simulation tools and experimental platforms – v2”, due date: June 2022.

[DDL15] B. Debaillie, C. Desset, F. Louagie, “A flexible and future-proof power model for cellular base stations,” in *Proc. IEEE VTC Spring*, Glasgow, Scotland, UK, pp. 1-7, 2015.

[DH16] J. L. Du, M. Herlich, “Software-defined networking for real-time Ethernet,” in *Proc. ICINCO*, Lisbon, Portugal, pp. 584-589, 2016.

[FVMS19] J. Farkas, B. Varga, G. Miklós, J. Sachs, “5G-TSN integration meets networkingr equirements”, *Ericsson Technol. Rev.,* pp. 1-11, 2019.

[GAS+17] M. Gutierrez, A. Ademaj, W. Steiner, R. Dobrin, S. Punnekkat, “Self-configuration of IEEE 802.1 TSN networks,” in *Proc. IEEE ETFA*, Limassol, Cyprus 2017, pp. 1-8, 2017.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

[GKD21] G. Ghatak, S. R. Khosravirady, A. De Domenico, "Stochastic geometry framework for ultra-reliable cooperative communications with random blockages," submitted to *IEEE INFOCOM*, 2021.

[GSS15] C. Gough, I. Steiner, W. Saunders, "*Energy Efficient Servers: Blueprints for Data Center Optimization"*. Apress, 2015.

[HCF20] D. Han, W. Chen, Y. Fang, "Joint channel and queue aware scheduling for latency sensitive mobile edge computing with power constraints," *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3938-3951, 2020.

[HMKS19] T. Hackel, P. Meyer, F. Korf, T. C. Schmidt, "Software-defined networks supporting time-sensitive in-vehicular communication," in Proc. IEEE VTC-Spring, Kuala Lumpur, Malaysia, pp. 1-5, 2019.

[IEEE17] IEEE, "IEEE draft standard for local and metropolitan area networks – timing and synchronization for time-sensitive applications," IEEE P802.1AS-Rev/D6.0, pp. 1-496, December 2017.

[IEEE19] IEEE, "Timing and synchronization for time-sensitive applications," IEEE P802.1AS-Rev/D8.3, pp. 1-434, October 2019.

[JDC20] J.-L. Jimenez Gonzalez, C. Dehos, N. Cassiau, "Channel bonding transceivers for 6G future networks," in *Proc. 6G Summit*, 2020.

[JDPM16] L. Jing, E. De Carvalho, P. Popovski, A. O. Martinez, "Design and performance analysis of noncoherent detection systems with massive receiver arrays," *IEEE Trans. Signal Process.*, vol. 64, no. 19, pp. 5000-5010, 2016.

[JKV18] R. Jurdi, S. R. Khosravirad, H. Viswanathan, "Variable-rate ultra-reliable and low-latency communication for industrial automation," in *Proc. CISS*, Princeton, NJ, USA, pp. 1-6, 2018.

[JPY+18] H. Ji, S. Park, J. Yeo, Y. Kim, J. Lee, B. Shim, "Ultra-reliable and low-latency communications in 5G downlink: Physical layer aspects," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 124-130, 2018.

[Jun+14] V. Jungnickel *et al.*, "The role of small cells, coordinated multipoint, and massive MIMO in 5G", *IEEE Commun. Mag.*, vol. 52, n. 5, pp. 44-51, 2014.

[KB17] D.P. Kingma J. Ba, "Adam: A method for stochastic optimization", *arXiv preprint*, 2017. Available: https://arxiv.org/abs/1412.6980

[KJG+19] M. Khoshnevisan, V. Joseph, P. Gupta, F. Meshkati, R. Prakash, P. Tinnakornsrisuphap, "5G industrial networks with CoMP for URLLC and time sensitive network architecture," *IEEE J. Sel. Areas Commun.*, vol. 37, n. 4, pp. 947-959, 2019.

[KRV+15] D. Kreutz, F. Ramos, P. Verissimo, C. E. Rothenberg, S. Azodolmolky, S. Uhlig, "Software-defined networking: A comprehensive survey," IEEE Proc., vol. 103, no. 1, pp. 14-76, 2015.

[LBDP19] C. Liu, M. Bennis, M. Debbah, H. V. Poor, "Dynamic task offloading and resource allocation for ultra-reliable low-latency edge computing," *IEEE Trans. Commun.*, vol. 67, no. 6, pp. 4132-4150, 2019.

[LGJG19] L. Li, Q. Guan, L. Jin, M. Guo, "Resource allocation and task offloading for heterogeneous real-time tasks with uncertain duration time in a fog queueing system," *IEEE Access*, vol. 7, pp. 9912-9925, 2019.

[LNSM14] M. Lauridsen, L. Noël, T. B. Sorensen, P. Mogensen, "An empirical LTE smartphone power model with a view to energy efficiency evolution," *Intel Technol. J.*, vol. 18, pp. 172-193, 2014.

[LUSS18] Z. Li, M. A. Uusitalo, H. Shariatmadari, B. Singh, "5G URLLC: Design challenges and system concepts," in *Proc. ISWCS*, Lisbon, Portugal, 2018, pp. 1-6.

[Man+15] K. Manolakis *et al.*, "Cooperative cellular networks: Overcoming the effects of real-world impairments", *IEEE Veh. Technol. Mag.*, vol. 10, n. 3, pp. 30-40, 2015.

[Mat60] B. Matérn, "Spatial variation. Stochastic models and their application to some problems in forest surveys and other sampling investigations," *Meddelanden från statens skogsforskningsinstitut*, vol. 49, no. 5, 1960.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276

[MB17] P. Mach, Z. Becvar, "Mobile edge computing: A survey on architecture and computation offloading," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1628-1656, 2017.

[MCG20] F. F. Manzillo, A. Clemente, J. L. Gonzalez-Jimenez, "High-gain D-band transmitarrays in Standard PCB technology for beyond-5G communications," *IEEE Trans. Antennas Propag.*, vol. 68, no. 1, pp. 587-592, 2020.

[MDBF20] M. Merluzzi, P. Di Lorenzo, S. Barbarossa, V. Frascolla, "Dynamic computation offloading in multi-access edge computing via ultra-reliable and low-latency communications," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 6, pp. 342-356, 2020.

[MdD+19] M. Merluzzi, N. di Pietro, P. Di Lorenzo, E. Calvanese Strinati, S. Barbarossa, "Network energy efficient mobile edge computing with reliability guarantees," in *Proc. IEEE GLOBECOM*, Waikoloa, HI, USA, pp. 1-6, 2019.

[MdD+20] M. Merluzzi, N. di Pietro, P. Di Lorenzo, E. Calvanese Strinati, S. Barbarossa, "D-MEC: Discontinuous mobile edge computing," *arXiv preprint,* 2020. Available: https://arxiv.org/abs/2008.03508

[MEC] ETSI Industry Specification Group. Web page: https://www.etsi.org/technologies/multi-access-edge-computing

[MJJB13] K. Manolakis, S. Jaeckel, V. Jungnickel, V. Braun, "Channel prediction by doppler-delay analysis and benefits for base station cooperation", in *Proc. IEEE VTC Spring*, Dresden, Germany, pp. 1-6, 2013.

[MPA19] L. Moutinho, P. Pedreiras, L. Almeida, "A real-time software defined networking framework for next-generation industrial networks," *IEEE Access*, vol. 7, pp. 164468-164479, 2019.

[MHZ11] B. Müller, M. Holters, U. Zölzer, "Low complexity soft-input soft-output Hamming decoder," in *Proc. FITCE Congress*, Palermo, Italy, pp. 1-5, 2011.

[MZSL17] Y. Mao, J. Zhang, S. H. Song, K. B. Letaief, "Stochastic joint radio and computational resource management for multi-user mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 9, pp. 5994-6009, Sep. 2017.

[Nan+17] Y. Nan *et al.*, "Adaptive energy-aware computation offloading for cloud of things systems," *IEEE Access*, vol. 5, pp. 23947-23957, 2017.

[PAU14] S. Paquelet, L. M. Aubert, B. Uguen, "An impulse radio asynchronous transceiver for high data rates," in *Proc. Joint UWBST & IWUWBS*, Kyoto, Japan, pp. 1-5, 2004.

[PKT+12] J. Park, S. Kang, S. V. Thyagarajan, E. Alon, A. M. Niknejad, "A 260 GHz fully integrated CMOS transceiver for wireless chip-to-chip communication," in *Proc. VLSIC*, Honolulu, HI, USA, pp. 48–49, 2012.

[PTSD18] P. Popovski, K. F. Trillingsgaard, O. Simeone, G. Durisi, "5G wireless network slicing for eMBB, URLLC, and mMTC: A communication-theoretic view," *IEEE Access*, vol. 6, pp. 55 765-55 779, 2018.

[PW09] G. K. Psaltopoulos, A. Wittneben, "Diversity and spatial multiplexing of MIMO amplitude detection receivers," in *Proc. IEEE PIMRC*, Tokyo, Japan, pp. 202–206, 2009.

[Rog+14] R. Rogalin *et al.*, "Scalable synchronization and reciprocity calibration for distributed MIMO", *IEEE Trans. Wireless Commun.*, vol. 13, n. 4, pp. 1815-1831, 2014.

[SKJ18] S. Schriegel, T. Kobzan, J. Jasperneite, "Investigation on a distributed SDN control plane architecture for heterogeneous time sensitive networks," in *Proc. IEEE WFCS*, Imperia, Italy, pp. 1-10, 2018.

[SMB19] T. Striffler, N. Michailow, M. Bahr, "Time-sensitive networking in 5th generation cellular networks - current state and open topics," in *Proc. IEEE 5GWF*, Dresden, Germany, pp. 547-552, 2019.

[SSB15] S. Sardellitti, G. Scutari, S. Barbarossa, "Joint optimization of radio and computational resources for multicell mobile edge computing," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 1, no. 2, pp. 89-103, Jun. 2015.

[Sun+16] S. Sun *et al.*, “Propagation path loss models for 5G urban micro- and macro-cellular scenarios,” in *Proc. IEEE VTC Spring*, Nanjing, China, pp. 1–6, 2016.

[Swa+15] V. N. Swamy *et al.*, “Cooperative communication for high-reliability low-latency wireless control,” in *Proc. IEEE ICC*, London, UK, 2015, pp. 4380-4386.

[TBB20] M. -T. Thi, S. Ben Hadj Said, M. Boc, "SDN-based management solution for time synchronization in TSN networks," in *Proc. IEEE ETFA*, Vienna, Austria, pp. 361-368, 2020.

[TE16] D. Thiele, R. Ernst, “Formal analysis based evaluation of software defined networking for time-sensitive ethernet,” in *Proc. DATE*, Dresden, Germany, pp. 31-36, 2016.

[THCY13] T.-H. Tsai, H.-Y. Huang, Y.-C. Chen, K.-J. Yang, “Simultaneous multiple carrier frequency offsets estimation for coordinated multi-point transmission in OFDM systems”, *IEEE Trans. Wireless Commun.*, vol. 12, n. 9, pp. 4558-4568, 2013.

[Tom+15] S. Tombaz *et al.*, “Energy performance of 5G-NX wireless access utilizing massive beamforming and an ultra-lean system design,” in *Proc. IEEE GLOBECOM*, San Diego, CA, USA, pp. 1-7, 2015.

[VPZ13] M. Voicu, D. Pepe, D. Zito, “Performance and trends in millimetre-wave CMOS oscillators for emerging wireless applications,” *International J. Microw. Sci. Technol.*, vol. 2013, pp. 1-6, 2013.

[WZYW19] S. Wang, X. Zhang, Z. Yan, W. Wang, “Cooperative edge computing with sleep control under nonuniform traffic in mobile edge networks,” *IEEE Internet Things J.*, vol. 6, no. 3, pp. 4295-4306, 2019.

[ZC08] B. W. Zarkoff, J. K. Cavers, “Multiple frequency offset estimation for the downlink of coordinated MIMO systems”, *IEEE J. Sel. Areas Commun.*, vol. 26, n. 6, pp. 901-912, 2008.

[ZC10] B. W. Zarikoff, J. K. Cavers, “Coordinated multi-cell systems: Carrier frequency offset estimation and correction”, *IEEE J. Sel. Areas Commun.*, vol. 28, n. 9, pp. 1490-1501, 2010.

[ZHYZ19] R. Zeng, H. Huang, L. Yang, Z. Zhang, “Joint estimation of frequency offset and Doppler shift in high mobility environments based on orthogonal angle domain subspace projection”, *IEEE Trans. Veh. Technol.*, vol. 67, n. 3, pp. 2254-2266, 2019.

[ZSF+20] Q. Zhang, H.Sun, Z. Feng, H.Gao, W. Li, “Data-aided Doppler frequency shift estimation and compensation for UAVs”, *IEEE Internet Things J.*, vol. 7, n. 1, pp. 400-415, 2020.

This project has received funding from the ECSEL Joint Undertaking (JU) under grant agreement No 826276