IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 22 September 2023, accepted 18 October 2023, date of publication 23 October 2023, date of current version 1 November 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3326846

SURVEY

Non-Orthogonal Multiple Access for Offloading in Multi-Access Edge Computing: A Survey

ROMAIN DULOUT[1,2], LEO MENDIBOURE[2], (Member, IEEE), YANNIS POUSSET[1], VIRGINIE DENIAU[3], AND FREDERIC LAUNAY[4], (Member, IEEE)

[1]XLIM Laboratory (UMR CNRS 7252), University of Poitiers, 86000 Poitiers, France
[2]COSYS-ERENA Laboratory, Université Gustave Eiffel, 33067 Pessac, France
[3]COSYS-LEOST Laboratory, Université Gustave Eiffel, 59650 Villeneuve-d'Ascq, France
[4]LIAS Laboratory (EA 6315), University of Poitiers, 86000 Poitiers, France

Corresponding author: Romain Dulout (romain.dulout@univ-eiffel.fr)

ABSTRACT Multi-Access Edge Computing (MEC) architectures now seem to represent the future of data processing architectures. Indeed, they have the potential to optimize the use of the backhaul network, guarantee the implementation of real-time applications and offer services adapted to the user's context. To further improve the performance of MEC architectures, it may be worth combining them with a new technology at the physical layer: Non Orthogonal Multiple Access (NOMA). During the task offloading process, this could enhance energy efficiency, maximize the number of users benefiting from MEC services and further reduce latency. That is why the article focuses on the use of NOMA for task offloading in MEC architectures and offers a comprehensive study on the subject. First, we define a taxonomy to ensure a systematic review of existing work. We then analyze and compare existing works, classifying them according to their purpose. Based on this, we then discuss the benefits and limitations of existing work, highlighting some good practices. Finally, we identify future research directions in the field of NOMA-assisted MEC architectures.

INDEX TERMS Massive connectivity, multi-access edge computing, NOMA, offloading, quality of service.

I. INTRODUCTION

The growth in computing capabilities, the introduction of new artificial intelligence based (AI) algorithms and more efficient communication systems have led to the emergence of new applications and services for the future [1]. These include: 1) Augmented/Virtual/Mixed Reality, which creates immersive environments allowing users to interact with objects and information in a more intuitive way; 2) Cooperative Intelligent Transport Systems (C-ITS), which aims to increase road safety and traffic management; 3) the multiplication of Internet of Things devices (IoTs) that can supervise and control entire cities or new industries by providing massive data to AI programs; 4) 8K streaming, which guarantees a fluid user experience for content viewing and 5) Telemedicine, which enables remote medical care for distant populations.

The associate editor coordinating the review of this manuscript and approving it for publication was Christos Anagnostopoulos.

However, the operation of these new applications and services requires high performance, both in terms of the computing resources needed to run them, and in terms of the network infrastructure that transports and processes the data. Reduced processing times to operate in real time, as well as limited power consumption to improve the lifetime of the devices are particularly mandatory. In recent years, the use of Mobile Cloud Computing (MCC) architectures has proven to be relevant to address these concerns [8]. This type of distributed architecture combining remote computing and storage capabilities in the cloud aims at improving the performance and capabilities of mobile network users, providing them with cloud services to run their new performance-intensive applications. However, in the MCC architecture the physical resources are located far from the users, increasing latency and processing delays [9]. These are important limitations to the optimal operation of these new applications. In addition, the higher power consumption required to transport data to the cloud is disadvantageous for

TABLE 1. Comparison of existing surveys.

Survey	NOMA oriented	MEC oriented	NOMA-MEC side	Number of papers reviewed
[2]	Yes	No	No	Less than 15
[3]	Yes	No	No	Less than 15
[4]	No	Yes	No	Less than 15
[5]	No	Yes	No	Less than 15
[6]	No	Yes	No	Less than 15
[7]	Yes	No	No	Less than15
Our survey	No	No	Yes	57

users with limited resources, such as IoTs, since they are running on batteries.

To overcome these limitations, a paradigm shift has occurred with the introduction of Multi-Access Edge Computing (MEC) architectures [10]. MEC consists in deploying computing and storage servers at the edge of the network, as close as possible to the end users. This type of deployment close to end users enables them to benefit from additional computing resources in a quasi-transparent manner. This ensures the functioning of new applications and services while limiting processing times and energy consumption compared to MCC architectures [11]. As a result, mobile users have the option of running some or all of their applications on the MEC server, known as offloading process. However, the latency and energy consumption during offloading depend on several factors, such as the underlying communication technology in the MEC infrastructure, the propagation channel conditions, the mobility and number of users, and so on. Despite the predisposition of MEC architectures to provide low latency and low power consumption, such factors can drastically reduce its performance.

Alongside this, a significant amount of work has been done on the physical layer, with a growing interest in recent years in Non-Orthogonal Multiple Access communications (NOMA) [12]. These technologies allow multiple users to share the same radio resources in time and frequency, using power domain with superposition coding (SC) at the transmitter side, and Successive Interference Cancellation (SIC) at the receiver side to distinguish the signals of the different users. This enables users to access the channel simultaneously and improves spectral efficiency.

Due to its attractive characteristics, researchers began formulating new proposals around MEC architecture in 2018, where NOMA is used as the physical layer providing an optimal solution to overcome the MEC limitations. Their intuition proved to be right, as NOMA permits multiple users to offload their tasks simultaneously without having to wait for a radio resource to free up, reducing processing times. In addition, superposition coding, consisting in weighting the signals to be transmitted, reduces the energy consumed to offload tasks. Thus, the integration of NOMA in MEC architectures has been a major contribution in improving the performance of these infrastructures, becoming a new type of architecture called NOMA-assisted MEC networks. Since then, the research community has formulated many proposals, assessing these architectures in new use cases, defining new algorithms to efficiently manage joint resources, and introducing new solutions based on emerging technologies.

To the best of our knowledge, no state of the art on offloading in NOMA-assisted MEC architectures has been proposed to date. Table 1 compares the existing surveys that partially explore this topic. Most of them address offloading proposals in a very oriented way, i.e. from a NOMA point of view in which MEC is an annex technology, or from a MEC point of view in which NOMA is part of a set of solutions allowing to improve MEC architectures. Purely NOMA-assisted MEC papers are only presented within one sub-section of the studies carried out, and a limited number of papers were compared. In this survey, we propose to analyze specifically the approaches and proposals made by the authors in the NOMA-assisted MEC architectures, considering MEC and NOMA a monolithic block. Our contributions can be summarized as follows:

- We define a taxonomy around the key concepts of NOMA-assited MEC architectures by defining a basic conceptual framework that allows us to understand all the papers in the literature as well as future proposals. The defined structure also contributes to the harmonization of terminologies and future works around these architectures, facilitating future collaborations, mutual understanding and sharing of research results;
- Considering the combination of NOMA and MEC as an indivisible block, we have conducted a comparative study and classified a large number of proposals structured around the main objectives of NOMA-MEC architectures;
- On the basis of this comparative overview, a discussion is conducted on the approaches and trends that emerge, identifying in particular the most effective methods for dealing with a particular problem, the limitations of certain models and approaches, and best practices;
- Finally, we have identified four perspectives on this research area that will allow us to overcome the bottleneck induced by a number approaches, in order to evaluate and propose new solutions in these architectures.

The rest of this paper is organized as follows. First, a brief introduction to MEC and NOMA communication architectures is proposed to understand the functioning of the global NOMA-assisted MEC architecture (Section II).

Thereafter, a taxomony is described to facilitate mutual understanding (Section III). On this basis, the proposals made in the literature are compared (Section IV). Then, a discussion is led to identify the more relevant approaches and the limitations of the proposals (Section V). Finally, we in proposals and perspectives around this field of research (Section VI).

II. BACKGROUND

This section defines a number of concepts that are essential to understanding the rest of the document. Unlike existing surveys, such as [13] and [14], NOMA-MEC architectures are analyzed as a complete and indivisible entity. Consequently, this section will not constitute a new knowledge base specific to each of these technologies, but will detail particular elements borrowed from them for the formulation of optimization problems centered around these NOMA-MEC architectures. MEC architectures are reintroduced and we explain how they are modeled in the authors' proposals, as well as the different offloading models used. We also describe how NOMA is defined and applied in this type of problem.

A. MULTI-ACCESS EDGE COMPUTING

MEC architectures have a long history that dates back to the convergence of mobile networks and cloud computing. This network architecture arose as a solution to meet the growing need for low-latency, high-bandwidth applications and services. Its origins can be traced back to the European Telecommunications Standards Institute's (ETSI) introduction of MEC in 2014, which was fundamental in the development of MEC as an industry standard, with the goal of enabling the deployment of computational resources closer to the network edge. A framework for the implementation of this architectures is compiled in the ETSI MEC standard which supports an open and interoperable ecosystem. It outlines the functional architecture, APIs, protocols, and interfaces necessary for interoperability among MEC platforms, networking hardware, and application developers. It addresses a number of topics, such as application lifecycle management, MEC application discovery, mobility support, and security issues [15].

Because of this proximity, real-time and context-aware services may be provided, thus lowering latency by eliminating the need to backhaul data to centralized data centers. Indeed, MEC architectures are based on the deployment of computing nodes at the network edge, known as MEC nodes, that can be hosted within base stations (BSs) and access points. These nodes run applications by utilizing virtualization and distributed computing technologies (cf. Figure 1). To efficiently manage resources and maximize service delivery, these architectures employ techniques as Network Slicing (NS) and Software Defined Networking (SDN) [16]. This way, MEC is able to support multiple functionalities by putting processing, storage, and networking capabilities closer to end devices, enabling data offloading. Particularly, this offloading stages reduce network congestion, improve user experiences, and provide support for new compute-intensive services, such as real-time analytics, video processing [17], content caching [18], and other edge-based services, while limiting the energy consumed by the users.

Several approaches are considered for MEC servers' tasks offloading, including binary offloading and partial offloading, which optimize resource usage and enhance performance. By offloading the complete execution of an application or task from an end device to a MEC server, binary offloading lowers the computational load on the device and exploits the computing capabilities at the network edge. As opposed to this, partial offloading offers a selective offloading method in which only a portion of an application is transferred to the MEC server and the rest is run locally on the device. This technique offers fine-grained control over the offloading process, enabling the optimization of resource usage by exploiting the capabilities of both the device and the MEC server.

MEC architectures are used in many different fields such as 1) C-ITS to manage for example task detection and path panning [19]; 2) Smart city to enable heterogeneous systems inteoperability [20]; 3) telemedecine to provide rea-time heathcare services [21]; and 4) Industry 4.0 to enhance machine to machine communications and decision-making processes [22].

B. NON-ORTHOGONAL MULTIPLE ACCESS (NOMA)

In 2017, to improve the performance of MEC architectures, researchers put forward proposals to integrate NOMA as the physical layer of the Radio Access Network (RAN). NOMA is a multi-user communication technique that allows multiple users to simultaneously access the same time-frequency resources in wireless networks [12]. Unlike traditional Orthogonal Multiple Access (OMA) schemes, NOMA employs power domain multiplexing, where users are allocated different power levels to transmit their signals. Users with weaker channel conditions are assigned higher power levels, enabling them to overcome their weaker links and ensuring reliable reception. **It's important to note that in the literature, the use of the terms NOMA uplink or NOMA downlink does not refer to communications directions from a network point of view** (uplink and downlink), but to the way offloading is managed. (cf. Figure 2):

- **Uplink-NOMA**: this version is based on the objective of simultaneously offloading the tasks of several users to the MEC server onto the same radio resources. This suggests that the BS coordinates centrally how the users will exploit the radio resources during the offloading phase. To do this, prior exchanges were necessary between the users and the BS. In the first phase, the users provide the BS with feedback on information such as the CSI, the size of the tasks to be offloaded and the local computing resources available, so that the BS can

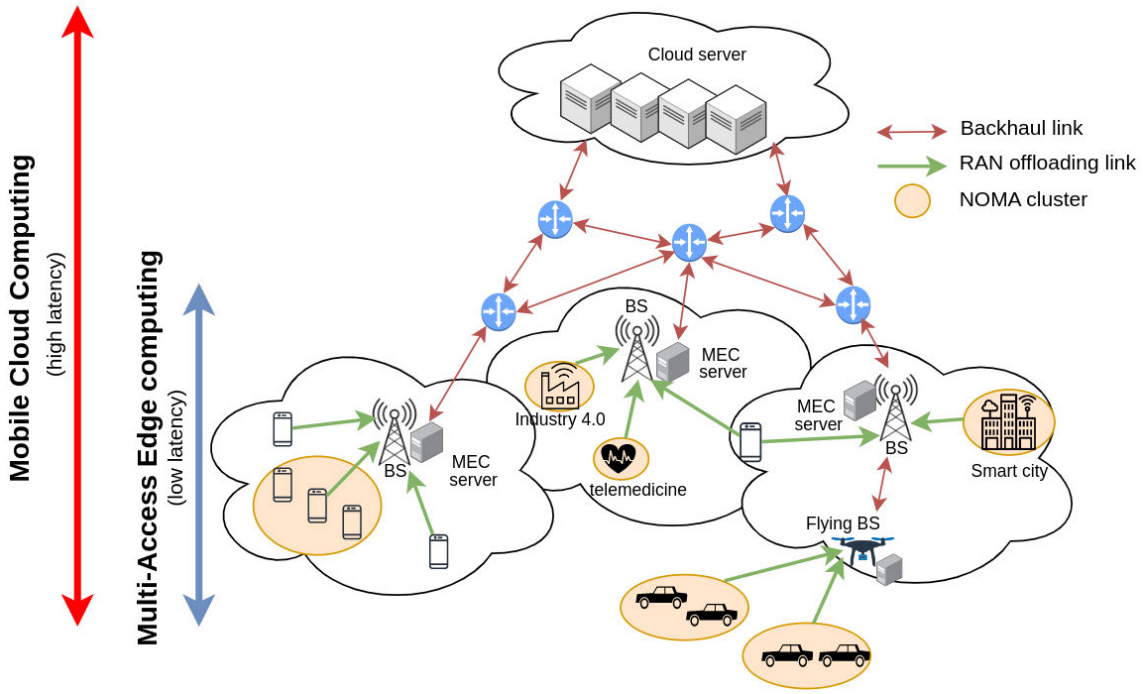

FIGURE 1. NOMA-assisted MEC architectures.

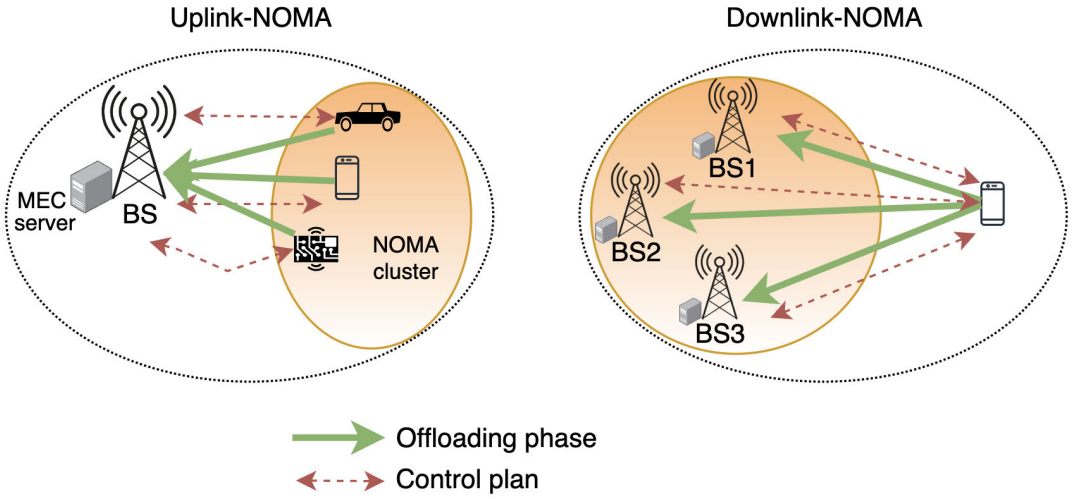

FIGURE 2. NOMA-Uplink and NOMA-Downlink communications.

accurately determine the most suitable offloading policy to adopt. The BS then sends the established parameters, power coefficient and selected radio resources, to each user for their offloading phase;

- **Downlink-NOMA**: this type of communication is also used for offloading phases (i.e. from users to the network, uplink direction), although this may be counter-intuitive due to its designation. Like Uplink-NOMA communications, Downlink-NOMA allows data to be offloaded to the MEC server, but enabling a user to simultaneously offload tasks to multiple MEC. This suggests that the decision and coordination of the offloading phase falls under the responsibility of the user. A preliminary phase therefore consists of several BSs equipped with MEC servers communicating their CSI and available computing resources to the user, so that the latter can establish the optimum offloading policy based on this information.

In all cases, NOMA uplink and downlink are based on the principle of SC on the transmit side, and SIC on the receive side. Figure 3 shows the principle of these two techniques in the context of transmission (SC) and reception (SIC). NOMA symbols are represented in a constellation, in which user1 (red arrow) aims to send bits 00, and user2

(blue arrow) wishes to send bits 01. These bits are then mapped into a 4-Quadrature Amplitude Modulation (QAM) digital communication, simultaneously on the same radio resource. Initially, the SC is based on equation (1) where the symbol transmitted (Sn - violet arrow) is equivalent to the sum of the signals from several users, weighted by a coefficient established according to the quality of their channel, so that condition (2) is respected, which in other words corresponds to a limited power budget (in our examples, the total power is fixed at 1).

$$S[n] = \sum_{i=1}^{N} \sqrt{\alpha_{u_i} \times P_{tot}} \times s_{u_i}[n] \tag{1}$$

$$\sum_{i=1}^{N} \alpha_{u_i} = 1 \tag{2}$$

These coefficients are then set so that the user with the worst channel can have the maximum power. Indeed, this permits to prevent its signal from being significantly modified by the poor conditions offered by its communication channel, enabling it to be decoded by the base station. In our case, user1 is assigned the $\alpha = 0.75$ coefficient because it has a worse channel than user2, which is assigned the $\alpha = 0.25$ coefficient, and will be considered as noise in user1's signal as in (3).

$$\underbrace{S[n]}_{superimposed\ symbol} = \underbrace{\sqrt{\alpha_{u_1}} s_{u_1}[n]}_{user1\ symbol} + \underbrace{\sqrt{\alpha_{u_2}} s_{u_2}[n]}_{user2\ symbol} \tag{3}$$

Secondly, during reception, the Sn symbol is partly filtered by the different user channels and Additive White Gaussian Noise (AWGN) noise is added to the received Yn signal, as shown in equation (4). In our example, these effects are negligible on Sn, which is equivalent to Yn.

$$\underbrace{Y[n]}_{received\ symbol} = \underbrace{h_{u_1} \sqrt{\alpha_{u_1}} s_{u_1}[n]}_{user1\ symbol} + \underbrace{h_{u_2} \sqrt{\alpha_{u_2}} s_{u_2}[n]}_{user2\ symbol} + \sigma^2 \tag{4}$$

At this point, SIC is used to find the symbol of each user:

- User1 only has to decode the symbol Yn received as his own to find the transmitted message (user2 is not considered to be noise);
- On user2's side, it first decodes user1's message. Once it has obtained user1's symbol, it then subtracts user1's weighted theoretical symbol from Yn, so as to extract the noise that carries its symbol, as shown in equation (5).

$$y_{u_2}[n] = Y[n] - h_{u_2} \sqrt{\alpha_{u_1}} s_{u_1}[n] \tag{5}$$

These two techniques allow NOMA communications to fully exploit the power domain, which significantly improves spectral efficiency, increases capacity, and enhances the overall system performance. However, the above formulas are not representative of the NOMA-assisted MEC domain. In fact, all the proposals made in this area are based on the formulation of a mathematical model representing the underlying architecture. NOMA is then expressed in terms of the maximum theoretical throughput that can be obtained for a user on a subcarrier, as shown in equation (6). Compared to the formulation of the theoretical throughput of an OMA communication, we find in particular the term of interference between symbols ($P_{tot} |h_{ui}|^2 \sum_{k=i+1}^{N} \alpha_{uk}$) specific to the superposition of several users on the same subcarrier.

$$R_{ui,sp_j} = \log_2 \left(1 + \frac{\alpha_{ui} P_{tot} |h_{ui}|^2}{P_{tot} |h_{ui}|^2 \sum_{k=i+1}^{N} \alpha_{uk} + \sigma_n^2} \right) \tag{6}$$

In the case of two users, user1 and user2, sharing the same subcarrier, the theoretical throughput for each is equivalent to equations (7) and (8) respectively.

$$R_{u1,sp_1} = \log_2 \left(1 + \frac{\alpha_{u1} |h_{u1}|^2}{\alpha_{u_2} |h_{u_1}|^2 + \sigma_n^2} \right) \tag{7}$$

$$R_{u2,sp_1} = \log_2 \left(1 + \frac{\alpha_{u2} |h_{u2}|^2}{\sigma_n^2} \right) \tag{8}$$

This formulation of throughput rates is used to obtain the delay in the task offloading phase, by expressing the ratio between the task's size (bits) and this theoretical throughput rate. Finally, the introduction of decision variables into the theoretical throughput formula, such as a parameter taking the value 1 or 0 in the numerator of the fraction present in the logarithm, makes it possible to manage the allocation or non-allocation of the subcarrier to a user. Similarly, the weighting of the ratio between the size of the task and the theoretical throughput can be used to define the offloading policy.

III. TAXONOMY

This section presents the taxonomy we defined to understand and analyze NOMA-assisted MEC offloading. Most of the studied papers were formulated as optimization problems. The structure of these documents can be summarized by formulating the following problem: the authors proposed a given environment composed of **use case** with a particular **RAN type** and some different **communication types**, where they aim to minimize/maximize an **objective**, implementing **algorithms** controlling **decision variables** and **new technologies**. These elements are shown in Figure 4 and detailed in this section.

A. OPTIMIZATION CRITERIA

The main objective of a NOMA-assisted MEC architecture is to provide additional computational capabilities to users, allowing them to process their tasks as quickly as possible (latency) while consuming as little energy as possible (energy). Most authors have therefore presented proposals to minimize these elements by using them as optimization criteria in the mathematical formulations of the optimization problems. Some authors have addressed other objectives, or have proposed new metrics to formulate these criteria. The main criteria we identified are as follows:

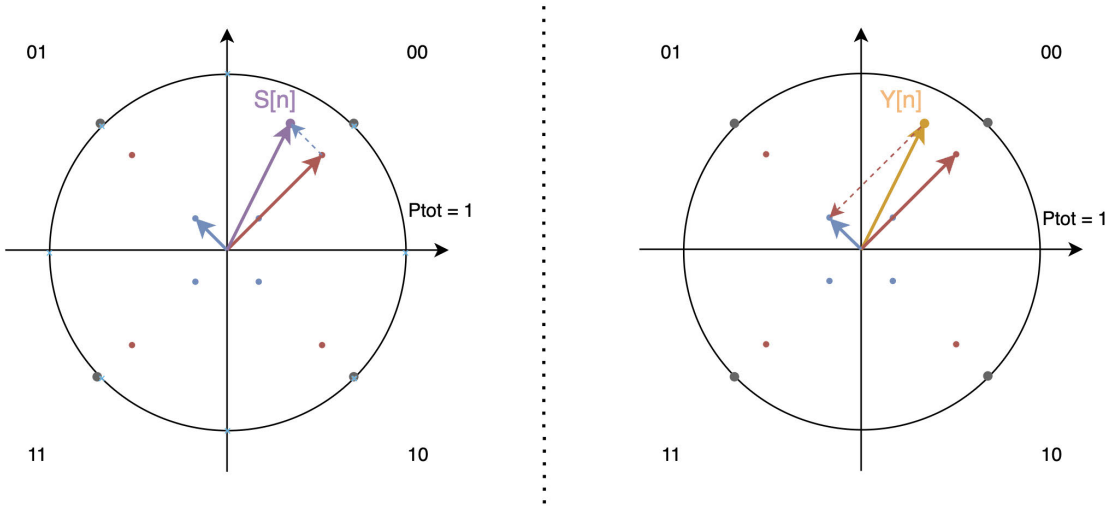

FIGURE 3. SC and SIC constellations.

- **Energy**: Energy minimization is one of the main criterion for the implementation of a NOMA-assisted MEC architecture. Indeed, the users of this environment are often IoT devices with low battery autonomy. The researchers propose to minimize energy at several levels, such as the energy consumed locally by each user, the energy consumed by the users to transmit and download their data to the MEC server, the energy consumed by the MEC server, or the combination of all these criteria which then corresponds to the overall energy consumed in a NOMA-assisted MEC architecture;
- **Delay/Latency**: Providing sufficient computing resources to users with limited computing capacity allows them to speed up the processing of demanding computing tasks and thus reduce their processing time. Authors aim particularly at providing transparent processing times to users, i.e. the shortest latency between the moment of offloading the task and the response containing the result of the task's calculation. The underlying objective is to minimize the processing times of the tasks (at device and MEC server level) in these architectures, and also the latency of the NOMA communications;
- **Joint Weighted Energy and Latency (JWEL)**: The above two objectives are often combined in the problem formulation although one is defined as an optimization criterion and the other is formulated as a constraint. This is why some authors have defined the JWEL criterion, so that the two objectives become simultaneously a single optimization criterion. In addition, by adjusting the weighting coefficient, the most significant metric can be defined according to the proposed use case;
- **Security**: Due to the properties and nature of the air-waves, RANs communications are broadcasted and are therefore subject to several external attacks. An attacker could conduct passive eavesdropping, compromising the confidentiality of such a system. Some authors proposed new mechanisms to improve security;
- **Quality of Service (QoS)**: Given that not all offloaded tasks have the same performance requirements, as they need shorter or longer processing times or higher or lower reliability, objective may be to determine how the MEC system can meet the QoS level of each task;
- **Processing rate**: The latency criterion can be reformulated by considering the capacity of the MEC server architecture to process as many tasks as possible. In this case, the objective is to maximize the processing rate;
- **Outage probability**: The probability of the system to crash can also be considered as a criterion allowing to quantify the probability that the network architecture fails to meet a certain threshold of reliability or QoS;
- **Amount of transferred data**: Maximizing the amount of data transferred to the MEC server, i.e. the capacity of the system to receive, manage and process data from multiple users, is another objective. However, this metric must be formulated with energy and latency constraints to be representative in NOMA-assisted MEC architectures.

B. COMBINATION OF NOMA AND OTHER TECHNOLOGIES

To improve the strength of the proposals, and/or to overcome some of the limitations of MEC architectures and NOMA communications, some authors propose to combine several emerging technologies on both the access and backhaul networks. While the introduction of these emerging technologies may overcome the limitations of these architectures, they introduce new management issues and bring their own shortcomings into the model. These include:

- **Content caching**: Content caching consists in storing frequently requested data on MEC server, close to the users. This allows users to access content much more quickly without having to go further up the network, thus reducing task processing time and ensuring the required

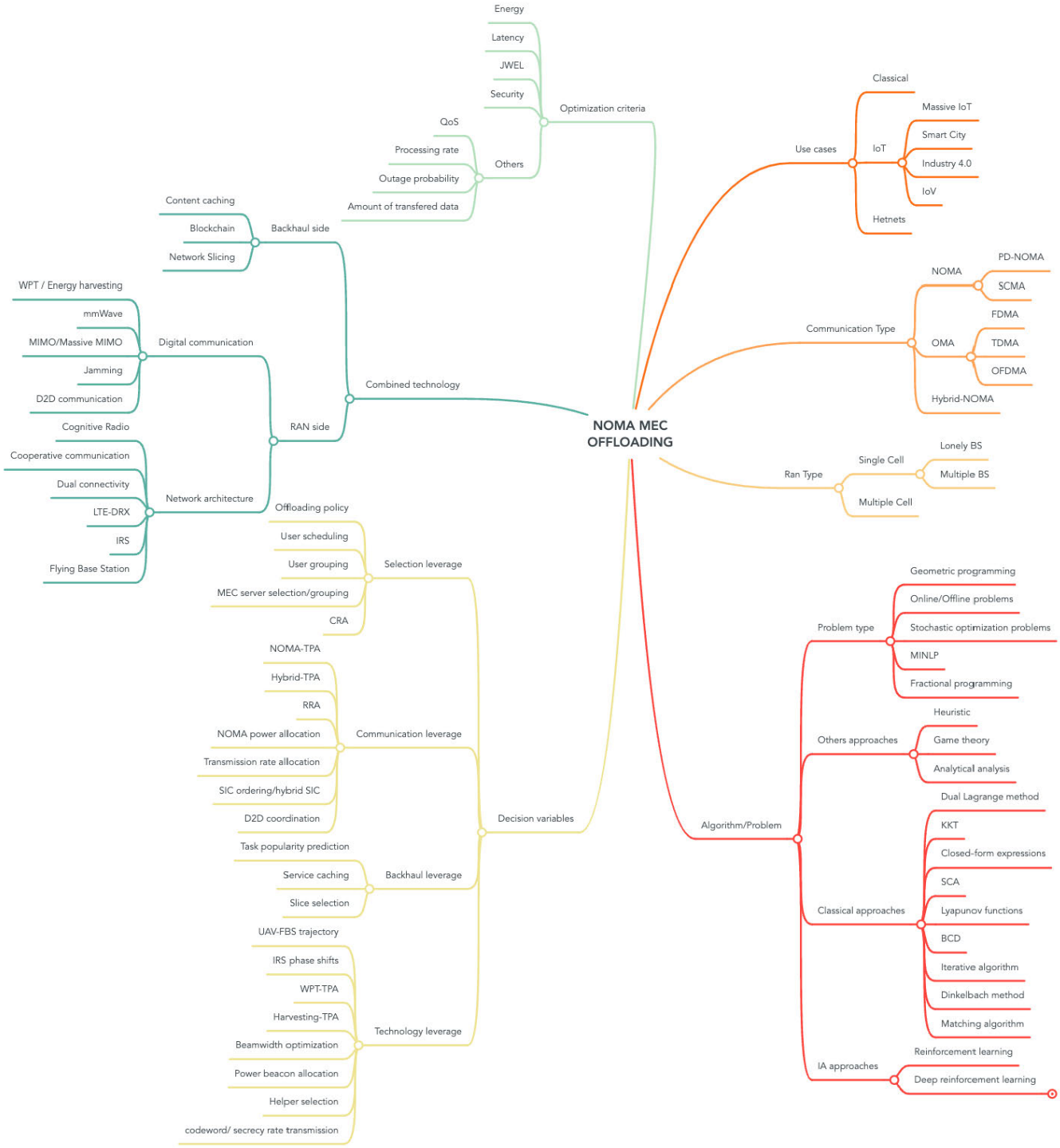

FIGURE 4. NOMA-MEC taxonomy overview.

QoS [23]. This also allows the backhaul to be unloaded by reducing the number of processing tasks to be carried out and the network congestion, minimizing the energy consumed in this type of architecture, and allowing to increase the number of supported users;

- **Blockchain**: Blockchain in MEC is a newly proposed strategy that aims to give users efficient, secure, and decentralized data sharing and processing capability [24]. Users can offload their task in a trustless environment without the need for a centralized entity. The capacity of Blockchain to improve data security, privacy, and transparency through the use of cryptographic techniques, making it more difficult for attackers to infiltrate the system. Blockchain can also support

resource sharing, and enable decentralized resource management. However, a number of factors, including the system's scalability, interoperability, and latency, affect how well Blockchain works in MEC. Indeed, this technology can be computationally demanding which increases the energy consumed and degrades the performance of NOMA-MEC systems;

- **Network slicing**: Network Slicing (NS) consist in splitting the MEC network infrastructure into several logical network slices that have their own key performance (slice with very low latency, slice providing high connectivity, slice providing high reliability etc.) to process tasks in the most optimal way, ensuring a sufficient QoS [25]. This provides NOMA-assisted MEC architectures with features such as flexibility, efficiency, and connectivity by supporting multiple user groups with heterogeneous QoS tasks on the same infrastructure. However, the efficiency of NS depends of the system's complexity, the slices' management, and the infrastructure's scalability;
- **Wireless Power Transfer (WPT) / Energy harvesting**: This technology allows through the use of wireless technologies to transfer power that will be gathered by the users of the MEC network to reduce the dependence on wired power sources or batteries and enabling cost-effective and sustainable MEC deployment [26]. The base station sends power beacons which are then harvested by the users to increase their lifetime, providing them with more power to process their tasks locally, or giving them energy to offload their tasks to the MEC server. From a technical point of view, WPT can be achieved through various techniques such as inductive coupling, resonant coupling, and electromagnetic radiation. Nevertheless, the effectiveness of these techniques depends on the variability and unpredictability of the energy sources, requiring advanced resource control and management;
- **mmWave**: Communications using the frequency bands between 30GHz and 300GHz are called Millimeter-wave communications because their wavelength decreases when the frequency increases. This type of communication provides high bandwidth and very low latency communications and could therefore be useful in a MEC context [27]. However, due to the small wavelengths, these communications are subject to strong attenuation phenomena and are susceptible to be hindered by obstacles such as buildings, degrading the transmitted signal and the communication reliability. However, by combining mmWave and massive MIMO it is possible to bypass these limitations [28]
- **MIMO/Massive MIMO**: Massive-MIMO enables digital communications to transmit to multiple users simultaneously by spatial multiplexing, formed by multiple beamforming, using a large number of antennas at the BS. In particular, it increases spectral efficiency and signal quality, which ensures better radio cell coverage, greater connectivity, higher data rates and better latencies [29]. Coupled with NOMA, it can greatly increase network capacity by superimposing user signals in beams, especially for massive IoT use cases. Nevertheless, the use of this technology generates energy overheads due to the complexity of demodulating this type of communication (especially with user mobility and inter-cell interference), which is particularly challenging in MEC-type architectures where we aim to minimize the energy consumed;
- **Jamming**: In security, jamming techniques are initially used by attackers wishing to affect the availability and proper functioning of the network. The attacker emits a strong signal over the entire operating frequency band to jam the channel, thus making communication impossible [30]. However this tool recently has been adapted to overcome the problems of passive listening in radio networks by an eavesdropper, in particular by combining it with NOMA, where two users join forces, one to transmit this data, the other to jam and prevent any passive listening;
- **Device-to-Device (D2D) communication**: D2D communications permit RAN users to communicate directly with each other without using the infrastructure such as BSs [31]. This type of communication is often used in operator networks such as 5G sidelink, or is particularly effective in vehicular networks with V2V communication type, i.e. in dense and high mobility environments, allowing to increase network capacity, reduce latency and traffic congestion at the BS. Combined with NOMA, it allows D2D communications to be overlapped on top of conventional RAN communications, which is beneficial to the NOMA-assisted MEC architecture. However, this type of overlapping increases the interference inside the cell, which weakens and degrades the performance of users when offloading their tasks;
- **Cognitive radio**: Cognitive Radio (CR) allows multiple users (secondary users) to dynamically choose optimal frequency bands for communication that are not occupied by other users (primary users and licensees) at that moment. By permanently reusing the radio spectrum, this greatly improves the spectral efficiency and capacity of the network [32]. However, the use of complex algorithms and hardware to operate this type of technology can increase latency and energy consumption in NOMA-assisted MEC architectures;
- **Cooperative communication**: To cope with bad channel conditions, which can be degraded due to the distance to the BS, the mobility of the user, and the multipath, the use of a relaying node, is employed to enable the offloading phases. This is especially the case for users at the cell edges, that will have insufficient conditions to offload their tasks and that will therefore need cooperative communications [33].

In a first step, this user will offload part of his tasks to another user, who will then have the possibility, depending on the availability of his computing resources, to process the task locally or to transmit it to the MEC server. This approach greatly improves the performance of NOMA-assisted MEC architectures by exploiting the available resources of all the actors as much as possible. Nevertheless, the complexity of such a system, in particular the coordination of all the actors, becomes limiting as the number of users increases;

- **Dual Connectivity (DC)**: This technology allows users to connect to multiple BSs simultaneously enabling faster and more reliable data transmission, improved coverage, and reduces latency [34]. It is mainly found in Non-StandAlone (NSA) 5G systems, where it allows a user to associate simultaneously with a 5G NR BS and an eNB BS. However, its implementation in network architectures increases complexity, signaling overhead and interference due to the need to coordinate the two base stations, which increases the energy consumption;
- **LTE-DRX**: LTE Discontinuous Reception is a power-saving technology used in 4G LTE networks to save energy on user side [35]. It enables a device to periodically turn off its receiver while still being connected to the network which helps to reduce energy consumption. However, as the user's device must re-establish connectivity with the network after each sleep period, this technique increases latency;
- **Intelligent Reflecting Surface (IRS)**: These are passive arrays of elements that reflect and direct signals to extend the radio coverage area, and/or bypass an obstacle, and/or increase the transmitted signal strength [36]. In NOMA-assisted MEC architectures, its passive nature does not increase energy consumption, which is a major advantage over other solutions. Nevertheless, its combination with NOMA requires more optimization efforts to limit interference. In addition, the management of this technology can be complicated in cases of high mobility;
- **Flying Base Station (FBS)**: The use of an Unmanned Aerial Vehicle (UAV) as a FBS in MEC architectures introduces a paradigm shift, where the access network comes to the users and not the contrary (such as cooperative communications). This UAV can either act as a MEC server providing computing power to users, or simply as a relay transmitting offloaded data to another BS [37]. It can therefore provide edge computing services in areas where there is no connectivity or in highly mobile environments. However, the use of such a tool does not provide much computational capacity. Moreover the use of a battery to power the FBS, run the computations on the embedded server, and manage the communications greatly limits the performance of this type of approach.

C. USE CASE

Although most papers are classical optimization proposals, some authors shape the optimization problem according to certain use cases with specific requirements and constraints:

- **Classical use case**: The user is a simple actor in the network that offloads a certain amount of data. These computing capacities (CPU frequency, power budget) are often dimensioned to look like a smartphone or a laptop;
- **IoT/Massive IoT**: This refers to physical devices with the ability to be connected to the Internet and communicate with each other via the MEC infrastructure. They generally provide extensive data feedback from their sensors to centralized applications feeding AI for monitoring and decision making [38]. However, these devices are limited in terms of computing power and autonomy, as they are battery powered. NOMA-MEC architectures can greatly improve the use of these devices, especially in cases of massive IoT requiring high connectivity and network capacity process the massive data traffic [39];
- **Smart City**: this is a broader use case, many IoTs are used to monitor a city to improve the quality of life of its inhabitants, the efficiency of public services (f.e. optimal rubbish collection) and to supervise the environment sustainability: pollution control, hygrometry, temperature, etc [40]. The smart city use case (high density, heterogeneity) requires communications that can adapt to and support such an environment. Interoperability, security and privacy issues are raised in this type of environment;
- **Industry 4.0**: the combination of IoT with IA, robotics and augmented reality constitutes the new industrial revolution, allowing to optimize production processes, production quality, and energy use [41]. However, it implies constraints such as very low operating latency to monitor and manage production in real-time;
- **Internet of Vehicles (IoV)**: IoV brings together multiple connected objects from the vehicular domain. These can be road infrastructure sensors as well as autonomous and connected vehicles. These technologies, based on the cooperation of all the actors that compose this environment, make it possible to improve road safety, manage traffic efficiently and reduce pollution [42]. Applications in this context require very low latency and very high levels of reliability (f.e. cooperative emergency braking);
- **Hetnets**: this is an environment made up of several overlapping radio cells that aim to serve a wide variety of use cases, such as IoV, industry 4.0, smart city, etc. A Macro Cell could be used to manage and connect users in a city, while several small cells could be used to retrieve data from an urban area and manage vehicular traffic. This type of environment requires the MEC architecture to be able to adapt to the wide range of QoS required.

D. COMMUNICATION TYPE

We were able to identify various NOMA and OMA techniques that have been studied in the literature:

- **Power Domain-NOMA (PD-NOMA)**: This is the most widely considered NOMA approach in the literature, which uses the principles of SC to assign multiple users to the same radio resource simultaneously;
- **Sparse Code Multiple Access (SCMA)**: This type of communication constitutes the second NOMA domain based on codes, more precisely on the sparcity properties of codes. It introduces new decision variables in the formulation of the authors' proposals such as the design of the codebooks or the design of the factor graph matrix;
- **Frequency Division Multiple Access (FDMA)**: This OMA technique divides radio resources into several frequencies, each of them being allocated to a different user;
- **Time Division Multiple Access (TDMA)**: This technique OMA consists of dividing the time into several time slots, which are then allocated to each user;
- **Code Division Multiple Access (CDMA)**: This type of communication, widely used in Universal Mobile Telecommunications Service (UMTS) networks, allows several users to share the same frequency band by assigning each user's signal a unique orthogonal code;
- **Orthogonal Frequency-Division Multiple access (OFDMA)**: Used in recent communications systems, it is a wireless communication technique that uses multiple orthogonal sub-carriers to transmit data simultaneously to multiple users in the same frequency band;
- **Hybrid-NOMA**: Several types of hybrid-NOMA communications are distinguished in the literature. Firstly, some authors consider that these communications are made up of several sub-carriers, some of which can be overloaded by several users, while others can support only one user. The other approach to hybrid-NOMA communications is to start by offloading on the basis of a NOMA communication and then, when there is only one user left to offload, switch to an OMA communication. This makes it possible to correct the mathematical model for calculating transmission rates in certain papers, where the value of the NOMA interference changes when a user stops offloading (which is not always taken into account).

E. DECISION VARIABLE

To optimize NOMA-MEC architectures, researchers propose the use of different leverages that become decision variables in the formulation of the proposed optimization problem. They can be classified into several categories, corresponding to the different components of the architecture.

First, we find the communication leverages allowing to modify, adapt, coordinate and control the NOMA and/or OMA type digital communications in the RAN:

- **NOMA Time Phase Allocation (NOMA-TPA)**: This leverage is used to set the transmission times that several users occupy when simultaneously offloading their tasks onto a NOMA radio resource. It is often set as a trade-off between several users to make fair use of the radio resource, i.e. without penalizing the user with the best channel conditions and without degrading the offloading performance of users with a weak channel. Also, by defining shorter offloading times, it can improve offloading security by limiting the ability of an eavesdropper to listen to the communication;
- **Hybrid-TPA**: Like NOMA-TPA, it allows to allocate the duration of the communication phase but for a Hybrid NOMA-OMA communication. In particular, it enables the first transmission phase to be set to NOMA, followed by the OMA phase when only one user remains, which often benefits users with good channel conditions by enabling them to increase their offloading performance;
- **Radio Resource Allocation (RRA)**: it is globally defined as the way to allocate the radio resources available at each moment to access the channel. However, due to the nature of the formulation of the optimization problems, these radio resources do not really have a physical meaning but only a mathematical one (functions or vectors), and are considered more as blocks of available resources, labelled by the authors according to the context and the nature of the underlying communications (NOMA, OMA). The main types of RRA are block resource allocation, sub-channel allocation, sub-carrier allocation and the factor-graph matrix design in SCMA;
- **NOMA power allocation**: This is the most widely used tool in the literature to manage NOMA communications. During the Uplink phases, it enables the allocation of transmission power levels to the base station for each user sharing the same radio resource, and according to the channel status of each one. Due to the weighting of the power level, it allows users to reduce their consumption during the transmission phase. However, it is set optimally by the optimization algorithms, to avoid degradation of the transmission performance and to respect a balance between energy minimization and task processing time. In the case of NOMA-downlink communication, it allows a user to offload his tasks more efficiently to several MEC servers while respecting the channel state of each one;
- **Transmission rate allocation**: This leverage refers to the process of dynamically allocating various transmission rates to many users based on their channel conditions and QoS needs. This permits effective use of the radio resources and improved capabilities for users with different communication demands;
- **SIC ordering/Hybrid SIC**: The SIC ordering algorithm designates the order in which user signals are detected and decoded at the BS. Both decoding accuracy and

interference cancellation are affected by the SIC ordering, which impacts system performance. The optimal SIC ordering decoding according to different objectives, such as processing time, is not necessarily the classical decreasing order of the channel qualities of each user (Hybrid SIC);
- **D2D coordination**: This leverage refers to the use of nearby mobile devices to facilitate the offloading process, which can enhance the system's efficiency and reduce communication delays by not using RAN's main radio resources. However, the challenge of this type of leverage is to limit the amount of interference that D2D use causes with other communications in the radio cell.

The management and coordination of the different nodes of NOMA-assisted MEC architectures are also tools considered:

- **Offloading policy**: This is one of the main leverages used to efficiently manage offloading to the MEC server to minimize the energy consumed and the processing tasks duration. Depending on the context, the offload policy could be partial offloading or binary offloading. In scenarios with several MEC servers, the offloading policy also corresponds to the selection of the servers to assign the different tasks to offload and process them as efficiently as possible;
- **User scheduling**: It defines the process of selecting the subset of users that will offload tasks to the MEC servers depending on their channel conditions, tasks requirements, and priority in stochastic optimization problems;
- **User grouping**: This mechanism relates to the clustering of users with comparable channel and computational characteristics so they can be scheduled for offloading together. By exploiting the variety of channel and task requirements, user grouping seeks at improving system efficiency while reducing interference and resource allocation overheads. It has a direct impact on the quality or degradation of the performance of NOMA communication. It would be especially true if a group had to be formed with a user at the edge of the radio cell and a user closer to the antenna, which corresponds to a OMA use case rather than a NOMA use case;
- **MEC server selection/grouping**: This is the mechanism that allows a user to define a group of MEC servers that will simultaneously be used to offload tasks. This MEC servers' selection depends on the quality of the channel of each MEC server BS, its workload and availability. A balance must be found between a large number of MEC servers, which would limit the transmission performance for offloading calculations, and a small number of servers, which would increase the task processing time;
- **Computation resource allocation (CRA)**: Since the primary goal of a MEC architecture is to provide additional computational resources to network users, the authors have defined different metrics to reflect the computational resources available on each network entity. From a local point of view, i.e. at the level of the RAN users, computational resources are often expressed as the definition of CPU frequencies or computational time (inversely proportional to the frequency, which consequently sets the CPU frequencies) to locally execute the task. At the MEC server level, even though they are often considered unlimited, some authors quantify these computational resources to fit real scenarios. They then define policies to manage the computational frequencies or define a limited number of blocks of computational resources to process the offloaded tasks (cf. Section IV).

Some authors focused on backhaul network optimization:

- **Service caching**: Like CDNs, MEC architectures have the ability to store data close to users to deliver real-time services [18]. Thus, this metric aims at defining optimal policies for storing tasks as close to the users as possible, particularly in architectures consisting of several radio cells equipped with MEC servers;
- **Task popularity prediction**: In real systems, the arrival and nature of tasks are stochastic. Thus, to improve the adaptability of MEC architectures to efficiently handle offloaded tasks according to their QoS constraints, it is necessary to anticipate and predict the nature of the tasks to prioritize their processing. In caching systems, this allows to store tasks/data that will be requested or exchanged several times by several users;
- **Slice selection**: This leverage allows a task to be assigned to a network slice according to the QoS expexted by this task. The aim of this process is to match the performance provided by each slice to the QoS constraints of each task.

Combining NOMA with new technologies, new leverages are also introduced to drive them efficiently without degrading the functioning of the underlying NOMA-MEC architecture:

- **UAV-Flying Base Station (UAV-FBS) trajectory**: The introduction of FBSs to best serve the users of the network is proposed by several authors. They define the optimal trajectories (x,y,z coordinates in space) that the UAVs must follow to cover a specific area of users, taking into account their CSI and the mobility of each of them [43]. This allows to improve the CSI of each user and thus to enhance the performance of task offloading thanks to an optimal placement. Nevertheless the calculation of this trajectory is defined by several constraints such as a limited energy resource, UAV's displacement capacities, the obstacles and the mobility of the different users;
- **IRS phase shifts**: This leverage refers to an IRS capacity to modify the phase of electromagnetic waves that reflect off of it. Thus, it can direct and focus the reflected waves by modifying the phase shifts of the different elements that compose it, which enhances the signal quality and coverage of wireless

communication systems, especially in large or obstacle-constrained environments [44];
- **WPT-TPA**: This metric defines the time during which the power beacons sent by the BS will be broadcast to distribute energy to the RAN users. This must be defined in such a way that it does not interfere with the transmission phases, while ensuring sufficient energy delivery to increase IoTs' lifespan;
- **Harvesting-TPA**: Like WPT-TPA, it allows the user to define the period of time during which an IoT will harvest the energy transmitted by the network architecture via power beacons. It is also defined to ensure a balance between using an external power source to increase lifetime and minimize energy consumption, and performance during offloading and task processing phases;
- **Beamwidth optimization**: It consists in modifying the width of the electromagnetic wave beam sent out by antenna to improve the performance of the communication system in mmWave communication. This optimization is essential for addressing the significant path loss and the high directional nature of mmWave signals, in particular for the NOMA communications which contain several superimposed signals;
- **Power beacon allocation**: This leverage consist in assigning power sources, known as power beacons, to specific locations in a wireless charging network. Thus, by figuring out how many power beacons are required, where they should be placed, and how much power they should have, it is possible to maximize charging efficiency while reducing energy waste and interference;
- **Helper selection**: Cooperative-NOMA systems are made up of users who will offload their tasks through nodes called Helpers. A user must select a helper that will enable him to improve his performance during the offloading phase. This selection is based on the distance from the user to the helper, the quality of the channel between user/helper and helper/BS, and the computational and radio resources available on the helper;
- **Codeword/secrecy rate transmission**: It defines the maximum rate at which secure communication can be achieved in a wiretap channel using a specific coding scheme. It quantifies the amount of secret information that can be transmitted from a sender to a legitimate receiver while keeping the information confidential.

F. RAN TYPE

The experiments carried out in existing proposals are evaluated in several types of radio environments such as:

- **Single cell - One Base Station**: This is the environment most used to demonstrate the proposals made, consisting of a conventional radio cell with a BS. However, it does not allow the evaluation of user mobility;
- **Single cell - Multiple Base Stations**: This is a more complex version of the previous model to represent Hetnets models, cognitive radio based models, dual connectivity based models, or models which require user mobility;
- **Multiple cells**: It constitutes the most complex environment to model because it takes into account new parameters such as interference between radio cells, mobility in different cells and so on. It is particularly used in papers presenting high mobility environments such as IoV.

G. PROBLEMS TYPE AND ALGORITHMS

The description of a context with particular use cases, objectives and technologies, leads to formulate a mathematical problem. Although in most cases this leads to a classical optimization problem with a non-convex function to be minimized/maximized under constraints, some papers propose other alternatives such as game theory or AI tools. In any case, each problem is solved using a certain algorithm/solver specific to each problem formulation. It is therefore important in this section not to dissociate the type of problem formulated and the solution used to solve it. First, there are several types of problems depending on the formulation of the model and the decision variables used by the authors:

- **Geometric programming**: It is a particular class of optimization problem where the constraints and objectives are defined as convex monomials and posynomials [45]. Geometric programming can be utilized to formulate power allocation issues in the NOMA-MEC use case;
- **Online/Offline problems**: An offline problem requires that all information is provided up front, whereas an online problem in optimization includes making decisions sequentially and adaptively given incomplete knowledge. In the NOMA-MEC use case, online optimization can be used to allocate resources in real-time based on the state of the network such as task arrival, channel state information, while offline optimization can be used to design the system parameters, such as the user transmission power or resource allocation in order to maximize the system performance under a variety of hypotheses or scenarios;
- **Stochastic optimization problems**: Stochastic optimization deals with uncertain variables that have probabilistic distributions [46]. Similar to online problems, stochastic optimization requires sequential and adaptive decision-making based on incomplete information, but with the complexity of dealing with randomness and uncertainty in the decision-making process;
- **Mixed Integer Non-Linear Programming (MINLP)**: It is a type of optimization problem that is computationally difficult because it combines continuous and discrete variables [47]. MINLP can be employed to

address resource allocation issues in the NOMA-MEC use case that involve both continuous variables such as power and bandwidth allocation and discrete variables like user association and computing resource allocation;
- **Fractional programming**: It refers to a type of optimization problem that involves optimizing a ratio of two convex functions [48].

Most of the formulated problems are classical non-convex optimization problems such as MINLP which are solved using various methods. The main methods used are:

- **Dual Lagrange Method and KKT**: this approach is an optimization technique used to solve constrained optimization problems. It involves introducing Lagrange multipliers associated with the problem's constraints to create a new optimization problem called the Lagrangian. On the other hand, Karush-Kuhn-Tucker (KKT) conditions are a set of necessary conditions for a point to be a solution to a constrained optimization problem. Together, they are powerful tools in mathematical optimization, aiding in the analysis and solution of complex constrained optimization problems by introducing and leveraging the concept of Lagrange multipliers and their associated conditions.
- **Closed form expressions**: It refers to mathematical models that can be expressed in an explicit and analytical form, rather than iterative or numerical methods. It can be used to derive analytical solutions to optimize the resource allocation problem, such as the closed-form expression for the optimal power allocation;
- **Successive Convex Approximation (SCA)**: The SCA algorithm is a widely used method for solving non-convex optimization problems by iteratively approximating them with convex subproblems;
- **Lyapunov functions**: In stochastic optimization, Lyapunov functions are used to design stability criterion for the convergence of stochastic systems over time. They are used to analyze the long-term behavior of the system as the randomness of the problem evolves;
- **Block Coordinate Descent (BCD)**: This algorithm divides the decision variable into blocks and updates each block in a cyclic manner to solve a complex optimization problem. In NOMA-assisted MEC systems, this enables ideal radio and computational resource allocation policies to be found in a coordinated way;
- **Iterative algorithm**: It's an ensemble of optimization techniques that progressively get better at the solution until a good convergence criterion is reached;
- **Dinkelbach method**: It refers to a particular iterative algorithm that can solve fractional programming problems by transforming them into a sequence of linear programming problems;
- **Matching algorithm**: It belongs to a group of optimization algorithms used to determine the best combination of two sets of devices or entities (optimize resource allocation based on user demands and system constraints in NOMA-assisted MEC).

Other approaches can also be considered by implementing different algorithms and considering different formulations:

- **Heuristic**: It refers to a method for addressing issues that quickly and effectively finds approximations to difficult optimization problems;
- **Game theory**: It is a mathematical framework for examining how different actors interact strategically when each decision has an impact on the others. Game theory can be utilized in the NOMA-assisted MEC architecture to model interactions between users or MEC servers competing for limited resources and to identify the best course of action for maximizing system performance or achieving a predetermined level of fairness;
- **Analytical analysis**: It corresponds to a method used in optimization to employ mathematical methods and formulas to solve problems precisely. Under specific assumptions and conditions, analytical analysis can be used in the NOMA assisted MEC to obtain closed-form expressions for power allocation, bandwidth allocation, and resource allocation, and so on.

More recently, to address online models, new approaches based on artificial intelligence appeared:

- **Reinforcement Learning (RL)**: In NOMA-MEC, RL is a technique that instructs an agent to make decisions on rewards and punishments they receive from the environment. The agent gains knowledge on how to enhance the functionality of the NOMA-assisted MEC system in this situation. This is accomplished by teaching the agent's best policies using Q-learning approach or Policy Iteration algorithms (with markov decision process (MDP) as decision model);
- **Deep Reinforcement Learning (DRL)**: It is an extension of RL that uses Deep Neural Networks (DNNs) to learn complex representations of the environment. It includes many approaches which will be detailed in the following sections and which are used to adapt to the various environments and use cases proposed by the authors in the NOMA-assisted MEC systems;
- **SAQ-learning**: It refers to a variant of Q-learning that aims to learn the optimal action-value function using a single sample per episode, making it more sample-efficient than traditional Q-learning. It achieves this by using a soft-update rule to update the action-value function, rather than a hard-update rule;
- **Long Short-Term Memory (LSTM)**: It is a type of recurrent neural network that is able to remember long-term dependencies in data and avoid the vanishing gradient problem, making it well-suited for sequential data tasks such as time series prediction of NOMA-assisted MEC environment. It has the ability to handle variable-length sequences, uses backpropagation through time for effective training, and can selectively forget or remember certain information depending on the input;

- **Deep Deterministic Policy Gradient (DDPG)**: It refers to a model-free, off-policy algorithm used in DRL that combines the actor-critic architecture with the insights from deep Q-learning. In particular, it has the capacity to learn policies in contexts with delayed rewards and the handling of high-dimensional continuous action spaces.

IV. ANALYSIS AND COMPARISON OF STATE-OF-THE-ART SOLUTIONS

This section aims at presenting and comparing the main existing works applying the NOMA technology in a MEC context. The articles identified in the literature are classified into five broad categories according to their purpose: 1) energy minimization (Section IV.A), 2) delay minimization (Section IV.B), joint energy and latency minimization (Section IV.C), security (Section IV.D) and Others (Section IV.E).

A. ENERGY MINIMIZATION

Minimizing the energy required can be a first objective associated with the application of the NOMA technology. State-of-the-art solutions (cf. Table 2) aim at minimizing the energy consumed by users while respecting the time constraints imposed by the processing of their tasks. This can be particularly useful for devices with low energy capabilities that need to process a significant amount of data.

To the best of our knowledge, [49] is the first paper that introduced **NOMA in MEC architectures for energy minimization**. The authors have defined an offloading policy (local or remote computation) that involves 1) managing the computing power of users by modulating the frequencies of their CPUs, 2) allocating the transmission rate and power for each user, and 3) managing the SIC control at the base station. The proposes solution is based on a Dual Lagrange method and the presented results highlight the benefits of a partial offloading approach compared to a binary offloading approach to minimize the energy consumed by users, independently of the underlying type of communication. Nevertheless, the proposed scheme shows a very small gain in using NOMA over OMA to minimize the energy consumed.

Following on from this, the authors of [50] considered a **more specific scenario**: two users simultaneously trying to offload their tasks to a MEC server to reduce their energy consumption during an offloading phase. The notion of Hybrid-NOMA communications was introduced in this paper. This means that a user first offloads part of his calculations onto a radio resource already allocated to another user, and then offloads the rest of the calculation onto a time slot that is totally dedicated to him via an OMA communication. The authors defined both the allocation of the durations of the NOMA and OMA time phases (Phase 1 NOMA, then phase 2 OMA) as well as the transmission powers considering a Geometric Programming optimization problem. Although the hybrid scheme performs better than other schemes such as full NOMA or full OMA, this difference seems to decrease as the processing time constraint of the task to be offloaded is reduced.

Hybrid-NOMA was also considered in [51] to minimize the energy consumed by users under the constraint of respecting the **specific deadlines** of each task and the **transmission powers**. OMA communications can be specifically used depending on the critical nature of the processing time of its task. This allows the communication model to take into account the evolution of the NOMA interference value: when a user of a group completes the offloading phase, it leaves the group of users that is still offloading its tasks, modifying the interference value. In this paper, the use of the SCA method enables the treatment of the non-convex problem, and the use of the KKTs simplifies the problem and provides the closed-form expressions for the transmission powers and the time allocation of the offloading phases, i.e. the different NOMA phases and the OMA phases where only one user remains. The development of the analytical results leads to the understanding that the power allocation strategy for each user looks like a water-filling strategy which involves allocating more power to frequency bands with higher channel gains, and less power to frequency bands with lower channel gains.

In contrast to a global formulation of the problem, the **energy minimization** problem can also be formulated for **each user** during their offloading phases as proposed in [52]. This paper was the first to define the computational resources of the MEC server as a set of finite block resources, similarly to radio block resources. The heuristic algorithm then tackles the joint computation and communication resource allocation problem to cluster users that will share the same resource block, the allocation of transmission power, and the computational resources of the MEC server. The results show that the proposed algorithm provides a quasi-optimal solution with a much faster data processing. However, although the authors have mentioned 5G architecture several times, the formulation of the problem and the experiments conducted do not validate and demonstrate the use of NOMA in a 5G context.

Due to the complexity of formulating an optimization problem considering both uplink and downlink cases, the authors of [53] proposed for the first time to focus only on pure **downlink use cases**. Several data of different sizes dedicated to each user should be transmitted by the base station according to specific delay constraints. Dynamic user scheduling with transmission power allocation for SC is considered in this paper to minimize the long-term total energy consumed. The formulated stochastic problem is then transformed into a series of static optimization problem and treated with Lyapunov functions. Moreover, during phases when the user is not receiving, the authors propose to use the concept of discontinuous reception (LTE-DRX), in which the user goes into sleep mode, to enhance power reduction. The results show that the proposed algorithm quickly converges to an equilibrium point between the power consumption and the virtual queue backlog, and provides better performance in terms of energy minimization when the

system is underloaded, and provide a better user satisfaction when the system is overloaded. However, it should be noted that this proposed scheme can only be used up to a threshold of 60 users, beyond which the power constraint is no longer satisfied.

The identified problems are also related to specific models: **offloading model**, network model, etc. Regarding offloading model, a **binary offloading model** was for example considered in [54]. UEs offload all their computational tasks to the server without performing any local processing. Significant results were achieved in this paper using an efficient offloading policy and an SCA approach to efficiently allocate transmission powers, uplink and downlink phase transmission times. This is particularly the case when the size of the tasks to be offloaded is large and the deadline constraints are stringent. However, these results may be questioned due to the formulation of the underlying network model and the many assumptions made that do not match the actual operation of a network. For example, they considered that the channel does not change over the whole offloading period, i.e., the same gain over the entire uplink and downlink phases.

In [55], a **specific offloading model** was regarded. Considering that some users may only partially use their available computing resources, the authors proposed to rely on the availability of the computational resources of a neighboring node to perform task offloading. They defined a temporal division of the uplink communication into two phases. In the first phase, the user relies on NOMA to simultaneously offload to the MEC server and the neighboring node called the helper. In the second phase, the helper offloads part of its task to the MEC server and simultaneously executes user's task processing. To this end, the authors formulated two problems aimed at minimizing the consumed energy, increasing the number of data transmitted to the MEC server by partitioning tasks (policy offloading), managing transmit power, and managing the time of the various uplink phases. The use of KKT and BCD addresses the highly non-convex problem. The results showed that from an energy perspective, if the user consumes more energy than the helper, it will tend to offload its calculations. High SNR or distance between the helper and the user do not seem to be in favour of the proposed solution. It can be noted that the paper does not consider the case of feedbacks processed by the Helper and the MEC server during downlink phases, which would increase the energy consumption and the delay when transmitting the process results from the Helper to the user.

Regarding **network model**, many papers consider in their model that the prediction of task arrival and channel state can be accurately predicted. However, this assumption cannot be valid in real large-scale systems, where many IoTs offload tasks of different sizes to the MEC server in a dynamic and random way. Therefore, a new algorithm called NTORA based on Lyapunov functions was proposed in [56] to adapt to rapidly changing conditions in a smart city environment, where the main objective is to minimize the average energy consumed by each IoT in the system, constrained by the processing time of their tasks. Stochastic optimization techniques are used to transform the problem into a deterministic one, split into three sub-problems to address the offloading policies, and the allocation of computational resources of the MEC server. The results show that NTORA offers much better performance in terms of energy minimization and processing time than the full local processing or full offloading approach (binary), thus offering a scalable solution for the smart city and smart grid use cases.

Energy consumption can also be considered in a context of **user mobility**. For example, in [57], a dual-connectivity model has been proposed to describe an environment where a user connects to multiple base stations in the same cell that overlap through NOMA. To this end, they defined a reinforcement learning algorithm (Twin Delayed DDPG - TD3) that allows efficient management of transmission power and propose a heuristic method for segmenting the tasks to be offloaded. Although the idea of mobility is introduced in this paper, mobility management issues when a user moves from one radio cell to another, through the implementation of handover type mechanisms, were not addressed.

Beyond problems and models, energy consumption can also be attached to **specific use cases**. For example, **flying base stations (UAVs)** is today a major subject for increasing network coverage, and thus, in the long term, massively deploying MEC architectures. However, these technologies bring new constraints in terms of energy consumption. This is why many research works, such as [58] and [59], integrate into energy consumption models local consumption issues at the flying base station level. In this context, the authors of [58] defined a new framework allowing to efficiently pilot several UAVs to best serve all the users of a zone, while addressing the problems of radio and computing resources allocation. The non-convex optimization problem is divided into two sub-problems iteratively treated by SCA and quadratic approximation. The results obtained demonstrate the relevance of the proposed framework, reducing by 5 the total energy consumption of the network compared to OMA schemes. In [59], the authors proposed a unique flying base station (UAV) to best serve all terminals during their offloading phases and consider the time allocation problem along different user cluster. A global energy minimization problem was formulated, including the local energy consumption on each mobile users and the UAV energy consumption. As the previous papers, the NLP formulated problem is subdivided into 2 subproblems and resolved using iterative algorithm and SCA to find the optimal UAV trajectory and resource allocation. The results show that the proposed scheme performs better results up to 33% in energy consumption minimization as compared to OMA and equal resource allocation schemes.

IoT is another possible use case. In [60], the authors considered an **IoT (smart camera processing video**

streaming) that has several tasks to process in parallel, each with specific time constraints. These tasks must be offloaded to several MEC servers, depending on their load level. The formulated problem aimed at minimizing the total energy consumed by the IoT while respecting the processing time limit of each task. Two steps were defined to solve this problem: 1) by fixing the tasks' allocation to the MEC servers, authors defined local computation-rate allocation and NOMA transmission times to the MEC servers by implementing a layered algorithm and 2) by defining the optimal allocation of tasks to the different MEC servers as an optimal ordering problem they solved it with an index-swapping algorithm. The proposed scheme outperforms the NOMA-Heuristic scheme and the FDMA scheme in terms of energy minimization of the IoT.

The **massive IoT use case** was also considered in [61]: the minimization of the energy consumed by many IoTs to fully offload their computational tasks to several MEC servers is studied. The proposed model takes into account both intra-cell and inter-cell interference, which enables a more realistic evaluation of the impact of NOMA. The authors of this paper proposed to manage, in an energy-efficient way, the allocation of computational resources of the servers, as well as the communication resources (allocation of NOMA transmission power and association with the subchannels). The formulated problem is of MINLP nature, and is subdivided into two sub-problems to address the optimal management of the previously mentioned decision variables, handled by sequential convex programming algorithms and Knapsack method respectively. The NOMA scheme proposed by the authors outperforms the others in terms of fair distribution of network resources among its users thanks to the equitable and more reasonable power allocation method.

Industry 4.0 is a third potential use case. In this context, the authors of [62] proposed to optimize the energy consumed by an IoT with several workloads of tasks to be offloaded on different MEC servers using NOMA-downlink communications, while respecting the processing time constraints of each of these tasks. To do this, they defined two types of approaches depending on the channel model. For a static model, the use of a distributed algorithm addresses the non-convexity of the problem formulated. DRL can be used in a time-varying channel context to efficiently define the offloading policy, the NOMA power allocation, and the allocation of local computational resources and at the MEC servers level. This second approach allows to evaluate the solution proposed by the authors in a real-life context. Moreover, the policies obtained using the DRL show significant performances, with a relative error with respect to the optimal solution not exceeding 3%, which validates the approach.

Content caching can be seen as another use case: storing in MEC servers content that will be used in the near future by other users will reduce the number of exchanges and the computational overhead while limiting latency and energy consumption. This is why the authors of [63] have proposed a framework based on DRL, allowing them to efficiently manage content caching (CDNs), policy offloading and computation resource allocation in order to minimize energy consumption over the long term. The authors argue that optimizing content caching is a matter of estimating and identifying task popularity to respond to user dynamic requests. Using an LSTM algorithm and collected time series data, the authors accurately predict task popularity by identifying traffics patterns at different timescales. Finally, an SAQ-learning algorithm and a BLA are used to address the problems of computing resources and policy offloading respectively. The results show that the framework proposed by the authors outperforms the classical approaches using MAQ-learning algorithm. Nevertheless, in the proposed scenario, the results tend to show that the increase in computational resources of the MEC server has more impact in reducing the energy consumption of the network in the long-term compared to approaches with content caching.

This **content caching** issue was also considered by the authors of [64] in a specific context: **multiple radio cells each with a MEC server**. Service caching is used to reduce the load on the backhaul network and the cloud server. The authors take into account the popularity of the services, their size and the caching capabilities available on the MEC server and propose a DRL strategy to maximize service caching. The algorithm based on the DDPG DRL is also used to obtain the policy offloading, and the allocation of computational resources. The proposed approach surpasses the baseline algorithms in terms of energy consumption and QoS, according to the results. In particular, the suggested method saves up to 30% more energy than the baseline algorithms while still achieving users QoS criteria and a high hit rate for service requests.

Beyond problems, models and use cases, **specific architectures** can also be considered. For example, the authors of [65] proposed to optimize the energy consumption of all users of a heterogeneous network consisting of a **Macro Base Station (MBS)** with a MEC server and several **Small Base Stations (SBS)** also equipped with servers. Users can compute tasks locally or remotely. Tasks can be offloaded to the server of the macro base station via OMA communications, or to the SBSs by superimposing themselves on the sub-channel of the macro cell via NOMA. In this context, they formulated a joint problem in which they manage the allocation of radio resources and the offloading policy with the aim of minimizing the energy consumed by the users while respecting the quality of service required by each user task. This division into two sub-problems to deal with the original MINLP problem allows, via an iterative algorithm, to find the optimal solutions concerning the task offloading policy, transmission powers, subchannel resource allocation, the management of the CPU frequencies of each user for the local processing and the allocation of the computing resources of the MEC servers. The authors demonstrate that their scheme performs near optimal performance with a low complexity. The results obtained with tasks having a delay tolerance of

TABLE 2. Comparison of state-of-the-art solutions focusing on energy minimization.

Proposition	Use Case/ technology	Approach	Algorithm	Offloading type	RAT
[49]	Classical	Off. Policy, SIC ordering Transmission rate allocation NOMA power allocation	Dual Lagrange method	Partial	Single cell PD-NOMA Uplink
[50]	Classical	NOMA power allocation NOMA-TPA	Geometric programming KKT	Binary	Single cell Hybrid-NOMA Uplink
[51]	Classical	Hybrid-TPA NOMA power allocation	Closed-form expressions SCA KKT	Binary	Single cell Hybrid-NOMA Uplink
[52]	Classical	User grouping NOMA power transmission RRA remote CRA	Heuristic	Binary	Single cell PD-NOMA Uplink
[53]	IoT LTE-DRX	User schedluing NOMA power allocation	Lyapunov functions	Partial	Single cell PD-NOMA Downlink
[54]	Classical	NOMA-TPA Offloading policy NOMA power allocation	SCA	Partial	Single cell PD-NOMA Uplink/Downlink
[55]	Cooperative NOMA	NOMA power transmission NOMA power transmission NOMA-TPA	BCD KKT	Partial	Single cell PD-NOMA Uplink
[56]	Smart City	Offloading Policy remote CRA	Stochastic optimization Problem Lyapunov functions	Partial	Single cell PD-NOMA Uplink
[57]	Dual connectivity	NOMA power allocation Offloading Policy	Reinforcement learning Heuristic algorithm	Binary	Single cell PD-NOMA Uplink
[58]	Flying Base Station	RRA UAV trajectory Offloading Policy	Iterative algorithm SCA	Partial	Multiple cell PD-NOMA Uplink
[59]	Flying Base Station	NOMA-TPA Offloading policy UAV trajectory	Iterative algorithm SCA	Partial	Single cell PD-NOMA Uplink
[60]	IoT	Off. policy, NOMA-TPA Local CRA	Layered algorithm TopLS algo BotforDual algo Index swapping	Partial	Single cell multiple BS PD-NOMA Uplink
[61]	IoT	NOMA power allocation RRA Local & remote CRA	MINLP SCP Knapsack method	Binary	Multi cell PD-NOMA Uplink
[62]	Industry 4.0	Offloading policy NOMA power transmission Local & remote CRA	Distributed algorithm DRL	Partial	Single cell multiple BS PD-NOMA Uplink
[63]	Content caching	Task popularity prediction Offloading policy remote CRA	DRL with LSTM SAQ-learning BLA	Partial	Single cell PD-NOMA Uplink
[64]	Content caching	Service caching remote CRA Offloading Policy	DRL DDPG	Partial	Multiple cell PD-NOMA Uplink
[65]	Hetnets	Offloading policy NOMA power transmission RRA Local & remote CRA	MINLP Itterative algorithm SCA	Partial	Hetnets PD-NOMA Uplink
[66]	Hetnets mmWave Cooperative NOMA	MEC server selection ON-OFF switching RRA Local & remote CRA	Dual decomposition SCA	Binary	Hetnets MBS & SBS PD-NOMA & PD-SCMA Uplink
[67]	WPT Energy harvesting	NOMA power allocation Harvesting time duration Power beacon allocation Local & remote CRA	Stochastic optimization Problem Lyapunov functions	Partial	Single cell PD-NOMA Uplink
[68]	IRS	IRS phase shifts, Off. Policy Transmission rate allocation NOMA power allocation	Dual decomposition method Penalty method BCD	Partial	Single cell PD-NOMA Uplink

100ms, also demonstrate the effectiveness of the authors' approach in minimizing further the global energy consumed by the users compared to schemes based on OMA, binary offloading and local computing.

Similarly, the authors of [66] considered a **heterogeneous network** in which there are several **small base stations (SBS)** within a **macro base station (MBS)** to serve a high density of users. Each SBS acts as a relay to send information back to the MBS. The energy considered corresponds both to the energy consumption within each SBS and at the backhaul level. The authors formulated an optimization problem in which they wish to minimize the overall energy consumed in the system

while guaranteeing a certain fairness between each actor. The non-convex problem defined aims to handle the selection of users matching the right SBSs, the allocation of subcarriers, and the allocation of computational resources. It was solved using dual decomposition and ACS respectively. To further increase the minimization of energy consumption, they also propose to determine the ON/OFF transitions of small cells with limited computational complexity. Using this method allows for dynamic energy saving and reduced interference with other SBS, which, at the same time, improves and reduces the transmission power of other SBS. The results indicate that for a very large number of users (about 1000), the authors' proposal outperforms other schemes under different traffic models while guaranteeing the service level agreement. Moreover, it seems that the use of PD-NOMA compared to PD-SCMA communication seems preferable in terms of energy efficiency and computational complexity as the number of available subcarriers increases.

Architectural solutions can also integrate **new technologies** such as **wireless energy harvesting technologies**. For example, in [67], the authors aim to minimize the overall energy consumption of the system, by maximizing the ratio between offloaded bits and energy consumption using wireless energy harvesting technologies. A nonconvex fractional programming problem was formulated and solved using a Dinkelbach-based algorithm to determine: 1)the computational frequencies and execution times of the server and user tasks, 2) the Uplink transmission powers, 3) the harvesting times for the terminals and 4) the power transmitted by the power beacon. The results show that the energy consumed decreases significantly when the computational frequencies on the MEC server and user sides are reduced. Furthermore, the total number of task bits offloaded should be equal to the maximum number of computational bits for the MEC server during the task execution stage, while the MEC server and users use the maximum allotted time to complete the tasks, so as to minimize the energy consumed by the system.

Reconfigurable Intelligent Surfaces (RIS) are another solution that has already been considered in NOMA-MEC architectures. The authors of [68] describe an environment in which the direct link between the users and the base station where the MEC server is located is hindered by obstacles, thus providing poor conditions for offloading information. To overcome this, the user signals are sent to a RIS which passively relays to the base station the tasks to be offloaded. The authors then formulated a problem of minimizing the overall energy consumed in such a system, by first managing the passive reflecting beamforming of the RIS (phase shifts) to transmit these uplink signals in the best possible conditions. Each UE is assigned an offloading policy, a transmission rate, a transmission power and a transmission time allocation. SIC ordering management is proposed to optimize the decoding phase at the base station. The authors also proposed an iterative algorithm that treats the problem in two sub-problems using the dual decomposition method for the first one (phase shifts and SIC ordering), then the Penalty method for the second one which subdivides the problem into several convex problems to obtain the sub-optimal solutions. The use of a BCD is then applied to address in continuity the treatment of these two sub-problems. The results show that the use of NOMA is beneficial compared to the approach using TDMA communications.

B. DELAY MINIMIZATION

Delay minimization can be another objective associated with the application of the NOMA in MEC architectures (cf. Table 3. The main idea of such papers is to minimize the time required to process the tasks requested by the end users (delay). This may be necessary for critical applications with high latency constraints: UAV, connected vehicles for example.

Similarly to the work related to energy minimization, the **first studies** on delay minimization were based on **binary offloading models**. For example, the authors of [69] proposed a new NOMA resource allocation scheme for users who want to fully offload their computational tasks to a MEC server to minimize the overall delay (task transmission time and task processing time). Due to the complexity of the problem formulated as a MINLP, they divided it into 3 sub-problems tackled using a heuristic algorithm called NCORA that defines in a greedy manner the allocation of radio resources (power and sub-carriers) and the allocation of computational resources of the MEC server (CPU frequency). The experiments were conducted in a scenario where 30 users are seeking to offload their computational tasks. The results show that NCORA performs much better than other schemes, in particular the NOMA-NCORA version compared to OMA-based schemes. Nevertheless, this evaluation is insufficient to validate the authors' hypothesis that NCORA is a scalable algorithm able to handle a large number of users.

Reference [70] is also one of the first papers that aimed to minimize task execution time by managing task offloading policies and NOMA transmission powers. However, as many papers focusing on **partial offloading**, the authors assumed that their users have the possibility to fragment their tasks into several fragments of any size (processed locally or remotely). In real scenarios this would not be possible as only specific partitions would be possible. Due to the particular structure of the problem, the authors were able to formulate it in a convex manner and to apply a bisection search iterative algorithm. Closed-form expressions for the optimal policy offloading and power allocations are studied to reduce the complexity of the proposed algorithm using the Karush-Kuhn-Tucker (KKT) approach when two users want to process data. The evaluations were carried out considering that the computer resources of the MEC server are limited and allowed the authors to demonstrate the relevance of their solution in terms of delay minimization.

More realistic data processing models have been considered in some papers such as [71]. Under real conditions, the

arrival of tasks and their nature is variable. Different types of services imply different processing delays. This is why the authors formulated a problem where they seek to optimize the global task processing delay by the MEC server, given the constraint that these delays are variable and specific to each task. They first evaluated the relationship between the differentiated uploading delay and the co-channel interference between NOMA users. This allowed them to detect the close link between interference management and overall system latency, and thus to define the optimal power allocation. In particular, they noticed that when the completion order of uploading tasks is not consistent with their SIC decoding order, reducing co-channel interference cannot decrease the users' offload times. Secondly, the use of KKT conditions allowed them to find the close-form expression of the MEC server's computational resource allocation. Finally, using semidefinite relaxation approach to determine a lower bound on the average offload delay helped them determine the offloading and user pairing policy. The results showed that the proposed scheme improves the overall processing time of the whole set of tasks.

Some studies have also considered **more complex scenarios** in which data is not transmitted to a single MEC server but to a set of servers. For example, the authors of [72] tried to minimize the processing time of a user's task groups by offloading them to several MEC servers using downlink NOMA transmissions with superposition coding. Due to the transmission power limit of the user and the channel quality of each MEC server, the idea is to define a group of servers that is neither too large nor too small to handle the user's task groups. A group that is too small results in too much data being processed, which would considerably increase the processing time. On the contrary, too many MEC servers increase co-channel interference and therefore transmission power. To address the challenges of time varying channels and variable MEC servers' capabilities, the authors defined transmission times and offloading policies. In a static channel scenario, they proposed an algorithm based on the combination of the Golden Section search method and the CQR algorithm to obtain the optimal task offloading policy. When channels are time-varying a DRL-like algorithm was considered to get the optimal transmission time on this channel, coupled with the use of the CQR algorithm to define the task offloading policy. It allowed them to reach performances that converge quickly to find the optimal solution, much more than the CVX (convex) tool and the FDMA schemes.

A similar problem was considered in [73] with a **specific use case**: an **IoT** (camera) aiming to offload tasks to several MEC servers. The authors formulated a joint optimization of the computation resource allocations at the MEC server, the radio resource allocations, and the offloading policy (assigning part of the job to a specific MEC server depending on the availability of its resources). It is solved by implementing a three-layered algorithm to find the optimal offloading solution exploiting the hidden convexity of the problem. A more complex use case with different IoT devices simultaneously offloading their computational tasks to MEC servers was also considered there. An algorithm based on the Nash stability concept that allows the optimal allocation of MEC server clusters to each IoT was defined to solve that. The results showed that the use of NOMA in such a context optimizes the performance compared to OMA communication.

The **IoT use case** was also studied in [74]: the authors placed their work in a **NB-IOT** context with a limited number of resources. To miminize delay (task transmission time and task processing time), they optimized the SIC ordering and the allocation of computing resources (CPU frequency) at the MEC server formulating a job-shop scheduling problem. The results obtained in a RAN containing 30 users validated the authors' intuition. Indeed, using the SIC in a classical way (ascending order of channel conditions) is suboptimal compared to the scheme proposed by the authors. Moreover, the considered approach outperforms the other schemes as the number of users increases, perfectly adapted to IoT.

Combining different technologies can also be a solution to minimize latency in MEC architectures and **Hybrid-NOMA** could be a potential approach. In [75], considered two users aiming to offload tasks to a MEC server. Users switch, depending on an energy threshold, either to simultaneous transmission on the same block resource via NOMA or via OMA communication. This problem formulation is a fractional programming problem and is therefore addressed using Dinkelbach's method and Newton's method. The results showd that Newton's algorithm converges faster than Dinkelbach's algorithm in this environment. Furthermore, the higher the energy constraint, the more the use of a scheme based on pure NOMA communication is preferable to OMA and H-NOMA schemes, to reduce offloading delays in the case of two users sharing a block resource.

Going further in terms of technological solutions, **NOMA** could also be combined with **MIMO**. In this way, massive-MIMO was used in [76] to improve the global task processing delay. By combining the properties of massive MIMO with NOMA, the authors aimed to serve and improve the performance of all users in the cell, in particular those at the cell-edge, which usually benefit from degraded performance during offloading phases. To this end, they proposed to optimize the user pairing policy, the transmission duration (defined by the strong user), the offloading policy and the computation resource allocation (CPU frequency) using KKT and the interior-point method. The results showed that for a large number of antennas the delay minimization performance is much better than the OMA-MIMO schemes while using less energy for the transmission phases. Nevertheless, the proposed algorithm only allows the allocation of two users per NOMA cluster which under-utilizes the NOMA technology.

Similarly, the authors of [77] used **massive-MIMO** and **NOMA** in a specific use case: computationally-heavy and latency-critical IOT tasks. The use of a distributed Alternating

TABLE 3. Comparison of state-of-the-art solutions focusing on delay minimization.

Proposition	Use Case/ technology	Approach	Algorithm	Offloading type	RAT
[69]	Classical	RRA NOMA power transmission remote CRA	MINLP Heuristic (NCORA)	Binary	Single cell PD-NOMA Uplink
[70]	Classical	Offloading policy NOMA power transmission	Bisection search iterative algo KKT	Partial	Single cell PD-NOMA Uplink
[71]	Classical	Offloading policy NOMA power transmission User grouping remote CRA	SDR KKT	Partial	Single cell PD-NOMA Uplink
[72]	Classical	Server MEC grouping Offloading policy NOMA-TPA	CQR DRL	Partial	Multiple cell multiple BS PD-NOMA Uplink
[73]	IoT	CRA Offloading policy RRA Server MEC grouping	Three-layered algorithm	Binary	Single cell multiple BS PD-NOMA Uplink
[74]	IoT	SIC ordering remote CRA	Job-shop scheduling problem	Binary	Single cell PD-NOMA Uplink
[75]	Classical	Communication selection NOMA or OMA	Fractional programming Dinkelback method Newton method	Binary	Single cell Hybrid-NOMA Uplink
[76]	Massive-MIMO	User grouping NOMA-TPA Offloading policy CRA	Interior point method KKT	Partial	Single cell PD-NOMA Uplink
[77]	IoT Massive-MIMO	Offloading policy NOMA power transmission Transmission rate allocation	Fractional programming ADMM	Partial	Single cell PD-NOMA Uplink
[78]	MIMO	NOMA power transmission	Dinchelbach GSVD KKT	Binary	Single cell PD-NOMA Uplink
[79]	mmWave	Beamwidth optimization User scheduling NOMA power transmission	MINLP Decoupled in subproblems	Binary	Single cell PD-NOMA Uplink
[80]	IoT WPT Energy harvesting	Offloading policy WPT-TPA NOMA-TPA	DRL	Binary	Single cell PD-NOMA Uplink

Direction Method of Multipliers (ADMM) approach allowed them to manage offloading policy, transmission powers and offloading rates. The results of the numerical study showed that their method effectively minimizes offloading time and optimizes the energy for an uplink NOMA-MEC network.

On their side, the authors of [78] proposed to combine the advantages of **Hybrid-NOMA** communications and **MIMO** technologies to reduce latency in such systems. To solve the non-convex problem, they proposed to decompose it into two sub-problems in which they first deal with the NOMA communication part and then with the OMA part using the Dinkelbach method, the KKT conditions and the generalized singular value decomposition GSVD method. This allowed them to optimally define the allocation of transmission powers. Moreover, the GSVD allows them to decompose the MIMO channels between users and MEC into SISO channels to simplify the problem. Simulation results demonstrate that their proposed system can achieve better delay performance and lower energy consumption compared to OMA but only in the case of two users.

Finally, new architectures have also been considered to minimize latency. The authors of [79] proposed to define new **mmWave-based** architectures to improve NOMA-MEC offloading environments. The advantage of NOMA scheme-based mmWave is that it significantly improves the accessing efficiency for crowded networks by allowing users to share time and spectrum resources in the same spatial layer with NOMA. They defined a problem in which they minimize the task processing time by optimizing the beamwidth, the user scheduling, and the transmission powers. To solve this problem, the authors proposed an alternative optimization which decouples the MINLP problem into a series of solvable subproblems by iteratively optimizing each variable while fixing the others. They prove that this approach can converge to close-optimum solutions with low complexity and achieve up to 50% reduction in average delay compared to other schemes, such as random access and joint beamforming transmission power schemes.

In [80], a **WPT NOMA-MEC** environment was considered. Multiple users harvest the transferred energy to increase their durability and thus take advantage of more energy to solve the delay minimization problems in these systems. The authors proposed a DRL approach to optimize the policy offloading, the WPT duration, and the transmission duration to the MEC server. By exploiting DRL and optimization techniques, they obtained near-optimal offloading solutions with low computational complexity. The numerical results showed that the proposed algorithm outperforms the benchmark algorithms by up to 50% in terms of total computation delay

reduction which improve the overall system performance. Nevertheless the non consideration of the users' CSI is not realistic in practice, and therefore weakens the proposed solution.

C. JOINT ENERGY AND LATENCY MINIMIZATION

The authors of [81] were the first to demonstrate analytically that **NOMA**, compared to OMA, could offert **significant gains in terms of energy and latency minimization**. They considered two scenarios (uplink and downlink) and proposed a radio resource management and offloading policy that enabled them to show that NOMA outperforms OMA, especially when the signal-to-noise ratio is deferable.

Many studies were then interested in the **joint minimization of the consumed energy and the delay** in NOMA-MEC architectures (cf. Table 4. Most articles define a **new criterion**: the weighted sum of these two metrics. This coefficient is used to determine and quantify the degree of importance of one metric over the other. Usually, this coefficient is equal to 50%, meaning that latency and energy have the same degree of importance in the problem formulation. Depending on the particular use case/scenario, it is also possible to prioritize one metric over the other. This approach differs from the problem formulation described in the previous sections where a single objective was coupled with constraints. In the next section, we use the **JWEL** minimization notation, which stands for Joint Weighted Energy and Latency criterion.

For example, the authors of [82] proposed to **minimize the JWEL** in a **dynamic system**, i.e. an environment with time-varying channels. They defined a type of hybrid NOMA/OMA communication, consisting in allocating resources on several types of subcarriers simultaneously. Some can be overloaded and therefore used by several users, such as NOMA subcarriers, while others can only be allocated to one user. They then defined a non-convex optimization problem and addressed it using a DRL algorithm to tackle the radio resource allocation problem in the hybrid NOMA/OMA system, as well as the offloading policy. The DRL algorithm combines the advantages of actor-critic and Deep Q Learning, resulting in a low-complexity algorithm. The algorithm and the Hybrid NOMA/OMA system can then achieve up to 69% energy and latency reduction in such a system compared to non-task offloading, binary and random offloading using NOMA.

Hybrid-NOMA was also considered in [83]. The authors proposed to **minimize the JWEL** in MEC architectures under the constraints of respecting the **specific rates and delays** of each user's task. The authors' strategy is to use hybrid NOMA communications, in which several users start transmitting their messages simultaneously on the same subcarrier, and when a user finishes its offloading phase, the remaining user finishes offloading its task using a classical OMA communication (without changing subcarrier). Nevertheless, it allows to correct and improve the model proposed in some papers in the literature which do not consider variation over time of the interference between users using the same NOMA subcarrier (unsynchronized task completion). In this context, using the KKT approach, they obtain the closed-form expressions of the optimal power allocation and offloading time for each group of users sharing the same subcarrier. Finally, using these closed-for expressions, they define a matching algorithm that allows to select the different user groups. The solution proposed by the authors shows much better results in terms of joint minimization of energy and latency in such a system compared to OMA schemes. Moreover, the closed-form expressions permit to determine the use cases where it is preferable to use OMA or hybrid NOMA communication. Indeed, in situations with specific and constrained delays, it will be preferable to consider NOMA-hybrid, whereas OMA will be considered when there is no particular delay to respect in order to have higher performances.

While most of the proposed solutions considered PD-NOMA, other NOMA technologies could also be applied. For example, in [84], the authors formulated a MINLP minimization problem of joint latency and energy consumption in an **SCMA-assisted MEC system**. The use of SCMA compared to PD-NOMA necessarily involves the introduction of new variables to manage to efficiently allocate these radio resources. Indeed, it is essential to efficiently distribute the codebooks to each user by taking into account their CSI. However, as the codebooks remains the same during the whole communication, the problem of resource allocation in an SCMA system consists in designing the factor graph matrix, i.e. to which sub-carriers each user is attached. The use of the bidirectional matching method makes it possible to determine the factor graph matrix, which in turn allows the transmission powers to be fixed using the lagrange multiplier and KKT methods. Finally, the problems of allocating the computational resources (MEC servers' and users' CPU management), and the task policy offloading are tackled using the interior point method and a heuristic method respectively. The results show that the use of SCMA in conjunction with the use of transmission powers achieves significantly higher performance than other schemes. Also by evaluating scenarios with different values of the JWEL coefficient, it seems that the more energy minimization is taken into account compared to processing delays, the more the total system cost is minimized for a large number of users.

Beyond that, other papers considered **more complex scenarios**. The authors of [85] proposed to minimize the JWEL in a system based on **cooperative-NOMA**, in which several helpers are available to assist other users with task offloading on several MEC servers. To solve the formulated MINLP problem, the authors propose to optimize user association, resource block (RB) allocation, power allocation, task allocation and computational resource allocation. First, the problem is decoupled into a sub-problem of power allocation, task assignment and computational resource allocation processed using a machine learning algorithm (Incremental principal component analysis - IPCA). Then,

TABLE 4. Comparison of state-of-the-art solutions focusing on joint energy and delay minimization.

Proposition	Use Case/technology	Approach	Algorithm	Offloading type	RAT
[81]	Classical	RRA CRA Offloading policy	analytical study	Partial	Single cell PD-NOMA Uplink
[82]	Classical	RRA Offloading policy	DRL Deep Q-network	Partial	Single cell PD-NOMA Uplink
[83]	Classical	NOMA power transmission Hybrid-TPA User grouping	Closed-form expressions KKT	Binary	Single cell Hybrid-NOMA Uplink
[84]	IoT	RRA CRA Offloading policy NOMA power transmission	MINLP KKT Bidirectional matching method Interior point method Heuristic	Partial	Single cell SCMA Uplink
[85]	Cooperative NOMA	User grouping RRA NOMA power transmission CRA Offloading policy	MINLP Machine learning Four-sided matching algo	Partial	Single cell multiple cell PD-NOMA Uplink
[86]	IoV MIMO	NOMA power transmission Offloading policy	DRL DDPG	Partial	Single cell PD-NOMA Uplink
[87]	Classical	Offloading policy RRA	MINLP Game theory	Partial	Single cell PD-NOMA Uplink
[88]	Classical	Offloading policy NOMA power transmission remote CRA	Stackelberg game	Partial	Single cell PD-NOMA Uplink

with the obtained solution, the original problem is equivalently reformulated as a discrete user association and RB assignment processed with a four-sided UE-RB-helper-server matching algorithm, where a sequential permutation operation is designed. Numerical results presented in the paper show that the proposed four-sided algorithm achieves a near-optimal solution with much lower complexity, achieves the minimum fairness between user equipment (UEs) and outperforms other schemes for JWEL.

Decentralized decision processes could also be used to minimize the JWEL. Some works, such as [86], proposed to extend the use of **DRL**. The classical use of DRL in MEC architectures is based on a centralized approach in which the algorithm is applied at the base station level, which have previously collected all the global information of the network architecture needed. This includes conditions and metrics specific to each MEC user, which necessarily leads to additional delays. In [86], the authors proposed to deal with that thanks to a decentralized DRL for the IoV. It allows each car to make its decision based on its own local observations concerning the transmission power to offload, and the power needed to execute the task locally, which in a way consists in defining an offloading policy. They demonstrated the relevance of such an approach to minimize in the long term the energy consumed and the processing delay of the different tasks, in a highly mobile environment with MIMO-NOMA communications and the stochastic task arrival to depict a real scenario context. Thus, the deep deterministic policy gradient (DDPG) algorithm is adopted to learn the optimal power transmission and local execution allocation decision based on the DRL framework. The results show that the approach proposed by the authors outperforms the other schemes, and the further a car is from the base station, the more it will tend to execute its tasks locally and transmit less data. Nevertheless, this paper, as different studies focusing on this issue, assumes that the resources of the MEC server are infinite, which does not fit with real scenarios. It could be sometimes preferable to compute data locally due to the load of the MEC server.

DRL is not the only solution that could be used to provide a distributed decision process in NOMA-MEC architectures. **Game Theory** is another approach, considered for example in [87]. The authors modeled complex interactions and collaborations between several users wishing to offload tasks to the MEC server, using several NOMA subcarriers and game theory. UEs are then considered here as players who cooperatively form coalitions in order to optimize the network performance (subcarriers are regarded as coalitions). To this end, using a distributed algorithm based on this game model, they solve the MINLP problem by jointly optimizing the policy offloading, corresponding to the execution of all the user's task locally or remotely on the server, and the allocation of subcarriers, to minimize the total computational overhead, corresponding to the JWEL presented in the other papers. The presented algorithm outperforms in terms of energy minimization and system latency, compared to a Full Local/Binary Offloading/OMA Heuristic mode.

Game Theory was also an approach considered in [88]. The context presented in this paper is a two-user scenario of a NOMA-MEC network where a **Stackelberg Game**, in which the users are considered as the leads and the server MEC as a follower, is formulated to minimize energy consumption and execution time, i.e. JWEL. Specifically, the leader aims to minimize the total energy consumption for task offloading and local computing by optimizing the policy offloading (coefficients to determine how many

percent in local/remote the task is processed) and transmit power, while the follower aims to minimize the execution time by optimizing its computation resources. Moreover, closed-form expressions for the optimization variables are derived to obtain the Stackelberg equilibrium solution. The insights gained from considering the Stackelberg equilibrium in solving the formulated problem include a low-complexity solution and improved performance compared to other existing solutions.

D. SECURITY

Beyond energy and latency minimization, existing studies also focus on another important issue: security (cf. Table 5.

One of the first criterion defined to evaluate security in NOMA-assisted MEC network architectures is the **secrecy outage probability**, which quantify the probability that a secure communication link fails due to an inability to maintain the desired level of secrecy. In other words, it is the probability that an eavesdropper can successfully decode a transmitted message despite the use of encryption techniques. This criterion is formulated in this way because it is almost impossible for the base station to determine the attacker's CSI.

In this context, the authors of [89] proposed a **passive defense mechanism**. They presented a framework that allows both the minimization of the energy consumed by the terminals and the reinforcement of the security of the communication with the MEC server. They proposed to manage the offloading policy, the transmission power, the codeword transmission rates and the confidential data rates to prevent an attacker from conducting passive eavesdropping in the radio cell. Two models are then proposed. The first is to minimize the overall energy consumed by all users, under time constraints, and not to interrupt the secrecy of the communication by the attacker. The second is to consider only the minimization of the secrecy outage probability. To solve these models, they derive the different allocation schemes towards optimal solutions in closed form expressions that permit them to achieve performances better than the OMA schemes.

With the same objective of linking both energy minimization and improving communication security to protect against passive eavesdropping, the authors of [90] have defined a new metric to quantify both objectives simultaneously: the **Secrecy Energy Efficiency (SEE)** which is the ratio between the total number of secure computations offloading bits per Joule. This is equivalent to maximizing the number of bits to be offloaded to the MEC server without being intercepted, while minimizing at the same time the energy during the transmission phase. To maximize this new criterion, the authors formulate a model in which they adjust the computation resources, the transmission powers in the uplink phases, and the subchannel allocations. The use of a Dinkelbach-type algorithm allows them to address the problem of radio resource allocation, which lets them get the closed-from expressions of the transmission powers for each subchannel. Finally, the Knapsack algorithm is used to solve the computation resource allocation problem. The results show that the scheme combined with NOMA outperforms the others as the number of users increases, thus validating the use case of massive IOT networks presented by the authors.

To quantify the intrinsic ability of a user to be antieavesdropping, i.e. to make it difficult for a malicious user to listen in, the authors of [91] defined a new metric called **AntiEavesdropping Ability (AEA)**. This is the first paper to consider the eavesdropper's CSI in their approach. The formulation of the fractional programming optimization problem consisting in minimizing the AEA is then solved using an adaptation of an interative algorithm and the Dinkelbach method. This approach allows to solve for 2 users the secrecy rate, the policy offloading (i.e. in this context the number of locally processed bits), and the power allocation for transmission. The formulation of the problem and the results obtained tend to show that it is impossible to satisfy both low transmission latency and increased offloading security for a given amount of power. Thus, the increase in security during offloading phases in this system is achieved at the expense of the degradation of transmission delays. Nevertheless, the larger the energy budget, the better the authors' approach will be in terms of security and delay.

On their side, the authors of [92] considered **both security and latency**. They designed a solution that minimizes the latency of the system to improve its overall performance and security in the presence of an eavesdropper. Indeed, reducing these delays shortens the time window for eavesdroppers to intercept offloaded task, and benefits to the global performance improvement of the NOMA-MEC system. To do so, they considered power allocation, policy offloading, and computational resource allocation, which will be defined through the use of a reinforcement learning algorithm based on **Q-learning**. Furthermore, the proposed algorithm aims to minimize latency while ensuring secure transmission of confidential tasks in NOMA-MEC systems with hybrid SIC decoding. The results show that the proposed approach achieves lower latency and thus higher physical layer security compared to existing approaches. Finally, the use of RL-type algorithms in these architectures outperforms existing algorithms in terms of convergence speed and solution quality.

Generally, papers seeking to improve the security of their communications aim to put in place various solutions to protect themselves, allowing them to conceal their communications, which we will refer to in the rest of this paper as passive defenses. On the other hand, some papers [93] and [94] propose **active defenses**, in which users attack the eavesdropper so that it cannot decipher the communications, thus considering that the best way to defend oneself is to attack. In fact, one of the possible attacks in an environment where an attacker would carry out passive eavesdropping would be to jam the eavesdropper's channel in such a way that it would be impossible for him, to listen to any signal in these frequency bands: this is called jamming. In [93], the

TABLE 5. Comparison of state-of-the-art solutions focusing on security.

Proposition	Use Case/technology	Approach	Algorithm	Offloading type	RAT
[89]	IoT	NOMA power transmission Codeword rates transmission Offloading policy	Closed-form expressions	Partial	Single cell PD-NOMA Uplink
[90]	Massive IoT	NOMA power transmission CRA RRA	MINLP Dinckelbach Knapsack	Partial	Single cell PD-NOMA Uplink
[91]	Classical	Secrecy rate Offloading policy NOMA power transmission	Fractionnal programming Dinkelback Iterative algorithm	Partial	Single cell PD-NOMA Uplink
[92]	Classical	Hybrid SIC NOMA power transmission Offloading policy CRA	DRL Q-learning	Partial	Single cell PD-NOMA Uplink
[93]	Jamming security	User grouping CRA RRA	Nash bargaining game Gale-Shapley theory	Partial	Single cell PD-NOMA Uplink
[94]	Jamming security WPT Energy harvesting	NOMA-TPA NOMA power transmission Offloading policy	Dinckelbach SCA	Partial	Single cell PD-NOMA Uplink
[95]	Jamming security Ground Jammer Flying base station	CRA NOMA power transmission Offloading policy UAV trajectory Technology coefficient allocation	SCA BCD	Partial	Single cell PD-NOMA Uplink
[96]	Blockchain D2D	Blockchain transaction D2D coordination	Cooperative Game theory	Partial	Single cell PD-NOMA Uplink & D2D
[97]	IoV	NOMA power transmission	MINLP KKT Frank and Wold algorithm	Partial	Multiple cell PD-NOMA Uplink

authors propose to define pairs of cooperative users, which join together so that one transmits these data, and the other carries out **artificial jamming** to prevent the eavesdropper from deciphering the different communications. However, the use of jamming leads to energy consumption overheads. They then formulate an optimization problem in which they wish to minimize the total energy consumption of the MEC server users. In order for these users to benefit from this cooperating jamming in a fairness manner, they then formulate a system of cooperation between pairs of users based on the NASH bargaining game. Finally, they applied an algorithm based on Gale-Shapley theory to form the user pairs in an optimal way.

The authors of [94] also used jamming but in a different way. They proposed to use the **wireless power transfer technology**, using the power beacon sent to allow the user to offload his tasks, but also to perform the **jamming** of the channel to prevent **passive eavesdropping**. They then formulate a problem whose objective is to maximize the SEE by adjusting the powers allocations (beacon cells, during offloading, and jamming), time allocations (beacon emission duration and offloading duration) and the offloading policy. The problem formulated in this way is non-convex, and is treated by an algorithm that uses Dinkelbach's technique and SCA to find the optimal decision variables. The proposed scheme presents better performances, even though like the others it tends to decrease as the size of the tasks increases.

More complex scenarios have also been considered, including flying base stations **(UAVs)** to best serve all users in MEC architectures. In this context, the authors of [95] studied security issues. Due to the nature of line-of-sight radio communications, it is easy to carry out eavesdropping attacks. They proposed to set up several active defenses such as a jamming ground station, and passive defenses to protect against a possible UAV attacker doing passive eavesdropping. They formulated an optimization problem in which they attempt to maximize the average security computation capacity, under the constraint of minimizing the energy consumed by the system, minimizing the computational capacity of each actor in this architecture, and avoiding collisions between UAVs. To achieve this, they managed the CPU computation frequency, the transmission power, the offloading policy, and the trajectory of the UAV. In addition, to deal with the coexistence and simultaneous operation of the UAV and the ground jammer, they introduced the varying channel relations coefficient allocation in such a way that the flying base station is not affected by it. The formulated problem is then iteratively solved using SCA and BCD to find the optimal solutions.

To overcome the unavailability of the MEC server, UAVs are not the only solution. Some researchers considered **distributing the calculations to neighboring nodes** with sufficient computing resources. In this context, the authors of [96] proposed a Blockchain-based solution to secure the cooperation between different nodes so that they can cooperate securely and authentify each other. A cooperative game theory is applied to maximize the total sum rate and secrecy capacity. The use of NOMA with Blockchain allows to improve the latency and security of such a system by improving its resilience. Nevertheless, such a proposition does not consider active attacks such as jamming or availability attacks which, despite the use of a Blockchain system, would degrade its performance in terms of latency.

In [97], a specific use case was considered by the authors: an **IoV environment**, i.e. a highly mobile environment in which several vehicles move from radio cell to radio cell. This environment consists of a macrocell which contains several microcells in which the vehicles move, and where an eavesdropper can listen to the communications. Each of these microcells is equipped with a RoadSide Unit (RSU) which relays the offloaded tasks from the vehicles to the base station of the macrocell where the MEC server is located. In this context, the authors aimed at preserving the security of the PHY layer of communications while minimizing the offloading delay of the tasks by adjusting the allocation of transmission power. The underlying model proposed differs from the traditional paper using Jake's model (based on a Rayleigh fading model by summing up sinusoids), using an imperfect CSI and a queuing model. The formulated problem is then of the MINLP type and is solved by using Karush-Kugn-Tucker (KKT) conditions and the Frank-and-Wold algorithm to define the optimal transmission powers. For vehicle speeds up to 150km, the proposed solution is much more efficient than solutions based on OFDMA communications, in terms of processing time while guaranteeing the security of the communications. This demonstrates the overall superiority of using NOMA to improve latency and system security in highly mobile environments.

E. OTHERS

Beyond these issues common to many papers, there are also other relevant ideas that are addressed in a limited number of studies (cf. Table 6). These are highlighted in this section, whether they are objectives, architectures or models.

For example, **specific objectives/criteria** could be considered. The authors of [98] introduced a new criterion for evaluating the performance of offloading in MEC architectures: the **task processing capability** of the system. This metric measures the ratio between the size of the tasks to be offloaded and the delay to process them for each MEC user, which, compared to the notions of delay minimization, introduces a notion of **tradeoff between task size and delay**. They investigated a scenario in a heterogeneous network, where several users offload their computational tasks to several MEC servers. Thus, they formulated a MINLP problem, divided into two sub-problems to deal first with the allocation of radio and computational resources, and then with the task offloading problem. Finally, they implemented a low complexity sub-optimal matching process to deal with the resource allocation, and the Lagrangian multiplier method to cope with the convex transmission power allocation problem. The results show that the use of NOMA and several MEC servers in offloading scenario increases significantly the task processing capability of the system, and at the same time decreases the energy and latency of the system.

Such a criterion was also considered in [99]. To overcome the decreased offloading efficiency for users at the cell edge, the authors of this paper used **cooperative NOMA** technologies to offload their latency-intensive and critical tasks in an optimal manner. They proposed a new framework based on the **cooperation of several helpers** nodes, coordinated in two phases. In the first phase (first timeslot), the user at the cell edge offloads his tasks to several helpers via a NOMA-Downlink communication. During the second phase, the helpers pre-process the task and transmit it to the MEC server via a NOMA-Uplink communication. They then formulate a problem in which they maximize the amount of data transmitted to the MEC server under latency constraints, i.e. the amount of user's executed data in the system (overall system performance). To solve it, they first studied the optimal distances between the different helpers and the user, which maximize the amount of data transmitted, allowing them to define their helper selection. They then use the interior point method to address the optimal times of the two phases (user to helpers, helpers to MEC server), the offloading policy and the transmission powers. The results show up to 60% more performance in terms of task offloading compared to single helper and TDMA based schemes.

A similar idea was considered in [100]. **SCMA-assisted MEC architectures** are mostly evaluated as classical optimization problems. However, SCMA communications involving large numbers of parameters such as codebook and factor matrix design, or codebook distribution, have never been considered in stochastic systems, where the user demand and the propagation channel vary randomly over time. For this reason, the authors of [100] presented a framework in which the long-term maximization of the processing rate of the SCMA-MEC architecture with constraints on the processing time for each task is achieved, by using DRL. The LSTM and DQN algorithm allows them to address the joint problem of policy offloading and the SCMA radio resources allocation (codebook allocation and power distribution). Moreover, each IoT user acts as an agent in the algorithm (since it is able to observe only a part of its environment), which, using the LSTM, predicts the state of the other devices to define the optimal policies and allocations in such a system. The results show that the proposed framework achieves significant improvements (up to 30% higher) than schemes based on OMA communications or those based on random distribution of SCMA radio resources.

On their side, the authors of [101] focused on another objective: **processing rate maximization**. They presented a solution that uses the advantages of DRL to maximize the computation rate of an NOMA multi-carrier MEC system, which means improving the server's long-term capacity to process more and more tasks in a quicker manner. Thus, they designed an online model in which the inputs correspond to the channel gains at time t of every subcarrier. The proposed DRL algorithm then solves for each user the optimal allocation of the subcarriers as well as its remote or local task processing mode, using an optimal offloading policy. The use of replay experiences in the algorithm improves the learning efficiency of the agent. Indeed, by storing and

TABLE 6. Comparison of other state-of-the-art solutions.

eProposition	Use Case/technology	Approach	Algorithm	Offloading type	RAT
[98]	Hetnets	NOMA power transmission RRA Offloading policy	MINLP Matching algorithm Lagrangian multiplier method	Partial	Hetnets PD-NOMA Uplink
[99]	Cooperative NOMA	Helper selection NOMA-TPA NOMA power transmission Offloading policy	Interior point method	Partial	Single cell PD-NOMA Uplink
[100]	IoT	Offloading policy SCMA-RRA NOMA power transmission	DRL LSTM DQN	Partial	Single cell SCMA Uplink
[101]	Global	Offloading policy NOMA-RRA	DRL Offline DNN Online Q-learning	Partial	Single cell PD-NOMA Uplink
[102]	Cognitive radio	RRA CRA Offloading policy	ADMM	Partial	Single cell multiple BS PD-NOMA Uplink
[103]	IoT WPT Energy harvesting Cooperative NOMA	Power beacon allocation	Analytical analysis Closed-form expressions	Binary	Single cell PD-NOMA Uplink
[104]	IoV Network slicing	NOMA power transmission RRA Slice selection User grouping	DRL DQL	Binary	Single cell PD-NOMA Uplink
[105]	IoV	CRA	DRL Deep Q-learning Replay method	partial	Single cell PD-NOMA Uplink

reusing past experiences to feed the DNN, the agent learns from a multitude of diverse experiences. This not only allows it to break the correlation between consecutive experiences, but also to learn from rare and important experiences, which helps in the choice of optimal action policies. The algorithm presented by the authors outperforms those based on TDMA schemes by achieving near-optimal results for maximizing the system's computation rates.

Processing rate maximization was also an objective in [102] but in **another context**: a NOMA-MEC architecture, consisting of several servers MEC, based on **Cognitive Radio (CR)**. In this environment, there are two classes of users, including licensed Primary Users (PUs) with priority and absolute access to radio resources, and Secondary Users (SUs) wishing to access the computing resources of the MEC server. Several small Cognitive Base Stations (CBS) with limited computational resources then allow the SUs to offload their computational tasks while sharing the radio spectrum with the PUs. In this context, the authors formulate an optimization problem whose objective is to maximize and improve the global computational capacity of users (and thus maximize the utility of such a system). To this end, the authors propose to manage the allocation of radio and computational resources, and offloading decision in a distributed manner using an ADMM algorithm (Alternating Direction Method of Multipliers). This algorithm achieves much better performance than the other schemes, but is slightly inferior to the centralized solution. Nevertheless, the efficiency of such a system by using cognitive radio is not very relevant due to the under-utilisation of PUs.

Another critera was also introduced in [103]. The use of power beacons is a relevant solution to extend the operating time of IoT devices, in particular those whose battery is difficult or impossible to change. The use of power beacons in NOMA-MEC architectures seems promising to reduce energy consumption and decrease the outage rate of devices. This is why the authors of [103] proposed a **beacon-assisted NOMA-MEC** environment whose objective is to minimize the outage probability of the system. Moreover, the use of a relay, also supplied by the powers beacons, helps the distant IoTs to offload their computing tasks to the MEC server (cooperative-NOMA). The ability of wireless power transfer from the best power beacon to serve distant IoT devices in uplink phases is assessed by the authors using closed-form expressions of the outage probability. The results show that the use of NOMA provides better performance compared to OMA schemes. In addition, increasing the number of antennas and power beacons greatly improves the performance of this beacon-assisted NOMA-MEC system.

To guarantee a high level of performance, it could also be interesting to consider **new network architectures** promoted by 5G networks. This is the case of the solution proposed in [104], which relies on **Network Slicing** to guarantee a certain level of QoS. This paper focuses on the optimization of resource allocation in MEC-enabled IoV networks based on network slicing. The authors introduced a model-free approach based on DRL to solve the resource allocation problem, which jointly addresses channel and power allocation of NOMA transmissions, slice selection, and vehicle grouping. The selection process consists in choosing between two slices allowing to ensure respectively the reliability and the delay of the communication. The problem is modeled as a single-agent Markov decision process, and a Deep Q-Learning (DQL) algorithm is developed that outperforms other benchmark algorithms based on global and offline decisions. The proposed DQL algorithm is proven robust and effective against various system parameters, including the high-mobility characteristics of IoV networks. However, the

paper does not discuss the implementation of the proposed solution in a real-world scenario, which could be a limitation. Additionally, the paper only considers a single-agent Markov decision process and does not explore multi-agent scenarios which leads to more robust solutions by better decision making.

Finally, the authors of [105] considered a **specific use case**, Internet of Vehicles (**IoV**), taking into account the specific constraints of this environment. Offloading solutions for computing tasks are more difficult due to the diversity of communication quality in the current IoV and the **high-speed mobility** of vehicles. To this end, the authors proposed a computational resource allocation scheme based on a DRL-type algorithm to solve the complex problem of task offloading strategies in the IoV. The scheme takes into account the computational power of the service nodes and the speed of movement of the vehicles as constraints and uses a DQL and experience replay method to solve the mathematical resource allocation model to minimize the total system cost. For a number of 15 users with tasks of the order of 10Kbits, the proposed scheme shows excellent performance in terms of reduced network overhead and task processing time.

V. DISCUSSIONS

The taxonomy defined in section III and the comparison carried out in section IV allowed us to identify particularities of some NOMA-MEC approaches, highlighting both good practices and limitations. In this section, we discuss the findings of this study from two angles: the formulation of the problem and the advantages of NOMA technology and its combination with other technologies.

A. PROBLEM FORMULATION

From the problem formulation point of view, the following points seem important:

- **NOMA was one of the candidates studied for being integrated into the 5G standards**. However, due to its implementation complexity, other technologies such as MU-MIMO have been preferred. Due to its non-integration in a more global network, and the compartmentalization of research in telecommunication networks (research being conducted at three levels which are the application layers, the network layers, and the physical layers), this technology found its way into the hands of researchers in signal processing, limiting its study and integration into more global networks. Indeed, apart from a few papers, such as [106], [107], [108], and [109], which focused on linking the network layers and the NOMA physical layers, there is no real framework for applying it in a global network. This explains the fact that the majority of papers on NOMA-assisted MEC architectures rely only on simplified mathematical models to represent the communication system. NOMA is integrated in the formulation of the optimization problem using the formula for calculating the throughput of a user on a sub-carrier. This lack of realism leads to several problems. On the one hand, when a user offloads a task, the whole channel access aspect is abstracted, i.e. the whole communication chain of transforming information (tasks) into radio symbols and then signal is omitted, so that the user offloads in a linear and uninterrupted way his tasks over the same duration. This suggests that the channel is invariant over the offloading period, which seems unrealistic for a large task that will take several seconds to be offloaded with the same channel gain. Moreover, although some authors have introduced Hybrid-NOMA communications (NOMA then OMA communication), a large part of the papers do not modify the value of the inter interference between NOMA users in the throughput formula when one of the users completes its offload, which is very limiting to really judge the performance of NOMA compared to other digital communications. Moreover, this abstraction induced by the implementation of such a model simplifies the use of this technology by minimizing the additional costs generated by its control and management. Indeed, during the uplink phases (offloading of users to the MEC server), neither the allocation of radio resources (transmission power, selection of the sub-carrier, etc.) by the base station, nor the transmission from the users to the base station of the metrics on which the calculation of the optimal policies is based are considered, which greatly underestimates the impact of the complexity of such a system on the performance of NOMA communications;
- By extending the NOMA communication model (throughput formula) to offloading and network models, **the MINLP type optimization problem formulation has been one of the most considered approaches since the emergence of this research area**. This is due to the nature of the decision variables employed, such as user allocation (discrete decision variables), and continuous non-linear resource allocation, as in the power allocation problem which includes non-convex constraints on power and NOMA interference. This formulation leads to the solution of an optimization problem by setting up algorithms, which establishes the optimal solutions and policies to adopt. However, these optimal strategies do not take into account the stochasticity of a real system, alternating between periods of high demand and periods of low demand, making it impossible to guarantee a long-term objective. This is why the recent implementation of solutions such as DRL is a major evolution for the research field, offering new properties of flexibility, scalability, and management of complex and dynamic environments, thus guaranteeing long-term performance, regardless of the underlying scenario. Being model-free learning, it avoids relying on explicit modelling of the environment by learning directly from observations, thus avoiding the need to formulate an accurate and tractable mathematical model, due to the nonlineartity and complexity of the problem;

- The **majority of the papers** are also criticized for **considering that** the BS supports all connectivity simultaneously, i.e. that **all users offload their tasks at the same time**. This lacks realism because the model (throughput formulation) does not take into account the degradation of NOMA performance induced by a large number of simultaneous users, which then increases the complexity of algorithms such as the SIC. On the one hand, the latter will not be able to ensure the same performance for a large number of users due to the increase in inter-symbol interference, and on the other hand, it will increase the energy and delay consumption to carry out the demodulation. Only papers with online models or those considering user scheduling have considered the connectivity limit of the BS;
- Finally, in digital communications, **knowing the CSI of a radio link is fundamental**. It describes the way in which the transmitted signal will be distorted/attenuated during transmission, due to numerous effects such as scattering, fading and power decay. It is therefore essential to know this information to adapt the parameters of its digital communication to ensure that the transmitted signal can be decoded at the receiver. Two approaches are usually considered: the short term CSI based on the estimation of the channel impulse response, and the long term CSI based on a static characterization of the channel using some distribution models such as: Rayleigh model, Rice model. However, the acquisition of short-term CSI can be difficult in fast fading systems, where the channel conditions vary very rapidly compared to the estimation time required. In a security context, where an eavesdropper would conduct passive eavesdropping, knowing the channel properties of the eavesdropper also seems to be a fundamental issue because of the broadcast nature of radio signals. Indeed, knowing this information would allow the different users to adapt their communications so that once the transmitted signal is filtered by the eavesdropper's channel, it is impossible for it to decode the information. The first approaches considered were to assume the CSI of the eavesdropper which is a limited and unrealistic assumption, while others considered the definition of new metrics such as the AEA to take into account the CSI.

B. NOMA, MEC AND LINKED TECHNOLOGIES LIMITS/BENEFITS

From the point of view of NOMA technology, the following points seem important:

- Despite the shortcomings of the models proposed by the simplified handling of NOMA and the lack of realism of the solutions, **NOMA seems to offer much better performance than other OMA schemes**. NOMA could be the optimal solution to maximize spectral efficiency, improving user fairness by using power weighting to distribute network resources more equitably, reducing latency by eliminating queuing and channel access times, increasing connectivity, and improving energy efficiency by minimizing energy consumption. Like 5G networks which are polymorphic systems aiming to ensure a few KPIs in a global set (slice URLLC, EMBB, UMTC), NOMA does not provide all its performances simultaneously. Indeed, the minimization/maximization of one criterion leads to the degradation of the others. For example, the reduction of latency is often done at the expense of more energy consumption. This is why the authors define constraints such as a limited energy budget in delay optimization problems, otherwise NOMA would provide very low latencies at the expense of huge energy consumption;
- In some use cases, **NOMA seems to reveal some limitations compared to other schemes**. NOMA offers ideal performance conditions if and only if the resource allocation is optimally managed. This digital communication is particularly sensitive to the quality of the transmission channel. Indeed, superimposing several users with bad gains on the same radio resource presents less benefits compared to the use of a classical orthogonal scheme, because of the difficulty of demodulating the signals during the SIC. A common use case is for users located at the cell edge, i.e. users with a low signal-to-noise ratio. NOMA is non-energy efficient in this case where a user is forced to offload all its tasks onto a radio resource occupied by this edge cell user;
- **The superposition of users**, generating intra-cell interference, **could also constitute a real bottleneck limiting the performance of NOMA-MEC systems**. This limitation is however overcome by the implementation of optimal power allocation, user scheduling and grouping policies. Furthermore, although the experiments were conducted with a small number of users (fifteen on average), NOMA is still more profitable than the other schemes as the number of users increases, even though this complicates the resource allocation control;
- **Other resources than radio resources are considered in NOMA-assisted MEC architectures**. The main goal of such an architecture is to provide a supplement of computational resources to the users. This is achieved by defining an offloading policy which consists in segmenting a task, part of which will be executed locally and the other part remotely on the MEC server. Once again, due to the nature of the formulation of the optimization problem, the definition of this policy, as well as its application, appears to be unrealistic. On the one hand, the model does not take into account the multiple exchanges allowing the metrics on which the decisions are based to be centralized, and on the other hand, the segmentation of the task, i.e. the number of bits to be offloaded, is carried out under the constraint of the available resources without taking into account the

nature of the task (the task is split in two at any point, which is not coherent);

- **MEC architectures consist in placing servers at the edge of the network**. These servers have limited resources compared to the pooling of several servers in the cloud. Thus, it can be seen that the more computing resources are available on the MEC server side, the more the energy consumed by each user decreases. This is because users can offload much more data to the MEC server, and thus limit the execution and calculation of energy-intensive local tasks. To fit a real environment, it is therefore essential to limit the number of resources available on the MEC server in the problem formulation, otherwise users will tend to only offload their tasks. Moreover, with this constraint, it is sometimes preferable to execute a task locally rather than offloading it onto an overloaded server, when the aim is to reduce task processing times;
- One of the other major roles of **MEC architectures** is to **provide storage resources to network users**, which has been **little considered in the field of NOMA-assisted MEC**, apart from a few papers on content caching. However, the benefits of such a solution may be limited due to the popularity of content, the storage capacity of MEC servers and the caching policy. Indeed, if the stored content is not very accessible, this solution does not present major benefits. Finally, managing these large amounts of data also becomes resource intensive, which for small MEC servers can degrade their performance. The current state of the art does not therefore allow us to observe any real advantages in exploiting storage resources to significantly improve these architectures in terms of energy and delay minimization for the use cases assessed;
- **Other approaches could be relevant for the improvement of NOMA-MEC systems**, in particular **technologies combined with NOMA**. Some solutions coupled with NOMA-assisted MEC architectures improve some performances already provided by NOMA, while accentuating the degradation of others (polymorphic system). This is particularly the case for active systems such as the use of flying base stations or the cooperative NOMA, which introduce new elements degrading some performances that are expected by such systems (like the energy consumption of UAV base stations). On the other hand, the incorporation of passive solutions such as IRS limits this impact, but nevertheless retains a certain complexity of implementation.

VI. PERSPECTIVES

After comparing a large part of the literature on NOMA-assisted MEC architectures, we were able to conduct a discussion to identify the good practices and limitations of all the approaches considered. Nevertheless, some aspects still unexplored to date are revealing themselves as real challenges to optimize these environments. Although many perspectives have been identified so far in MEC surveys, our positioning focuses NOMA-assisted MEC architectures as a monolithic block. Thus, our perspectives concern improvements and proposals specific to this environment, i.e. they engage NOMA and the higher layers of the MEC in equal proportions.

A. DOWNLINK AND FULL DUPLEX TRAFFIC

Initially, the main feature of NOMA-assisted MEC architectures consists in having several users, offloading their tasks on the MEC server in order to benefit from additional computing resources to process them, which is called the offloading phase. So far, the authors have mainly made proposals around the uplink network side. Only [53] considered a strictly downlink context. Other papers such as [54] use the term downlink to refer to NOMA-downlink, which is from a network point of view an uplink communication to several MEC servers. In this context, NOMA-assisted MEC architectures will have to be able to simultaneously support several types of new services that require both types of communication, such as 8K streaming, Augmented Reality, or the results of MEC server calculations. The latter are considered negligible in research papers compared to the size of the offloaded tasks, but their consideration will be essential in new use cases such as IoV applications [110]): HD maps, digital twins, etc.. It will therefore be interesting to develop new problem formulations involving both downlink and uplink voices. Moreover, due to the abstract nature of the optimization problems, the coordination and management of resources (i.e. the feedback of information to the BS to decide on the best policies to adopt, then their coordination via the downlink channel) is not yet considered. Therefore, it would be interesting to exploit the uplink and downlink channels to match the real working of these architectures, and thus measure the impact of the additional costs generated. Finally, as the base stations are also used to manage the incoming and outgoing traffic of other types of users (not using the MEC services), it would be interesting to include them in the simulations in order to measure their impact on the performance of the MEC services.

B. TASK DEPENDENCY

In the formulation of the different approaches proposed in the literature, the nature of the tasks is greatly oversimplified. They are represented in a very basic way by a number of bits to be offloaded to the MEC server. Moreover, it is possible, when defining the offloading policies, to segment at any point of the task without worrying about the consistency of the task. This suggests that there is no relationship between the task segments, and that it is a simple file to be transferred. However, the future applications that will need to be handled by MEC servers are significantly more complex than just data. Indeed, they are often made up of indivisible parallel processes (threads). Moreover, some of them have interdependent relationships, i.e. the result of the

calculation of one thread is often necessary to process the other, establishing a precise processing order. This modelling of relations between tasks greatly complexifies offloading policies, as they also require an orchestration of their different segments. Some authors, such as [111], [112], [113], and [114], have studied this type of model in the context of MEC architectures and have shown that it has an impact on the offloading performance. In this context, the use of a NOMA-assisted MEC architecture would make it possible to overcome these limitations. Indeed, a user could unload several interdependent threads on different MEC servers thanks to downlink-NOMA. The optimal offloading policy would then be a trade-off between the performance of each MEC server, the local processing performance, and the return time of the results, in order to optimize the global processing time of the tasks, by limiting the waiting time between each thread according to the available resources. Moreover, it would be possible to limit the overload due to the return of calculation results on the downlink, by sharing them directly via the backhaul. For example, a user could offload several interdependent threads, and whose intermediate results are exchanged via the backhaul. Thus, a range of such approaches could be designed to efficiently handle more complex and realistic task dependency models using NOMA-assisted MEC architectures.

C. OTHER SECURITY ISSUES

Some of the authors of our comparative study were interested in the security of these new NOMA-assisted MEC architectures. They considered the case where an attacker would carry out passive eavesdropping in the RAN, which would compromise the confidentiality of the transmitted information. To prevent this, the authors proposed passive and active defensive measures to guarantee the confidentiality of their communications. However, these approaches remain limited because the challenges of computer security are also based on other equally important criteria such as integrity and availability. This last point seems to be the most important in this type of architecture. Indeed, if the MEC server becomes unavailable, the whole usefulness of this architecture is called into question, as it cannot guarantee access to the surplus of computing resources for its users. In this context, attacks could come from the RAN, by an attacker jamming the entire channel of the radio cell, or from the backhaul by a Denial of Service attack on the MEC server [115]. In both cases, a NOMA-assisted MEC architecture would be more suitable than a MEC architecture. Indeed, thanks to the downlink-NOMA, the user has the possibility to offload several times his computation tasks simultaneously on several MEC servers, so that his task is processed by an available server. This additional redundancy would provide resilience in these architectures, in order to mitigate the increasing attacks. Nevertheless, a reflection on the detection mechanisms of availability attacks combined with preventive measures such as NOMA-downlink should be studied.

D. NEW EVALUATION METHODS/ENVIRONMENTS

As mentioned in the discussion section, all the proposals made around NOMA-assisted MEC architectures are based on simplified mathematical models solved using optimization algorithms. However, these models do not allow for a detailed evaluation of the real performances in these architectures, in particular with the abstraction of the communication model (by the use of a flow formula), limiting the experiments conducted. This is also the case with the abstraction of the MEC server model, which in most cases is relegated to a simple relationship between the size of the task and the frequency of calculation, which does not allow the full benefits of the MEC architecture standardized by ETSI to be exploited (based on multiple orchestrators and virtualization managers). As there are no real connections between NOMA and the upper layers of the network, there is no complete environment for evaluating the performance of NOMA-assisted MEC architectures. Moreover, due to the compartmentalization of networks, many simulators are available today to evaluate each technology independently. To evaluate NOMA as a physical layer, there are many Matlab scripts, or simulators such as 5G VIENNA simulator [116]. On the other hand, MEC architectures can be evaluated in simulators/emulators such as SIMU5G (Omnet++) [117], EdgeCloudSim [118], NS-3 [119], or Emu5GNet [120]. Thus, one of the most interesting prospects for accelerating the research field would be to offer it a new environment for conducting its experiments. This will necessarily involve the development of new simulators, the combination of several, or the integration of NOMA in MEC simulators, which will lead to thinking about how to interface NOMA with the higher layers of a network. This new environment will allow a more accurate evaluation of the approaches considered in the literature until now.

VII. CONCLUSION

The emergence of new services requires the development of new network architectures to guarantee their operation. Multi-Access Edge Computing has been introduced to meet the growing needs of applications in terms of computing capacity. It enables users to offload part of their tasks, limiting the energy consumed by end-users and the applications' processing time. At the same time, the new NOMA digital communications have attracted the interest of researchers for their innovative properties. Defining NOMA-assisted MEC architectures could be a way to overcome numerous limitations (connectivity, spectral efficiency) while improving performance (latency, energy, security).

In this paper we analyze and compare state-of-the-art solutions proposed for NOMA-MEC architectures. To achieve that, we have introduced a taxonomy providing an overview of the field and a framework for the scientific community to develop new approaches around these architectures. We then carried out a comparative study of the papers, grouping them

according to the targeted performance improvement objectives, enabling us to identify in more detail the approaches considered by the authors. Following this, we conducted discussions on the formulation of the models as well as the advantages and limitations observed during experiments with NOMA-assisted MEC architectures. Finally, we have identified a number of perspectives that will help to overcome certain limitations and accelerate research concerning these architectures.

REFERENCES

[1] C. D. Alwis, A. Kalla, Q.-V. Pham, P. Kumar, K. Dev, W.-J. Hwang, and M. Liyanage, "Survey on 6G frontiers: Trends, applications, requirements, technologies and future research," *IEEE Open J. Commun. Soc.*, vol. 2, pp. 836–886, 2021.

[2] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2192–2235, 4th Quart., 2020.

[3] S. A. H. Mohsan, Y. Li, A. V. Shvetsov, J. Varela-Aldás, S. M. Mostafa, and A. Elfikky, "A survey of deep learning based NOMA: State of the art, key aspects, open challenges and future trends," *Sensors*, vol. 23, no. 6, p. 2946, Mar. 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/6/2946

[4] Q.-V. Pham, F. Fang, V. N. Ha, Md. J. Piran, M. Le, L. B. Le, W.-J. Hwang, and Z. Ding, "A survey of multi-access edge computing in 5G and beyond: Fundamentals, technology integration, and state-of-the-art," *IEEE Access*, vol. 8, pp. 116974–117017, 2020.

[5] K. Sadatdiynov, L. Cui, L. Zhang, J. Z. Huang, S. Salloum, and M. S. Mahmud, "A review of optimization methods for computation offloading in edge computing networks," *Digit. Commun. Netw.*, vol. 9, no. 2, pp. 450–461, Apr. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864822000244

[6] C. Feng, P. Han, X. Zhang, B. Yang, Y. Liu, and L. Guo, "Computation offloading in mobile edge computing networks: A survey," *J. Netw. Comput. Appl.*, vol. 202, Jun. 2022, Art. no. 103366. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804522000327

[7] U. Ghafoor, M. Ali, H. Z. Khan, A. M. Siddiqui, and M. Naeem, "NOMA and future 5G & B5G wireless networks: A paradigm," *J. Netw. Comput. Appl.*, vol. 204, Aug. 2022, Art. no. 103413. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804522000728

[8] A. S. Alahmad, H. Kahtan, Y. I. Alzoubi, O. Ali, and A. Jaradat, "Mobile cloud computing models security issues: A systematic review," *J. Netw. Comput. Appl.*, vol. 190, Sep. 2021, Art. no. 103152.

[9] T. H. Noor, S. Zeadally, A. Alfazi, and Q. Z. Sheng, "Mobile cloud computing: Challenges and future research directions," *J. Netw. Comput. Appl.*, vol. 115, pp. 70–85, Aug. 2018.

[10] H. Li, G. Shou, Y. Hu, and Z. Guo, "Mobile edge computing: Progress and challenges," in *Proc. 4th IEEE Int. Conf. Mobile Cloud Comput., Services, Eng. (MobileCloud)*, Mar. 2016, pp. 83–84.

[11] M. De Donno, K. Tange, and N. Dragoni, "Foundations and evolution of modern computing paradigms: Cloud, IoT, edge, and fog," *IEEE Access*, vol. 7, pp. 150936–150948, 2019.

[12] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. IEEE 77th Veh. Technol. Conf. (VTC Spring)*, Jun. 2013, pp. 1–5.

[13] N. Abbas, Y. Zhang, A. Taherkordi, and T. Skeie, "Mobile edge computing: A survey," *IEEE Internet Things J.*, vol. 5, no. 1, pp. 450–465, Feb. 2018.

[14] S. M. R. Islam, N. Avazov, O. A. Dobre, and K. S. Kwak, "Power-domain non-orthogonal multiple access (NOMA) in 5G systems: Potentials and challenges," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 2, pp. 721–742, 2nd Quart., 2017.

[15] F. Giust, X. Costa-Perez, and A. Reznik, "Multi-access edge computing: An overview of ETSI MEC ISG," *IEEE 5G Tech Focus*, vol. 1, no. 4, p. 4, Dec. 2017.

[16] B. Blanco, J. O. Fajardo, I. Giannoulakis, E. Kafetzakis, S. Peng, J. Pérez-Romero, I. Trajkovska, P. S. Khodashenas, L. Goratti, M. Paolino, E. Sfakianakis, F. Liberal, and G. Xilouris, "Technology pillars in the architecture of future 5G mobile networks: NFV, MEC and SDN," *Comput. Standards Interfaces*, vol. 54, pp. 216–228, Nov. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0920548916302446

[17] M. A. Khan, E. Baccour, Z. Chkirbene, A. Erbad, R. Hamila, M. Hamdi, and M. Gabbouj, "A survey on mobile edge computing for video streaming: Opportunities and challenges," *IEEE Access*, vol. 10, pp. 120514–120550, 2022.

[18] S. Safavat, N. N. Sapavath, and D. B. Rawat, "Recent advances in mobile edge computing and content caching," *Digit. Commun. Netw.*, vol. 6, no. 2, pp. 189–194, May 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2352864819300227

[19] W. Duan, J. Gu, M. Wen, G. Zhang, Y. Ji, and S. Mumtaz, "Emerging technologies for 5G-IoV networks: Applications, trends and opportunities," *IEEE Netw.*, vol. 34, no. 5, pp. 283–289, Sep. 2020.

[20] A. Tufail, A. Namoun, A. Alrehaili, and A. Ali, "A survey on 5G enabled multi-access edge computing for smart cities: Issues and future prospects," *Int. J. Comput. Sci. Netw. Secur.*, vol. 21, no. 6, pp. 107–118, 2021.

[21] M. Chen, W. Li, Y. Hao, Y. Qian, and I. Humar, "Edge cognitive computing based smart healthcare system," *Future Gener. Comput. Syst.*, vol. 86, pp. 403–411, Sep. 2018.

[22] Y. Wu, H.-N. Dai, and H. Wang, "Convergence of blockchain and edge computing for secure and scalable IIoT critical infrastructures in industry 4.0," *IEEE Internet Things J.*, vol. 8, no. 4, pp. 2300–2317, Feb. 2021.

[23] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1710–1732, 3rd Quart., 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8327582

[24] R. Yang, F. R. Yu, P. Si, Z. Yang, and Y. Zhang, "Integrated blockchain and edge computing systems: A survey, some research issues and challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 2, pp. 1508–1532, 2nd Quart., 2019.

[25] S. D. A. Shah, M. A. Gregory, and S. Li, "Cloud-native network slicing using software defined networking based multi-access edge computing: A survey," *IEEE Access*, vol. 9, pp. 10903–10924, 2021.

[26] E. Mustafa, J. Shuja, S. K. uz Zaman, A. I. Jehangiri, S. Din, F. Rehman, S. Mustafa, T. Maqsood, and A. N. Khan, "Joint wireless power transfer and task offloading in mobile edge computing: A survey," *Cluster Comput.*, vol. 25, no. 4, pp. 2429–2448, Aug. 2022.

[27] Y. Niu, Y. Li, D. Jin, L. Su, and A. V. Vasilakos, "A survey of millimeter wave (mmWave) communications for 5G: Opportunities and challenges," 2015, *arXiv:1502.07228*.

[28] S. Kutty and D. Sen, "Beamforming for millimeter wave communications: An inclusive survey," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 2, pp. 949–973, 2nd Quart., 2016.

[29] L. Lu, G. Y. Li, A. L. Swindlehurst, A. Ashikhmin, and R. Zhang, "An overview of massive MIMO: Benefits and challenges," *IEEE J. Sel. Topics Signal Process.*, vol. 8, no. 5, pp. 742–758, Oct. 2014. [Online]. Available: http://ieeexplore.ieee.org/document/6798744/

[30] K. Grover, A. Lim, and Q. Yang, "Jamming and anti-jamming techniques in wireless networks: A survey," *Int. J. Ad Hoc Ubiquitous Comput.*, vol. 17, no. 4, p. 197, 2014. [Online]. Available: http://www.inderscience.com/link.php?id=66419

[31] P. Gandotra, R. Kumar Jha, and S. Jain, "A survey on device-to-device (D2D) communication: Architecture and security issues," *J. Netw. Comput. Appl.*, vol. 78, pp. 9–29, Jan. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804516302727

[32] W. S. H. M. W. Ahmad, N. A. M. Radzi, F. S. Samidi, A. Ismail, F. Abdullah, M. Z. Jamaludin, and M. N. Zakaria, "5G technology: Towards dynamic spectrum sharing using cognitive radio networks," *IEEE Access*, vol. 8, pp. 14460–14488, 2020.

[33] M. Liaqat, K. A. Noordin, T. Abdul Latef, and K. Dimyati, "Power-domain non orthogonal multiple access (PD-NOMA) in cooperative networks: An overview," *Wireless Netw.*, vol. 26, no. 1, pp. 181–203, Jan. 2020. http://link.springer.com/10.1007/s11276-018-1807-z

[34] T. Sylla, L. Mendiboure, S. Maaloul, H. Aniss, M. A. Chalouf, and S. Delbruel, "Multi-connectivity for 5G networks and beyond: A survey," *Sensors*, vol. 22, no. 19, p. 7591, Oct. 2022.

[35] K.-H. Lin, H.-H. Liu, K.-H. Hu, A. Huang, and H.-Y. Wei, "A survey on DRX mechanism: Device power saving from LTE and 5G new radio to 6G communication systems," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 156–183, 1st Quart., 2023.

[36] S. Kumar, P. Yadav, M. Kaur, and R. Kumar, "A survey on IRS NOMA integrated communication networks," *Telecommun. Syst.*, vol. 80, no. 2, pp. 277–302, Jun. 2022, doi: 10.1007/s11235-022-00898-y.

[37] S. M. A. Huda and S. Moh, "Survey on computation offloading in UAV-enabled mobile edge computing," *J. Netw. Comput. Appl.*, vol. 201, May 2022, Art. no. 103341. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1084804522000108

[38] P. Porambage, J. Okwuibe, M. Liyanage, M. Ylianttila, and T. Taleb, "Survey on multi-access edge computing for Internet of Things realization," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 4, pp. 2961–2991, 4th Quart., 2018.

[39] F. Guo, F. R. Yu, H. Zhang, X. Li, H. Ji, and V. C. M. Leung, "Enabling massive IoT toward 6G: A comprehensive survey," *IEEE Internet Things J.*, vol. 8, no. 15, pp. 11891–11915, Aug. 2021.

[40] A. Kirimtat, O. Krejcar, A. Kertesz, and M. F. Tasgetiren, "Future trends and current state of smart city concepts: A survey," *IEEE Access*, vol. 8, pp. 86448–86467, 2020.

[41] Y. Lu, "Industry 4.0: A survey on technologies, applications and open research issues," *J. Ind. Inf. Integr.*, vol. 6, pp. 1–10, Jun. 2017. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452414X17300043

[42] L. Mendiboure, M.-A. Chalouf, and F. Krief, "Edge computing based applications in vehicular environments: Comparative study and main issues," *J. Comput. Sci. Technol.*, vol. 34, no. 4, pp. 869–886, Jul. 2019.

[43] M. Alzenad, A. El-Keyi, F. Lagum, and H. Yanikomeroglu, "3D placement of an unmanned aerial vehicle base station (UAV-BS) for energy-efficient maximal coverage," 2017, *arXiv:1705.03415*.

[44] X. Wang, Z. Fei, J. Huang, and H. Yu, "Joint waveform and discrete phase shift design for RIS-assisted integrated sensing and communication system under Cramer–Rao bound constraint," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 1004–1009, Jan. 2022.

[45] M. Chiang, "Geometric programming for communication systems," *Found. Trends Commun. Inf. Theory*, vol. 2, nos. 1–2, pp. 1–154, 2005.

[46] D. Fouskakis and D. Draper, "Stochastic optimization: a review," *Int. Stat. Rev.*, vol. 70, no. 3, pp. 315–349, 2002.

[47] J. Kronqvist, D. E. Bernal, A. Lundell, and I. E. Grossmann, "A review and comparison of solvers for convex MINLP," *Optim. Eng.*, vol. 20, no. 2, pp. 397–455, Jun. 2019.

[48] I. M. Stancu-Minasian, *Fractional Programming: Theory, Methods and Applications*, vol. 409. Berlin, Germany: Springer, 2012.

[49] F. Wang, J. Xu, and Z. Ding, "Optimized multiuser computation offloading with multi-antenna NOMA," in *Proc. IEEE Global Commun. Conf. (GLOBECOM) Workshop*, Singapore, Dec. 2017, pp. 1–7. http://ieeexplore.ieee.org/document/8269088/

[50] Z. Ding, J. Xu, O. A. Dobre, and H. V. Poor, "Joint power and time allocation for NOMA–MEC offloading," *IEEE Trans. Veh. Technol.*, vol. 68, no. 6, pp. 6207–6211, Jun. 2019.

[51] Z. Ding, D. Xu, R. Schober, and H. V. Poor, "Hybrid NOMA offloading in multi-user MEC networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 7, pp. 5377–5391, Jul. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9679390/

[52] A. Kiani and N. Ansari, "Edge computing aware NOMA for 5G networks," *IEEE Internet Things J.*, vol. 5, no. 2, pp. 1299–1306, Apr. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8267072/

[53] D. Zhai, R. Zhang, L. Cai, B. Li, and Y. Jiang, "Energy-efficient user scheduling and power allocation for NOMA-based wireless networks with massive IoT devices," *IEEE Internet Things J.*, vol. 5, no. 3, pp. 1857–1868, Jun. 2018.

[54] Y. Pan, M. Chen, Z. Yang, N. Huang, and M. Shikh-Bahaei, "Energy-efficient NOMA-based mobile edge computing offloading," *IEEE Commun. Lett.*, vol. 23, no. 2, pp. 310–313, Feb. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8543183/

[55] Y. Huang, Y. Liu, and F. Chen, "NOMA-aided mobile edge computing via user cooperation," *IEEE Trans. Commun.*, vol. 68, no. 4, pp. 2221–2235, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8951269/

[56] K. Li, J. Zhao, J. Hu, and Y. Chen, "Dynamic energy efficient task offloading and resource allocation for NOMA-enabled IoT in smart buildings and environment," *Building Environ.*, vol. 226, Dec. 2022, Art. no. 109513. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S0360132322007430

[57] C. Li, H. Wang, and R. Song, "Mobility-aware offloading and resource allocation in NOMA-MEC systems via DC," *IEEE Commun. Lett.*, vol. 26, no. 5, pp. 1091–1095, May 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9721241/

[58] X. Zhang, J. Zhang, J. Xiong, L. Zhou, and J. Wei, "Energy-efficient Multi-UAV-Enabled multiaccess edge computing incorporating NOMA," *IEEE Internet Things J.*, vol. 7, no. 6, pp. 5613–5627, Jun. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9032145/

[59] I. Budhiraja, N. Kumar, S. Tyagi, and S. Tanwar, "Energy consumption minimization scheme for NOMA-based mobile edge computation networks underlaying UAV," *IEEE Syst. J.*, vol. 15, no. 4, pp. 5724–5733, Dec. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9442899/

[60] Y. Wu, B. Shi, L. P. Qian, F. Hou, J. Cai, and X. S. Shen, "Energy-efficient multi-task multi-access computation offloading via NOMA transmission for IoTs," *IEEE Trans. Ind. Informat.*, vol. 16, no. 7, pp. 4811–4822, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8854118/

[61] B. Liu, C. Liu, and M. Peng, "Resource allocation for energy-efficient MEC in NOMA-enabled massive IoT networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 4, pp. 1015–1027, Apr. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9174768/

[62] L. Qian, Y. Wu, F. Jiang, N. Yu, W. Lu, and B. Lin, "NOMA assisted multi-task multi-access mobile edge computing via deep reinforcement learning for industrial Internet of things," *IEEE Trans. Ind. Informat.*, vol. 17, no. 8, pp. 5688–5698, Aug. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9113721/

[63] Z. Yang, Y. Liu, Y. Chen, and N. Al-Dhahir, "Cache-aided NOMA mobile edge computing: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6899–6915, Oct. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9139263/

[64] H. Zhou, Z. Zhang, Y. Wu, M. Dong, and V. C. M. Leung, "Energy efficient joint computation offloading and service caching for mobile edge computing: A deep reinforcement learning approach," *IEEE Trans. Green Commun. Netw.*, vol. 7, no. 2, pp. 950–961, Jun. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9815021/

[65] C. Xu, G. Zheng, and X. Zhao, "Energy-minimization task offloading and resource allocation for mobile edge computing in NOMA heterogeneous networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16001–16016, Dec. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9272879/

[66] A. Mohajer, M. Sam Daliri, A. Mirzaei, A. Ziaeddini, M. Nabipour, and M. Bavaghar, "Heterogeneous computational resource allocation for NOMA: Toward green mobile edge-computing systems," *IEEE Trans. Services Comput.*, vol. 16, no. 2, pp. 1225–1238, Mar. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9806318/

[67] L. Shi, Y. Ye, X. Chu, and G. Lu, "Computation energy efficiency maximization for a NOMA-based WPT-MEC network," *IEEE Internet Things J.*, vol. 8, no. 13, pp. 10731–10744, Jul. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9312671/

[68] Z. Li, M. Chen, Z. Yang, J. Zhao, Y. Wang, J. Shi, and C. Huang, "Energy efficient reconfigurable intelligent surface enabled mobile edge computing networks with NOMA," *IEEE Trans. Cognit. Commun. Netw.*, vol. 7, no. 2, pp. 427–440, Jun. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9386259/

[69] Y. Dai, M. Sheng, J. Liu, N. Cheng, and X. Shen, "Resource allocation for low-latency mobile edge computation offloading in NOMA networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Abu Dhabi, United Arab Emirates, Dec. 2018, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/8647693/

[70] F. Fang, Y. Xu, Z. Ding, C. Shen, M. Peng, and G. K. Karagiannidis, "Optimal resource allocation for delay minimization in NOMA-MEC networks," *IEEE Trans. Commun.*, vol. 68, no. 12, pp. 7867–7881, Dec. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9179779/

[71] M. Sheng, Y. Dai, J. Liu, N. Cheng, X. Shen, and Q. Yang, "Delay-aware computation offloading in NOMA MEC under differentiated uploading delay," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2813–2826, Apr. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8972932/

[72] B. Zhu, K. Chi, J. Liu, K. Yu, and S. Mumtaz, "Efficient offloading for minimizing task computation delay of NOMA-based multiaccess edge computing," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3186–3203, May 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9741761/

[73] L. P. Qian, B. Shi, Y. Wu, B. Sun, and D. H. K. Tsang, "NOMA-enabled mobile edge computing for Internet of Things via joint communication and computation resource allocations," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 718–733, Jan. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8895824/

[74] L. P. Qian, A. Feng, Y. Huang, Y. Wu, B. Ji, and Z. Shi, "Optimal SIC ordering and computation resource allocation in MEC-aware NOMA NB-IoT networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2806–2816, Apr. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8486633/

[75] Z. Ding, D. W. K. Ng, R. Schober, and H. V. Poor, "Delay minimization for NOMA-MEC offloading," *IEEE Signal Process. Lett.*, vol. 25, no. 12, pp. 1875–1879, Dec. 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8492422/

[76] S. S. Yilmaz and B. Özbek, "Massive MIMO-NOMA based MEC in task offloading for delay minimization," *IEEE Access*, vol. 11, pp. 162–170, 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9999655/

[77] M. H. Alharbi, M. Jun, and H. Liu, "A time- and energy-efficient massive MIMO-NOMA MEC offloading technique: A distributed ADMM approach," in *Proc. IEEE Global Commun. Conf.*, Rio de Janeiro, Brazil, Dec. 2022, pp. 1199–1204. [Online]. Available: https://ieeexplore.ieee.org/document/10001104/

[78] Y. Dursun, F. Fang, and Z. Ding, "Hybrid NOMA based MIMO offloading for mobile edge computing in 6G networks," *China Commun.*, vol. 19, no. 10, pp. 12–20, Oct. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9867957/

[79] J. Shi, Y. Zhou, Z. Li, Z. Zhao, Z. Chu, and P. Xiao, "Delay minimization for NOMA-mmW scheme-based MEC offloading," *IEEE Internet Things J.*, vol. 10, no. 3, pp. 2285–2296, Feb. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9893789/

[80] K. Zheng, G. Jiang, X. Liu, K. Chi, X. Yao, and J. Liu, "DRL-based offloading for computation delay minimization in wireless-powered multi-access edge computing," *IEEE Trans. Commun.*, vol. 71, no. 3, pp. 1755–1770, Mar. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10019272/

[81] Z. Ding, P. Fan, and H. V. Poor, "Impact of non-orthogonal multiple access on the offloading of mobile edge computing," *IEEE Trans. Commun.*, vol. 67, no. 1, pp. 375–390, Jan. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8467377/

[82] V. D. Tuong, T. P. Truong, T.-V. Nguyen, W. Noh, and S. Cho, "Partial computation offloading in NOMA-assisted mobile-edge computing systems using deep reinforcement learning," *IEEE Internet Things J.*, vol. 8, no. 17, pp. 13196–13208, Sep. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9373579

[83] J. Zhu, J. Wang, Y. Huang, F. Fang, K. Navaie, and Z. Ding, "Resource allocation for hybrid NOMA MEC offloading," *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4964–4977, Jul. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9079198/

[84] P. Liu, J. Lei, and W. Liu, "An optimization scheme for SCMA-based multi-access edge computing," in *Proc. IEEE 93rd Veh. Technol. Conf. (VTC-Spring)*, Helsinki, Finland, Apr. 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9449082/

[85] M. Ren, L. Yang, H. Jiang, J. Chen, and Y. Zhou, "Energy-delay tradeoff in helper-assisted NOMA-MEC systems: A four-sided matching algorithm," 2023, *arXiv:2301.10624*.

[86] H. Zhu, Q. Wu, X.-J. Wu, Q. Fan, P. Fan, and J. Wang, "Decentralized power allocation for MIMO-NOMA vehicular edge computing based on deep reinforcement learning," *IEEE Internet Things J.*, vol. 9, no. 14, pp. 12770–12782, Jul. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9663149/

[87] Q.-V. Pham, H. T. Nguyen, Z. Han, and W.-J. Hwang, "Coalitional games for computation offloading in NOMA-enabled multi-access edge computing," *IEEE Trans. Veh. Technol.*, vol. 69, no. 2, pp. 1982–1993, Feb. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8917566/

[88] K. Wang, Z. Ding, D. K. C. So, and G. K. Karagiannidis, "Stackelberg game of energy consumption and latency in MEC systems with NOMA," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2191–2206, Apr. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9314080/

[89] W. Wu, F. Zhou, R. Q. Hu, and B. Wang, "Energy-efficient resource allocation for secure NOMA-enabled mobile edge computing networks," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 493–505, Jan. 2020. [Online]. Available: https://ieeexplore.ieee.org/document/8886467/

[90] S. Han, X. Xu, S. Fang, Y. Sun, Y. Cao, X. Tao, and P. Zhang, "Energy efficient secure computation offloading in NOMA-based mMTC networks for IoT," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5674–5690, Jun. 2019. [Online]. Available: https://ieeexplore.ieee.org/document/8666718/

[91] W. Wu, X. Wang, F. Zhou, K.-K. Wong, C. Li, and B. Wang, "Resource allocation for enhancing offloading security in NOMA-enabled MEC networks," *IEEE Syst. J.*, vol. 15, no. 3, pp. 3789–3792, Sep. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9151270/

[92] K. Wang, H. Li, Z. Ding, and P. Xiao, "Reinforcement learning based latency minimization in secure NOMA-MEC systems with hybrid SIC," *IEEE Trans. Wireless Commun.*, vol. 22, no. 1, pp. 408–422, Jan. 2023. [Online]. Available: https://ieeexplore.ieee.org/document/9850428/

[93] Y. Wu, G. Ji, T. Wang, L. Qian, B. Lin, and X. Shen, "Non-orthogonal multiple access assisted secure computation offloading via cooperative jamming," *IEEE Trans. Veh. Technol.*, vol. 71, no. 7, pp. 7751–7768, Jul. 2022.

[94] M. Wu, Q. Song, L. Guo, and I. Lee, "Energy-efficient secure computation offloading in wireless powered mobile edge computing systems," *IEEE Trans. Veh. Technol.*, vol. 72, no. 5, pp. 6907–6912, May 2023. [Online]. Available: https://ieeexplore.ieee.org/document/10015648/

[95] W. Lu, Y. Ding, Y. Gao, Y. Chen, N. Zhao, Z. Ding, and A. Nallanathan, "Secure NOMA-based UAV-MEC network towards a flying eavesdropper," *IEEE Trans. Commun.*, vol. 70, no. 5, pp. 3364–3376, May 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9734047/

[96] R. Gupta, T. Rathod, and S. Tanwar, "Block-D2D: Blockchain-enabled cooperative D2D-assisted fog computing scheme under imperfect CSI," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Montreal, QC, Canada, Jun. 2021, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9473524/

[97] X. Pei, H. Yu, X. Wang, Y. Chen, M. Wen, and Y.-C. Wu, "NOMA-based pervasive edge computing: Secure power allocation for IoV," *IEEE Trans. Ind. Informat.*, vol. 17, no. 7, pp. 5021–5030, Jul. 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9115882/

[98] J. Xue and Y. An, "Joint task offloading and resource allocation for multi-task multi-server NOMA-MEC networks," *IEEE Access*, vol. 9, pp. 16152–16163, 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9316648/

[99] S. S. Yilmaz and B. Özbek, "Multi-helper NOMA for cooperative mobile edge computing," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 7, pp. 9819–9828, Jul. 2022. [Online]. Available: https://ieeexplore.ieee.org/document/9566305/

[100] P. Liu, J. Lei, and W. Liu, "A deep reinforcement learning scheme for SCMA-based edge computing in IoT networks," in *Proc. IEEE Global Commun. Conf.*, Rio de Janeiro, Brazil, Dec. 2022, pp. 5044–5049. [Online]. Available: https://ieeexplore.ieee.org/document/10001088/

[101] M. Nduwayezu, Q.-V. Pham, and W.-J. Hwang, "Online computation offloading in NOMA-based multi-access edge computing: A deep reinforcement learning approach," *IEEE Access*, vol. 8, pp. 99098–99109, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9102308/

[102] F. Jia, H. Zhang, H. Ji, and X. Li, "Distributed resource allocation and computation offloading scheme for cognitive mobile edge computing networks with NOMA," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2018, pp. 553–557. [Online]. Available: https://ieeexplore.ieee.org/document/8641192/

[103] D.-T. Do, M. V. Nguyen, T. N. Nguyen, X. Li, and K. Choi, "Enabling multiple power beacons for uplink of NOMA-enabled mobile edge computing in wirelessly powered IoT," *IEEE Access*, vol. 8, pp. 148892–148905, 2020. [Online]. Available: https://ieeexplore.ieee.org/document/9164962/

[104] Z. Mlika and S. Cherkaoui, "Network slicing with MEC and deep reinforcement learning for the Internet of Vehicles," *IEEE Netw.*, vol. 35, no. 3, pp. 132–138, May 2021. [Online]. Available: https://ieeexplore.ieee.org/document/9318243/
[105] Y. Zhang, M. Zhang, C. Fan, F. Li, and B. Li, "Computing resource allocation scheme of IOV using deep reinforcement learning in edge computing environment," *EURASIP J. Adv. Signal Process.*, vol. 2021, no. 1, p. 33, Dec. 2021. [Online]. Available: https://asp-eurasipjournals.springeropen.com/articles/10.1186/s13634-021-00750-6
[106] E. Balevi, F. T. Al Rabee, and R. D. Gitlin, "ALOHA-NOMA for massive machine-to-machine IoT communication," 2018, *arXiv:1803.09323*.
[107] M. F. Uddin, "Throughput performance of NOMA in WLANs with a CSMA MAC protocol," *Wireless Netw.*, vol. 25, no. 6, pp. 3365–3384, Aug. 2019, doi: 10.1007/s11276-018-1730-3.
[108] J. Montalban, E. Iradier, P. Angueira, O. Seijo, and I. Val, "NOMA-based 802.11n for industrial automation," *IEEE Access*, vol. 8, pp. 168546–168557, 2020.
[109] M. Elkourdi, A. Mazin, E. Balevi, and R. D. Gitlin, "Enabling slotted aloha-NOMA for massive M2M communication in IoT networks," in *Proc. IEEE 19th Wireless Microw. Technol. Conf. (WAMICON)*, Apr. 2018, pp. 1–4.
[110] C. Schwarz and Z. Wang, "The role of digital twins in connected and automated vehicles," *IEEE Intell. Transp. Syst. Mag.*, vol. 14, no. 6, pp. 41–51, Nov. 2022.
[111] J. Yan, S. Bi, Y.-J. Angela Zhang, and M. Tao, "Optimal task offloading and resource allocation in mobile-edge computing with inter-user task dependency," 2019, *arXiv:1810.11199*.
[112] J. Liu, J. Ren, Y. Zhang, X. Peng, Y. Zhang, and Y. Yang, "Efficient dependent task offloading for multiple applications in MEC-cloud system," *IEEE Trans. Mobile Comput.*, vol. 22, no. 4, pp. 2147–2162, Apr. 2023.
[113] W. He, L. Gao, and J. Luo, "A multi-layer offloading framework for dependency-aware tasks in MEC," in *Proc. ICC IEEE Int. Conf. Commun.*, Jun. 2021, pp. 1–6.
[114] X. An, R. Fan, H. Hu, N. Zhang, S. Atapattu, and T. A. Tsiftsis, "Joint task offloading and resource allocation for IoT edge computing with sequential task dependency," 2021, *arXiv:2110.12115*.
[115] S. A. Bhat, I. B. Sofi, and C.-Y. Chi, "Edge computing and its convergence with blockchain in 5G and beyond: Security, challenges, and opportunities," *IEEE Access*, vol. 8, pp. 205340–205373, 2020.
[116] S. Pratschner, B. Tahir, L. Marijanovic, M. Mussbah, K. Kirev, R. Nissel, S. Schwarz, and M. Rupp, "Versatile mobile communications simulation: The Vienna 5G link level simulator," *EURASIP J. Wireless Commun. Netw.*, vol. 2018, no. 1, p. 226, Sep. 2018, doi: 10.1186/s13638-018-1239-6.
[117] G. Nardini, G. Stea, A. Virdis, D. Sabella, and P. Thakkar, "Using Simu5G as a realtime network emulator to test MEC apps in an end-to-end 5G testbed," in *Proc. IEEE 31st Annu. Int. Symp. Pers., Indoor Mobile Radio Commun.*, Aug. 2020, pp. 1–7.
[118] C. Sonmez, A. Ozgovde, and C. Ersoy, "EdgeCloudSim: An environment for performance evaluation of edge computing systems," *Trans. Emerg. Telecommun. Technol.*, vol. 29, no. 11, Nov. 2018. [Online]. Available: https://onlinelibrary.wiley.com/doi/pdf/10.1002/ett.3493
[119] S. Massari, N. Mirizzi, G. Piro, and G. Boggia, "An open-source tool modeling the ETSI-MEC architecture in the industry 4.0 context," in *Proc. 29th Medit. Conf. Control Autom. (MED)*, Jun. 2021, pp. 226–231.
[120] T. Sylla, L. Mendiboure, M. Berbineau, R. Singh, J. Soler, and M. S. Berger, "Emu5gnet: An open-source emulator for 5G software-defined networks," in *Proc. 18th Int. Conf. Wireless Mobile Comput., Netw. Commun. (WiMob)*, 2022, pp. 474–477.

ROMAIN DULOUT received the master's degree in telecommunications engineering from the ENSEIRB-MATMECA Engineering School. He is currently pursuing the Ph.D. degree with the University of Poitiers, under the supervision of Yannis Pousset, Virginie Deniau, and Leo Mendiboure. Since 2022, he has been a member of the XLIM Laboratory. His main research interests include multi-access edge computing, quality of service management, and NOMA communications.

LEO MENDIBOURE (Member, IEEE) is currently a tenured Researcher in computer science with Université Gustave Eiffel. He is also a member of the COSYS/ERENA Team (Bordeaux site). His work has led to several publications in international journals and conferences. His main research interests include cooperative and intelligent transport systems, security of wireless communication networks, and quality of service management.

YANNIS POUSSET was born in 1971. He received the Ph.D. degree in mobile radio communication from the University of Poitiers, in 1998. Since 2012, he has been a Professor with the Department of Electrical Engineering, University of Poitiers. He developed its research activities with the XLIM Laboratory. His research interest includes the study of adaptive links related to the optimal transmission of data over wireless spatio–temporal radio channels.

VIRGINIE DENIAU received the Ph.D. degree in electronics, in 2003. She is currently a Research Fellow with Université Gustave Eiffel. She conducts research in the field of electromagnetic compatibility (EMC) for land transport and its signaling and communication systems. For the past 15 years, the work has also focused on the hardening of terrestrial transport systems against cyber attacks, including intentional EM interference and protocol attacks against communication or signaling systems. She has coordinated several European and national projects. She is also a Coordinator of the DEPOSIA Project, which focuses on the development of AI for geolocating EM attack sources. She is the Chair of the commission ElectroMagnetic Interference of the International Union of Radio Science (URSI).

FREDERIC LAUNAY (Member, IEEE) received the Ph.D. degree in electronics, in 2003. He is currently an Assistant Professor with the LIAS Laboratory and teaches with the Networks and Telecoms Department, University of Poitiers, Poitiers. He developed its research activities with the XLIM Laboratory. His work has led to several publications in international journals and conferences.