IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received October 6, 2020, accepted October 23, 2020, date of publication November 10, 2020, date of current version November 23, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.3037183

On the Opportunities and Challenges of NOMA-Based Fog Radio Access Networks: An Overview

MEGUMI KANEKO[1], (Senior Member, IEEE),
ITSIKIANTSOA RANDRIANANTENAINA[2], (Student Member, IEEE),
HAYSSAM DAHROUJ[3], (Senior Member, IEEE), HESHAM ELSAWY[4], (Senior Member, IEEE),
AND MOHAMED-SLIM ALOUINI[2], (Fellow, IEEE)
[1]National Institute of Informatics (NII), Tokyo 101-8430, Japan
[2]Computer, Electrical and Mathematical Sciences and Engineering, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia
[3]Center of Excellence for NEOM Research, King Abdullah University of Science and Technology (KAUST), Thuwal 23955, Saudi Arabia
[4]Electrical Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Dhahran 31261, Saudi Arabia

Corresponding author: Megumi Kaneko (megkaneko@nii.ac.jp)

This work was supported in part by the King Abdullah University of Science and Technology (KAUST) and the Grant-in-Aid for Scientific Research (KAKENHI) from the Ministry of Education, Science, Sports, and Culture of Japan number 17K06453 and the NII MoU grant, and in part by the Center of Excellence for NEOM Research at KAUST.

ABSTRACT Future generations of wireless networks are expected to provide new services with an unprecedented level of diverse and stringent requirements. Fog Radio Access Network (FRAN) and Non-Orthogonal Multiple Access (NOMA) have emerged as complimentary enablers to meet such requirements. On the one hand, FRAN architecture is designed to reduce the delay caused by the fronthaul link by pushing control and storage to the network edge. On the other hand, in addition to increasing the spectral and energy efficiency and the number of connected devices, NOMA has the potential to improve network latency. This paper overviews the joint benefits of enabling NOMA schemes in an FRAN architecture, by means of examining the applicability and adequateness of the NOMA-based FRAN features in achieving specific objectives of next generation of mobile networks, mainly those related to enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable low latency communication (URLLC). The paper further depicts the challenges and future research directions that must be addressed in order to meet such opportunities.

INDEX TERMS FRAN, CRAN, NOMA, OMA, Beyond 5G, machine learning, Deep Reinforcement Learning.

I. INTRODUCTION

While the commercial deployment of the fifth generation of wireless systems (5G) [1] is at its premises in a number of countries, worldwide research efforts are already turned upon defining Beyond 5G (B5G) systems [2]. Towards achieving the diverse and stringent service requirements envisioned in B5G systems, this overview paper focuses on the potential benefits and technical challenges offered by the integration of Non-Orthogonal Multiple Access (NOMA) technology in the realm of the Fog Radio Access Network (FRAN) architecture. In this introductory section, we first recall two major cloud-based architectures, namely Cloud Radio Access Networks (CRAN) and FRAN, as well as the basics of Orthogonal Multiple Access (OMA) and NOMA technologies. Then, we highlight the features and potential advantages of the advanced NOMA-based FRAN architecture against conventional wireless system architectures.

A. CURRENT TECHNOLOGIES AND THEIR LIMITATIONS

1) CRAN AND FRAN

During the recent years, Cloud Radio Access Networks (CRAN) had been promoted as a promising network architecture to face the upcoming mobile data traffic deluge. In CRAN, the conventional Base Station (BS) functionalities are separated into two components: the cloud center, and a set of Remote Radio Heads (RRH). While the RRHs serve

The associate editor coordinating the review of this manuscript and approving it for publication was S. Sofana Reka.

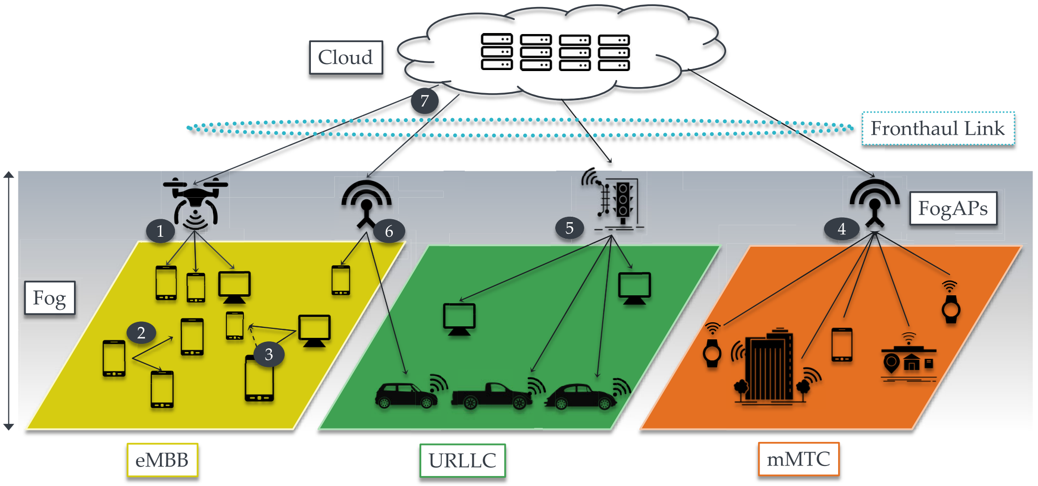


FIGURE 1. Integration of NOMA in FRAN for each 5G use case: eMBB, URLLC and mMTC. The arrows (1, 4, 5, 6) from each Fog Access Point represent users served by NOMA, the arrows (2, 3) indicate NOMA for D2D and user cooperation respectively, while arrow 7 marks NOMA for wireless fronthaul transmissions; see Section II for a detailed description.

as the data plane that grants wireless coverage, the cloud center is a centralized pool of Baseband Units (BBU), composed of processors with cloud computing capabilities. Such powerful pool of processors acts as the control plane that handles large-scale signal processing, Radio Resource Allocation (RRA), and Interference Management (IM) on a network-wide level. In a fully centralized CRAN, each RRH is a low-cost entity solely equipped with minimal functionalities such as Radio Frequency (RF) and Analog/Digital (A/D) conversion. Each RRH forwards to the cloud BBUs the baseband user signals of their associated mobile users, through wired or wireless fronthaul links. In theory, CRAN would be able to globally optimize the RRA and IM problems, as it allows for joint signal processing at the cloud level. However, the resulting performance heavily depends on the availability and accuracy of instantaneous Channel State Information (CSI) of all links between each RRH and user, which would generate excessive amounts of CSI feedback overhead. Furthermore, the CRAN performance depends on the capacity-limited fronthaul links, which often induces inevitable transport delays that are incompatible with the stringent requirements of the prospective B5G systems applications [3].

To circumvent the above design challenges while still benefiting from the powerful cloud processing capabilities, FRANs have recently garnered strong research interests from academia and industry [3]. The FRAN architecture is designed to overcome CRAN latency and signalling overhead issues by partially moving network intelligence, i.e., cloud computing and storage capabilities, closer to the network edge – access points, mobile users and devices – as illustrated in Fig. 1. Hence, in FRAN, the access points – herein called Fog Access Points (FAP) instead of RRHs – can perform distributed signal processing and RRA. The FAPs can also cache the most popular files for their served users. For instance, the benefit of cloud-edge processing in FRANs is highlighted by illustrating the joint improvement of network performance, e.g., throughput, latency, energy-efficiency and user fairness [4], [5].

2) OMA AND NOMA

In parallel to the active development on the network architecture level, advanced access schemes, such as NOMA, are bound to revolutionize the way of sharing and managing radio resources among multiple users [6]. Although NOMA has not been included in the initial 5G New Radio standard [7], it is still gathering keen interests from the research community given the plethora of NOMA-related open issues, as shown later in this paper. Conventionally, each basic radio resource unit in time-frequency-space-code is allocated to a unique user using OMA, as in legacy multiple access schemes such as Orthogonal Frequency Division Multiple Access (OFDMA) in 4G-LTE or IEEE 802.11ay. However, the theoretical suboptimality of OMA schemes, except under specific circumstances, is well established in the literature. For example, the capacity region of degraded broadcast channels is achieved by superimposing multiple user messages within each basic radio resource unit through Superposition Coding (SC) followed by Successive Interference Cancellation (SIC) [8]. Hence, by intelligently allocating multiple user messages to the same basic radio resource unit through SC, large performance gains are achieved compared to OMA schemes in terms of spectral and energy efficiency, connectivity and latency [6].

A large number of methods have been designed so far for implementing the theoretical concept of SC and SIC, such as power-based NOMA and code-based NOMA, whereby each user message sharing the same radio resource unit is differentiated through different power levels or different codewords. Upon signal reception, each user performs SIC, namely, each user sequentially decodes and subtracts other users' signals that are received with higher power than its own, before decoding its own signal while treating the remaining users' signals with lower power as noise [9]. A comprehensive survey on the existing variations of NOMA techniques may be found in [10].

One of the major advantages of the joint SC and SIC operation is its rate-fairness enhancing capabilities. More precisely, allocating low power to the ''strong'' users with high channel qualities and high power to ''weak'' users with low channel qualities, both types of users can still experience substantial data-rate levels simultaneously, albeit utilizing the same radio resource unit. In other words, SC allows to exploit the high channel disparities among users, for enhancing the weak users' rates without incurring much degradation of strong users' rates. More generally, SC schemes, such as NOMA, enable a flexible allocation of resources that can be tailored to various design objectives that involve different forms of fairness.

Despite the obvious advantages brought by NOMA, they heavily rely on the receiver's accurate knowledge of the involved CSIs, as imperfect CSI may incur severe error propagation during SIC decoding. This problem, however, is largely mitigated by superimposing only a few messages within each radio resource unit, which is why most NOMA-based schemes consider only two user messages, thereby hitting a reasonable trade-off between NOMA performance gain and receiver complexity [9]. NOMA's performance level is also contingent upon the decoding order of the superimposed user messages, as well as the network interference management methods. In fact, it has been reported in various works that NOMA is particularly vulnerable to inter-cell interference [11], [12]. Therefore, the true impact of integrating NOMA in future systems would strongly depend on the capability of the global network to address such issues. To this end, this paper examines the applicability and adequateness of NOMA-based FRAN features for realizing the prevailing B5G use cases, mainly those related to enhanced mobile broadband (eMBB), massive machine type communications (mMTC), and ultra-reliable low latency communication (URLLC).

B. THE POTENTIAL OF NOMA-BASED FRAN

Integrating NOMA technology in an FRAN architecture aims at merging and leveraging their mutual benefits. Such an approach has recently been adopted in terms of RRA and IM in [13], [14], and has shown promising features towards meeting the high expectations of future wireless systems in terms of capacity, latency, energy-efficiency, massive device connectivity, and deployment costs. More specifically, the IM algorithm proposed in [13], [15] for allocating radio resources and transmit power shows the remarkable merits of NOMA-based FRAN at providing user rate fairness in terms of Jain's index, without sacrificing global system throughput as compared to OMA-based FRAN. In the same context, power and channel allocation schemes are analyzed in [14], [16], resulting in substantial improvements in the NOMA-based FRAN data throughput. Energy efficiency optimization is investigated in [17], which proposes a matching-based resource allocation solution. Recent works also consider the joint resource allocation and content caching so as to capitalize on the edge-processing capabilities of NOMA-based FRAN [18], [19].

Along those directions, this overview paper provides an outlook of the opportunities and challenges resulting from integrating NOMA technology in a FRAN architecture. Within this framework, the major motivations and contributions of this paper can be listed as follows:

1) In the sequel, the advantages of NOMA-based FRAN architecture are presented in light of B5G systems selected use-cases, by advocating how this integration can help achieve the conflicting objectives of eMBB, mMTC and URLLC,
2) We then illustrate these potential merits and promising insights through numerical evaluations,
3) Lastly, we identify the main challenges and discuss open research directions in the context of NOMA-based FRAN.

II. NOMA-BASED FRAN ENABLING 5G AND BEYOND USE CASES

The International Telecommunication Union - Radiocommunication (ITU-R) sector defines three main classes of use cases for 5G, i.e., eMBB, mMTC and URLLC. In this section, we discuss how NOMA-based FRAN can help meeting or improving the QoS satisfaction level of each use case separately. Other opportunities outside these use cases are also presented. Each use case is characterized by specific Quality of Service (QoS) requirements in terms of peak data rate, latency, reliability and connectivity. These classes are expected to evolve in the context of B5G systems in two major ways: firstly by requiring more stringent Quality of Service (QoS) targets, and secondly, by generating overlapping QoS requirements among different use cases such as mMTC with eMBB-level peak data rate or eMBB with URLLC-level latency. The latter scenario is illustrated for instance by Extended Reality (XR), a new service envisioned for 6G combining augmented, mixed and virtual reality [2].

A. TOWARDS THE REQUIREMENTS FOR eMBB

The first use case, known as eMBB, includes applications that require ultra-high data rate and broad network coverage such as 3D video streaming on 4K screens, virtual reality and online gaming. eMBB peak data rates are expected to be 1000-fold faster compared to 4G.

Thanks to the FRAN architecture, the required application contents can be, in addition to be stored at the cloud, cached in the network edge in the proximity of the requesting users, i.e., either in local FAPs or neighbouring user devices. Compared to interference-limited small cell networks and delay-limited CRAN, the network edge caching feature of

FRAN helps boosting the experienced data rates and drastically reduces the incurring latency. From a medium access control perspective, NOMA further allows each FAP to serve multiple users simultaneously, within each basic Resource Block (RB) unit, as illustrated in Fig. 1-① (i.e., see label ① in Fig. 1). Along the same direction, reference [15] assesses the eMBB gains garnered through the integration of NOMA within an FRAN architecture. More specifically, [15] proposes an optimized RRA and IM method specifically designed for the downlink of a NOMA-based FRAN. The proposed RRA in [15] integrates specific constraints imposed by the FRAN architecture, namely 1) the fronthaul capacity limitation for each FAP, and 2) the local FAP operation whereby each user can be associated to a unique (local) FAP only. Note that, mathematically, this second constraint imposes a zero-norm constraint on the user and power allocation vector variables, which is one of the main technical distinctive features of FRAN optimization problems, as detailed in Section IV. Reference [15] shows that exploiting NOMA in the downlink transmissions of FAPs indeed increases user fairness without sacrificing data rates. Hence, NOMA-based FRAN is an attractive solution to increase the number of served users and devices, while simultaneously satisfying their individual eMBB QoS requirements.

Similarly, in D2D communications, NOMA would enable one user to share its cache content to multiple users simultaneously, as illustrated in Fig. 1-②. For instance, reference [20] shows that, unlike D2D OMA, D2D NOMA can be combined with users cooperation, i.e., instead of discarding the weak user's signal, the strong user sends it to the weak user for channel diversity (Fig. 1-③). As in [21], [22], future studies on the integration of NOMA and D2D communications in FRAN architecture, while taking into consideration the caching capabilities of the Fog edge devices, would indeed be useful for jointly providing eMBB high data rates and low latencies.

B. TOWARDS THE REQUIREMENTS FOR mMTC

Tens of billions of Internet of Things (IoT) devices are expected worldwide in the near future, and so mMTC help realizing such a massive connectivity of machines and sensors. mMTC devices often have low computational capabilities and low energy supplies, are mostly static, and communicate with sporadic and short packets, mainly for control signalling purposes.

Applying NOMA in the uplink transmission is, therefore, particularly suitable for mMTC applications, as the number of connected devices per basic RB unit can be increased, thereby enhancing the network scalability (Fig. 1-④). This is particularly illustrated in reference [23], which proposes a NOMA grant-free access where users transmit data without waiting for APs to grant them radio resources. Such grant-free access reduces the energy consumed in signalling. A comprehensive literature survey on NOMA-based grant free transmissions for mMTC applications can be found in [24], and on NOMA-based massive access in [25]. Compared to grant-free OMA, higher transmission success probabilities are achieved in grant-free NOMA as multiple user messages can share the same radio resource unit without suffering from collisions. Moreover, the FRAN and mobile edge computing (MEC) architectures' hierarchical networking and computing structures from edge to cloud would facilitate the processing of massive amounts of IoT data. Hence, NOMA-based grant-free access schemes in uplink FRAN are a promising, powerful solution for meeting the massive connectivity demands of future mMTC use cases.

C. TOWARDS THE REQUIREMENTS FOR URLLC

The third use case, URLLC, comprises a set of features that are suitable for mission-critical applications, e.g., telemedecine, self-driving vehicles, tactile Internet, etc. Such features are best described through low latency and ultra-high reliability requirements as highlighted in [26]:

- Extremely low latency: The end-to-end latency, i.e., the total delay caused by transmission, queuing, processing and re-transmissions – for URLLC applications is expected to be less than 1ms.
- Ultra-reliable communication: The required Bit Error Rate (BER) levels are in the order of 10^{-5} to 10^{-9} so as to guarantee extremely high packet transmission success. In critical applications, such as telesurgery, such BER requirement would be as low as 10^{-9}.

As mentioned earlier, FRAN supports low-latency real-time communications by carrying out data storage and computation at the network edge. Enabling NOMA-based allocation further reduces latency in both downlink and uplink. In the downlink, NOMA multiplexes multiple user messages on the same basic RB unit. In the uplink, NOMA allows several users to simultaneously access the same basic RB unit without collisions (Fig. 1-⑤). Moreover, cloud-based optimized packet scheduling, RRA and IM can jointly enhance the network reliability metrics of the users.

Despite the opportunities that NOMA-based FRAN has in supporting URLLC, the joint provision of high reliability and low latency is yet to be analyzed. First and foremost, the effect of SIC processing delays, SIC decoding error propagation, unpredictable interference caused by grant-free access, and imperfect CSIs should be analyzed in further details. More precisely, it must be clarified whether the performance gains offered by NOMA-based FRAN outweigh such additional burdens. In the next section, numerical evaluations are presented for illustrating the potential of NOMA-based FRAN in addressing eMBB, mMTC and URLLC requirements.

D. OTHER USE-CASES

The different use cases presented above can also be jointly realized non-orthogonally through the same FAP in NOMA-based FRAN (Fig. 1-⑥). Along this line, references [27] and [28] analyze the performance of a network where eMBB and URLLC coexist under NOMA, for FRAN and CRAN architectures, respectively. The eMBB-URLLC coexistence is possible thanks to their non-contradicting, complimentary requirements. While eMBB users are data-hungry, URLLC users require that their packets are successfully delivered within the required latency. Reference [27] proposes a hybrid CRAN-FRAN architecture where two 5G use cases (eMBB and URLLC) share the radio resource in a non-orthogonal manner. The eMBB users' traffic is processed in the cloud as in CRAN, while the URLLC users' traffic is processed at the edge nodes. Reference [27] further shows that despite the interference between the two use cases, NOMA offers low latency for URLLC and better spectral efficiency for eMBB. Reference [28] shows that serving eMBB and URLLC under NOMA yields a better rate for eMBB, while guaranteeing QoS for URLLC.

We note further that additional analysis is needed at this stage to determine whether it is also possible to serve the massive connectivity requests of mMTC use case in a non-orthogonal manner together with both eMBB and URLLC, without inducing mutual impairments.

Another important aspect to consider is related to the fronthaul links between the cloud BBUs and each FAP, which may cause severe bottlenecks both in terms of data rate and transport delay. In case of wireless fronthaul links, multiple FAPs can be connected to the BBU pool within the same radio resource through NOMA (Fig. 1-⑦), as described in [29] for CRAN. A similar setup can be implemented and analyzed for FRAN. Moreover, cooperation among FAPs can be established by sharing their decoded signals instead of discarding them without impairing the strong latency requirements of the applications handled in FRAN.

III. PROOF-OF-CONCEPT: NUMERICAL EVALUATIONS

This section presents numerical evaluations that illustrate the potential of NOMA-based FRAN in fulfilling the needs of each of the use-cases presented in the previous section, i.e., eMBB, mMTC and URLLC. The presented results are based on the algorithm proposed in [15], where user association, power and RB allocation, as well as NOMA power split optimizations are conducted for maximizing the weighted sum-rate of a downlink NOMA-based FRAN with multiple RBs. As the purpose of this section is to provide fundamental basic insights as to why NOMA-based FRAN would be an enabler of B5G systems, we do not delve into the technical and mathematical details, which can be found in [15]. Nevertheless, let us simply describe the main ideas of the RRA and IM method proposed in [15]. The problem is formulated as a weighted sum-rate maximization subject to FRAN-specific local processing and fronthaul capacity constraints, where the goal is to optimize the user-to-RB assignment, the power allocation to each RB, and the power split levels of the NOMA users jointly served in each RB. To solve this intricate mixed integer non-convex optimization problem, a feasible three-step decoupled approach is designed, which relies on the FRAN-specific cloud/edge computation capabilities. In particular, a Multiple Choice Knapsack (MCKP)-based method is devised for solving the assignment problem, followed by an auction algorithm to enforce the FRAN local processing constraint. The power allocation to RBs and the NOMA power split optimization steps are solved using the alternating direction method of multipliers (ADMM).

In the following, perfect Channel State Information (CSI) knowledge is assumed at each FAP and at the cloud BBUs. The main simulation parameters from [15] are illustrated in Table 1 for clarity. Specifically, 28 users are uniformly distributed over the network area served by 7 FAPs, one for each microcell. As heavy load conditions were of interest, full buffer eMBB traffic was assumed.

TABLE 1. Simulation parameters.

Parameter	Value
Environment	Urban micro-cell
Number of FAPs	7 with wrap around
Number of users	28
NOMA users	2 per resource block

Firstly, Fig. 2 compares NOMA and OMA in terms of the user rate cumulative distribution function (CDF). The figure shows that, in the lower rate region, NOMA performs better than OMA for both low and high fronthaul capacity values. In the higher rate region, NOMA and OMA perform similarly under low fronthaul capacity, while OMA serves more users with a higher rate under high fronthaul capacity. These results suggest that, compared to OMA-based FRAN, NOMA-based FRAN has the potential to support more users with their minimal required rates, thereby improving global user fairness. This tendency is especially pronounced in adverse network conditions where FAPs have low fronthaul capacity.

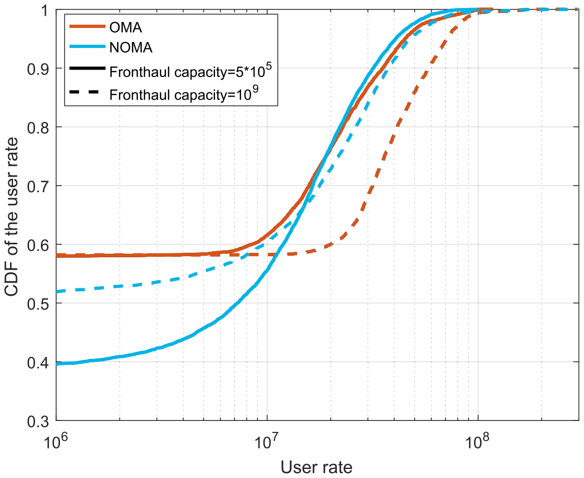


FIGURE 2. CDF of the user rate [bps] for NOMA-FRAN vs. OMA-FRAN, different values of the fronthaul capacity [bps].

User fairness is further illustrated in Fig. 3, which shows the variation of Jain's fairness index as a function of the number of resource blocks available at every FAP. Jain's fairness index is maximum (equal to 1) when all users experience an

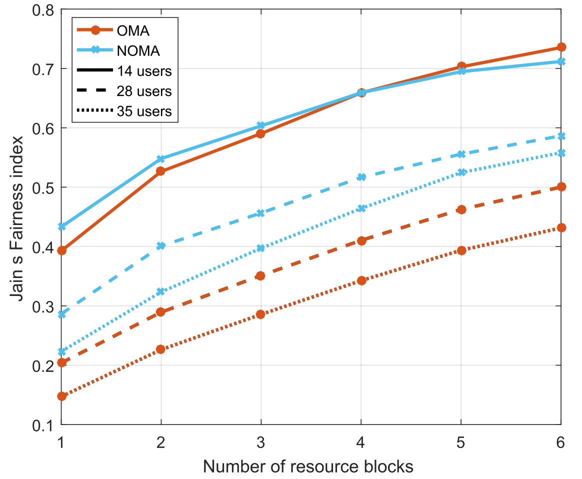


FIGURE 3. Variation of the Jain's fairness index according to the number of resource blocks (fronthául capacity = 1 Gbps).

equal rate. The major observation in Fig. 3 is that, the fairness gain of NOMA based-FRAN against OMA based-FRAN increases as the number of users increases. This behavior confirms the prominence of NOMA-based FRAN when the number of users in the network is high, a promising insight for mMTC transmissions. However, further analysis is needed to determine the optimal number of device packets that can be supported by NOMA based-FRAN within each resource block, under realistic mMTC network parameters.

Next, Fig. 4 depicts the CDF of the delay required for a user to receive a packet from its serving FAP, in downlink FRAN. Two packet sizes are considered: a small packet size of 1064 bits that is typical for IoT applications, and a standard packet size of 12 kbits. Fig. 4 shows that NOMA-based FRAN can largely reduce the experienced delays as compared to OMA-based FRAN. Indeed, for a target delay of 1 ms, NOMA-based FRAN enables 8% more small packets to be received within the deadline in the case of low fronthaul capacity, and up to 20% more with a larger fronthaul capacity.

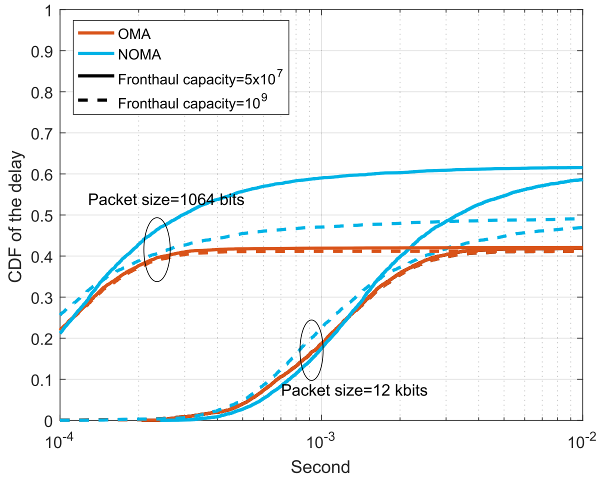


FIGURE 4. CDF of the time duration for transmission for NOMA-FRAN vs. OMA-FRAN, different packets sizes and values of the fronthaul capacity [bps].

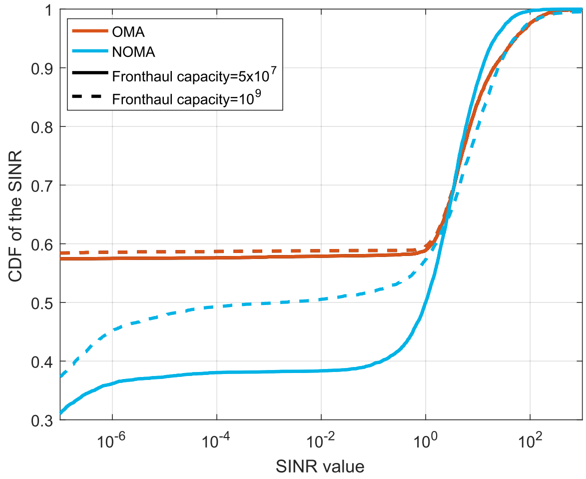


FIGURE 5. CDF of the SINR for a block error rate of 10^{-5} at the users for NOMA-FRAN vs. OMA-FRAN, different values of the fronthaul capacity [bps].

Similar gains can be observed in the case of standard packets, and for a delay deadline of 10ms. These results illustrate the promising possibilities for NOMA-based FRAN to realize significant latency reductions over a wide range of applications and delay requirements.

Lastly, the CDF of users' signal-to-interference-plus-noise ratio (SINR) is shown in Fig. 5. This figure illustrates that under NOMA-based FRAN, a larger proportion of users can achieve a given SINR target in the lower SINR regime, which corresponds to a certain block error rate level. For instance, consider a similar scenario to reference [30], where the user SINR should be at least −5dB to achieve a targeted block error rate of 10^{-5} for URLLC. Then, NOMA-based FRAN enables 5% more users to achieve this SINR target under low fronthaul capacity, and up to 15% more under high fronthaul capacity, as compared to OMA-based FRAN. Hence, NOMA-based FRAN can potentially support a significantly higher number of users while satisfying their minimal error rate requirements that are inherent to URLLC applications.

Despite the fact that the above results are obtained based on the weighted sum-rate maximization algorithms designed for a NOMA-based FRAN in [15], such numerical evaluations demonstrate the promising potential of NOMA-based FRAN in meeting the requirements of B5G systems use-cases, for which the joint gains in terms of rate, fairness, delay, and BER are essential. Our above numerical results fully illustrate the need for more advanced RRA and interference management methods in the context of NOMA-based FRAN, but tailored to the specific demands of each of the future applications. These challenges are further discussed in details in the following section.

IV. CHALLENGES AND RESEARCH DIRECTIONS

NOMA-based FRAN provides a powerful and promising platform for realizing future application requirements. Many fundamental issues, however, remain to be addressed so as to fully assess the gains which can be potentially harvested from

TABLE 2. Integration of NOMA in a FRAN architecture: Some opportunities and open research issues towards supporting B5G requirements.

NOMA use case		Opportunities	Challenges and research directions
Towards the requirements of the 5G use cases	eMBB	• Downlink NOMA, D2D NOMA, edge caching ⇒ 1) better spectral efficiency and users' fairness 2) increased number of served users satisfying their QoS	• Design a comprehensive RRA, IM, caching placement and strategy that includes downlink NOMA and D2D NOMA
	mMTC	• Grant-free NOMA in the uplink transmission ⇒ 1) Increased number of connected devices 2) Reduced energy consumed in signaling 3) Higher transmission success probabilities compared to OMA • MEC architecture ⇒ facilitated data processing	• Optimize the number of multiplexed users through NOMA
	URLLC	• Data storage and computation at the network edge ⇒ reduced latency • Downlink NOMA ⇒ increased network availability (more users are served simultaneously) • Uplink NOMA ⇒ reduced collision	• Analyze the joint provision of high reliability and low latency
B5G use cases		• Cloud-based optimized packet scheduling, RRA and IM ⇒ jointly enhanced user performance metrics • Intelligently multiplexing multiple use case with dissimilar requirements through NOMA ⇒ increased spectral efficiency • Multiplexing multiple access points via NOMA ⇒ improved fronthauling	• Design an optimal joint cloud-edge functional split • Incorporate use case-specific requirements into joint RRA, caching placement and caching strategy design • Design a hybrid and adaptive OMA/NOMA-based transmission scheme • Determine whether the performance gains of NOMA-based FRAN outweigh the costs of SIC processing delays, SIC decoding error propagation, unpredictable interference from grant-free access, imperfect CSIs • Design and analyze a deep learning-based joint cloud-edge processing

such an architectural integration. This section discusses the main open issues towards the design of NOMA-based FRAN. Few opportunities and challenges offered by this architecture are also summarized in Table 2.

A. CLOUD-EDGE FUNCTIONAL SPLITTING AND JOINT PROCESSING

Most of the related works on FRAN or MEC advocate running the delay-sensitive applications locally in the FAP closest to the user, and running the non-delay stringent applications at the cloud. However, in a dense network where a large number of users make use of delay-sensitive applications, this fully distributed and local FAP operation would create detrimental interference levels. Enabling high data rate transmissions in real-time is only possible by smartly and seamlessly distributing computation, storage, signal processing and radio resource management, among cloud and edge entities. Therefore, a comprehensive study of an optimized joint cloud-edge functional split in a FRAN that accounts for each use-case requirement is necessary. Such ‘‘splitting and joining’’ is expected to be seamlessly implemented in real-time using network virtualization and slicing technologies. Given the specific service requirements of each use case as previously discussed, the optimal joint cloud-edge functional split should be, in theory, use case- or even application-specific.

B. RADIO RESOURCE ALLOCATION AND INTERFERENCE MANAGEMENT

To best assess the performance of optimized NOMA-based FRAN, a multitude of RRA problems can be formulated as optimization problems, where the objective function can be chosen based on the underlying application, e.g., weighted sum-rate, weighted-sum power, max-min fairness, etc. Similarly, the optimization problems of interest may have both discrete variables (user-to-FAP association and RB allocation) and continuous variables (FAP power, beamforming vectors, power splitting ratios among different user messages served through the same RB). The constraints can be a combination of 1) maximum power constraint per FAP, 2) fronthaul constraint per FAP, 3) QoS constraint per required application, and 4) local edge FAP-operation as imposed by the extreme latency requirements of delay-sensitive applications. For example, let $\mathbf{x}_k$ be the FAP allocation vector (alternatively, beamforming vector) for user k. Let the size of $\mathbf{x}_k$ be equal to the total number of FAPs. Constraint 4 above can then be cast as a l_0-norm constraint over $\mathbf{x}_k$, i.e., $||\mathbf{x}_k||_0 = 1, \forall k$. In other words, to guarantee the local edge FAP-operation, only one element of $\mathbf{x}_k$ is non-zero, i.e., user k is served by a unique FAP. The discrete and combinatorial nature of this local edge-FAP operation constraint makes the RRA and IM more challenging in the FRAN case, as compared to the CRAN case where any number of RRHs may serve each user.

The weighted sum-rate maximization in a NOMA-based FRAN results into an NP-hard mixed-integer optimization problem. For instance, reference [15] proposes solving such a complex problem by separately solving each RRA building block, then iterating over the whole procedure until convergence to a suboptimal solution. The authors in [15] propose an assignment algorithm based on Multiple Choice Knapsack Problem combined with an auction approach to implement the local edge FAP-operation constraint. It is also shown that thanks to the same constraint, the objective function is separable per resource block. Therefore, the computational complexity of the problem is reduced by applying ADMM to solve the power allocation to each RB, and also the NOMA power-split among the users served through NOMA.

C. CACHING AND TRAFFIC OFFLOADING

Caching placement and traffic offloading strategies constitute major assets of the FRAN architecture. Indeed, the FAPs and end-devices are equipped with different degrees of storage, computation and communication functionalities, offering an unprecedented level of flexibility and adaptation to the space-time variations of user demands. The goal of the caching placement strategy is to determine the optimal FAP or end-device locations for caching, according to file popularity distributions and to available storages, and hence to optimize the files' distribution across FAPs and end-devices. This problem has been studied in [31] for a conventional FRAN without NOMA, where both centralized and distributed modes are proposed for minimizing average user download delay, under cache storage capacity constraints. There is also a big potential for improving caching and traffic offloading, through the hybrid use of the resources of nearby nodes, i.e., other end-devices, as well as of installed infrastructure, i.e., FAPs. Various studies have been conducted in such areas coined as ad-hoc cloud and device-enhanced MEC [32]. These existing methods should be further leveraged in the context of NOMA-based FRAN, which demultiplies the possibilities for joint caching, computation and wireless transmission optimization, in function of the available neighboring nodes and FAPs.

Although dealing with Small Cell Networks rather than FRAN, [33] designed a notable Reinforcement Learning (RL) framework whereby small cell BSs learn what and when to cache, and pre-fetch popular files during off-peak periods. The Q-Learning based algorithm enables to find the optimal caching policy through a joint consideration of local and global file popularity demands, under BSs' memory limitations, unknown popularity profiles and space-time popularity dynamics of user requests. Although this issue has been extensively investigated in the literature, the majority of the existing works deals with the classical FRAN architecture. One of the few works touching upon caching in the context of NOMA-based FRAN can be found in [34]. However, caching is only treated as a constraint, while the objective of this work is to optimize subchannel and power allocation so as to maximize the energy efficiency of a NOMA-based Unmanned Aerial Vehicle (UAV) network, where UAVs are flying FAPs. Hence, comprehensive studies are yet to be undertaken regarding this issue of optimized caching placement and traffic offloading strategies, in order to garner the increased opportunities unleashed by NOMA-based FRAN.

D. HYBRID OMA/NOMA IN FRAN

NOMA users that are far from their serving FAPs are more susceptible to interference from other FAPs, which may cause severe performance degradation if joint cloud-edge interference coordination is not feasible [11]. Moreover, NOMA receivers require good CSI accuracy to avoid error propagation in the SIC decoding process. However, retrieving highly accurate CSI might be challenging in fast dynamic networks. A hybrid and adaptive OMA/NOMA-based transmission scheme can overcome these limitations. For instance, static/low mobility users that are close to their associated FAP can be served through NOMA. However, highly mobile users and users that are farther from their serving FAP may benefit from OMA-based transmissions combined with Coordinated Multi-Point (CoMP)-like techniques [35] among FAPs through joint cloud-edge processing, as serving them through NOMA may degrade their performance due to their high CSI uncertainty.

E. POTENTIAL OF MACHINE LEARNING TECHNIQUES

In the presence of a large number of use cases, applications as well as heterogeneous users/devices with different mobility profiles, overcoming the above mentioned-challenges – joint cloud-edge processing, optimized RRA, caching, and adaptive OMA/NOMA – becomes quickly intractable. Moreover, solving the RRA and IM optimization problem requires accurate CSI knowledge of all user-to-FAP channels at all FAPs and at cloud BBUs, which results in unacceptable amounts of CSI feedback signalling. Therefore, the use of machine learning (ML) techniques, particularly deep learning, recently coined as "network-AI", becomes inevitable. For instance, a trained deep neural network (DNN) can replace the classical RRA and IM optimization problems. The DNN's inputs are selected features such as channel statistics, traffic statistics, required applications, user mobility profiles and its outputs are the user-to-FAP association, caching assignment, power allocation and NOMA power-split ratio between multiplexed users. The DNN's parameters, such as weights, number of layers, functions, should be fine-tuned and adapted to real-time network fluctuations. Massive control signalling exchange and instantaneous CSI feedback would be no longer required. Therefore, a large portion of the scarce radio resource would be freed.

Several research works investigate DNN issues in the context of CRAN and propose a fully centralized DNN located at the cloud [36]. Contrarily, in other research works [37], edge APs or user devices take autonomous decisions regarding resource access using Deep Reinforcement Learning (DRL) for MEC. For NOMA-based FRAN, a deep learning-based joint cloud-edge processing should be investigated. The cloud can perform DNN-based optimization for non-real-time eMBB applications, and DRL-based FAPs handle the delay-stringent applications. Broad avenues of research are yet to be explored to exploit ML techniques at all levels, ranging from network edges to the core cloud, in order to unleash the full potential of NOMA-based FRAN in enabling the awaited B5G services.

V. CONCLUSION

Worldwide research efforts are already turned upon defining B5G wireless systems architectures and service requirements. This article discussed the promising opportunities that result from the integration of NOMA in an FRAN architecture. The paper showed that NOMA-based FRAN has the strong potential of meeting the requirements of envisioned B5G use-cases, and illustrated such applicability in the context of eMBB, mMTC and URLLC. The paper also identified notable design and analysis research directions that are yet to be addressed, so as to fully harvest the various benefits of NOMA-based FRAN in the context of B5G systems. Up to the authors' knowledge, this is the first paper of its kind which numerically advocates for NOMA-based FRAN capabilities at meeting the requirements of B5G systems from a holistic perspective, which is in essence a step forward towards defining the objectives of the much anticipated sixth generation of wireless systems (6G).

REFERENCES

[1] V. Wong, R. Schober, D. Ng, and L. Wang, *Key Technologies for 5G Wireless Systems*. Cambridge, U.K.: Cambridge Univ. Press, 2017.

[2] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, May 2020.

[3] M. Peng, S. Yan, K. Zhang, and C. Wang, "Fog-computing-based radio access networks: Issues and challenges," *IEEE Netw.*, vol. 30, no. 4, pp. 46–53, Jul. 2016.

[4] S.-H. Park, O. Simeone, and S. Shamai, "Joint optimization of cloud and edge processing for fog radio access networks," in *Proc. IEEE Int. Symp. Inf. Theory (ISIT)*, Jul. 2016, pp. 315–319.

[5] N. Pontois, M. Kaneko, T. H. L. Dinh, and L. Boukhatem, "User pre-scheduling and beamforming with outdated CSI in 5G fog radio access networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[6] K. Yang, N. Yang, N. Ye, M. Jia, Z. Gao, and R. Fan, "Non-orthogonal multiple access: Achieving sustainable future radio access," *IEEE Commun. Mag.*, vol. 57, no. 2, pp. 116–121, Feb. 2019.

[7] *Technical Specification Group Services and System Aspects; Release 15 Description, Summary of Rel-15 Work Items (Release 15)*, document 3GPP TR 21.915 V15.0.0 (2019-09), Sep. 2019.

[8] M. Thomas Cover and A. Joy Thomas, *Elements of Information Theory* (Wiley Series in Telecommunications and Signal Processing). Hoboken, NJ, USA: Wiley, 2006.

[9] D. Tse, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, May 2005.

[10] L. Dai, B. Wang, Z. Ding, Z. Wang, S. Chen, and L. Hanzo, "A survey of non-orthogonal multiple access for 5G," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 2294–2323, 3rd Quart., 2018.

[11] K. S. Ali, H. Elsawy, A. Chaaban, and M.-S. Alouini, "Non-orthogonal multiple access for large-scale 5G networks: Interference aware design," *IEEE Access*, vol. 5, pp. 21204–21216, 2017.

[12] Z. Wei, L. Yang, D. W. K. Ng, J. Yuan, and L. Hanzo, "On the performance gain of NOMA over OMA in uplink communication systems," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 536–568, Jan. 2020.

[13] I. Randrianantenaina, M. Kaneko, H. Dahrouj, H. ElSawy, and M.-S. Alouini, "Joint scheduling and power adaptation in NOMA-based fog-radio access networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–6.

[14] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in NOMA-based fog radio access networks," *IEEE Wireless Commun.*, vol. 25, no. 3, pp. 110–115, Jun. 2018.

[15] I. Randrianantenaina, M. Kaneko, H. Dahrouj, H. ElSawy, and M.-S. Alouini, "Interference management in NOMA-based fog-radio access networks via scheduling and power allocation," *IEEE Trans. Commun.*, vol. 68, no. 8, pp. 5056–5071, Aug. 2020.

[16] B. Liu and M. Peng, "Joint resource block-power allocation for NOMA-enabled fog radio access networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2019, pp. 1–6.

[17] X. Wen, H. Zhang, H. Zhang, and F. Fang, "Interference pricing resource allocation and user-subchannel matching for NOMA hierarchy fog networks," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 467–479, Jun. 2019.

[18] J. Kim, D. Yu, S.-H. Moon, and S.-H. Park, "Grouped NOMA multicast transmission for F-RAN with wireless fronthaul and edge caching," in *Proc. 16th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2019, pp. 1–5.

[19] S. Yan, L. Qi, Y. Zhou, M. Peng, and G. M. S. Rahman, "Joint user access mode selection and content popularity prediction in non-orthogonal multiple access-based F-RANs," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 654–666, Jan. 2020.

[20] Z. Shi, S. Ma, H. ElSawy, G. Yang, and M.-S. Alouini, "Cooperative HARQ-assisted NOMA scheme in large-scale D2D networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 4286–4302, Sep. 2018.

[21] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "NOMA assisted wireless caching: Strategies and performance analysis," *IEEE Trans. Commun.*, vol. 66, no. 10, pp. 4854–4876, Oct. 2018.

[22] S. Gurugopinath, P. C. Sofotasios, Y. Al-Hammadi, and S. Muhaidat, "Cache-aided non-orthogonal multiple access for 5G-enabled vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8359–8371, Sep. 2019.

[23] H. Jiang, Q. Cui, Y. Gu, X. Qin, X. Zhang, and X. Tao, "Distributed layered grant-free non-orthogonal multiple access for massive MTC," in *Proc. IEEE 29th Annu. Int. Symp. Pers., Indoor Mobile Radio Commun. (PIMRC)*, Sep. 2018, pp. 1–7.

[24] M. Basit Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for IoT: A survey," 2019, *arXiv:1910.06529*. [Online]. Available: http://arxiv.org/abs/1910.06529

[25] X. Chen, D. Wing Kwan Ng, W. Yu, E. G. Larsson, N. Al-Dhahir, and R. Schober, "Massive access for 5G and beyond," 2020, *arXiv:2002.03491*. [Online]. Available: http://arxiv.org/abs/2002.03491

[26] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[27] R. Kassab, O. Simeone, P. Popovski, and T. Islam, "Non-orthogonal multiplexing of ultra-reliable and broadband services in fog-radio architectures," *IEEE Access*, vol. 7, pp. 13035–13049, 2019.

[28] A. Matera, R. Kassab, O. Simeone, and U. Spagnolini, "Non-orthogonal eMBB-URLLC radio access for cloud radio access networks with analog fronthauling," *Entropy*, vol. 20, no. 9, p. 661, Sep. 2018.

[29] T. M. Nguyen, W. Ajib, and C. Assi, "A novel cooperative non-orthogonal multiple access (NOMA) in wireless backhaul two-tier HetNets," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4873–4887, Jul. 2018.

[30] T. Fehrenbach, R. Datta, B. Goktepe, T. Wirth, and C. Hellge, "URLLC services in 5G low latency enhancements for LTE," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Aug. 2018, pp. 1–6.

[31] J. Liu, B. Bai, J. Zhang, and K. B. Letaief, "Cache placement in fog-RANs: From centralized to distributed algorithms," *IEEE Trans. Wireless Commun.*, vol. 16, no. 11, pp. 7039–7051, Nov. 2017.

[32] M. Mehrabi, D. You, V. Latzko, H. Salah, M. Reisslein, and F. H. P. Fitzek, "Device-enhanced MEC: Multi-access edge computing (MEC) aided by end device computation and caching: A survey," *IEEE Access*, vol. 7, pp. 166079–166108, Nov. 2019.

[33] A. Sadeghi, F. Sheikholeslami, and G. B. Giannakis, "Optimal and scalable caching for 5G using reinforcement learning of space-time popularities," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 1, pp. 180–190, Feb. 2018.

[34] Y. Li, H. Zhang, K. Long, S. Choi, and A. Nallanathan, "Resource allocation for optimizing energy efficiency in NOMA-based fog UAV wireless networks," *IEEE Netw.*, vol. 34, no. 2, pp. 158–163, Mar. 2020.

[35] M. S. Ali, E. Hossain, A. Al-Dweik, and D. I. Kim, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3982–3998, Sep. 2018.

[36] G. Gui, H. Huang, Y. Song, and H. Sari, "Deep learning for an effective nonorthogonal multiple access scheme," *IEEE Trans. Veh. Technol.*, vol. 67, no. 9, pp. 8440–8450, Sep. 2018.

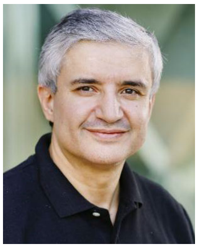
[37] T. Yang, Y. Hu, M. C. Gursoy, A. Schmeink, and R. Mathar, "Deep reinforcement learning based resource allocation in low latency edge computing networks," in *Proc. 15th Int. Symp. Wireless Commun. Syst. (ISWCS)*, Aug. 2018, pp. 1–5.

MEGUMI KANEKO (Senior Member, IEEE) received the B.S. and M.Sc. degrees in communication engineering in 2003 and 2004 from Télécom SudParis (French Grande Ecole), France, jointly with the M.Sc. degree from Aalborg University, Denmark, where she received the Ph.D. degree in 2007, and the H.D.R. degree (French Doctoral Habilitation for Directing Researches at professor position) from the University of Paris-Saclay, France, in May 2017. From January to July 2007, she was a Visiting Researcher with Kyoto University, Kyoto, Japan, and a JSPS Postdoctoral Fellow, from April 2008 to August 2010. From September 2010 to March 2016, she was an Assistant Professor with the Department of Systems Science, Graduate School of Informatics, Kyoto University. She is currently an Associate Professor with the National Institute of Informatics and the Graduate University of Advanced Studies (SOKENDAI), Tokyo, Japan. Her research interests include wireless communications, 5G and beyond, IoT wireless systems, and PHY/MAC design and optimization. Since September 2020, she has been a member of the Advisory Board for Promoting Science and Technology Diplomacy at the Ministry of Foreign Affairs of Japan. She received the 2009 Ericsson Young Scientist Award, the IEEE GLOBECOM 2009 Best Paper Award, the 2011 Funai Young Researcher's Award, the WPMC 2011 Best Paper Award, the 2012 Telecom System Technology Award, the 2016 Inamori Foundation Research Grant, and the 2019 Young Scientists' Prize from the Minister of Education, Culture, Sports, Science and Technology of Japan. She serves as an Editor for IEEE Transactions on Wireless Communications, IEEE Communications Letters, and *IEICE Transactions on Communications*.

ITSIKIANTSOA RANDRIANANTENAINA (Student Member, IEEE) received the Engineering Diplomat degree from the Institut National des Postes et Telecommunications (INPT), Rabat, Morocco, in 2013, and the M.Sc. and Ph.D. degrees in electrical engineering from the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, in 2015 and 2019, respectively.

HAYSSAM DAHROUJ (Senior Member, IEEE) received the B.E. degree (Hons.) in computer and communications engineering from the American University of Beirut (AUB), Lebanon, in 2005, and the Ph.D. degree in electrical and computer engineering from the University of Toronto (UofT), Canada, in 2010. In July 2020, he joined the Center of Excellence for NEOM Research, King Abdullah University of Science and Technology (KAUST), as a Senior Research Scientist. From June 2015 to June 2020, he was with the Department of Electrical and Computer Engineering, Effat University, as an Assistant Professor, and as a Visiting Scholar with the Computer, Electrical and Mathematical Sciences and Engineering (CEMSE) Division, KAUST, where he also was a Research Associate, from April 2014 and May 2015. Prior to joining KAUST, he was an Industrial Postdoctoral Fellow at UofT, in collaboration with BLiNQ Networks Inc., Kanata, Canada, where he worked on developing practical solutions for the design of non-line-of sight wireless backhaul networks. His contributions to the field led to five patents. During his doctoral studies at UofT, he pioneered the idea of coordinated beamforming as a means of minimizing intercell interference across multiple base stations. The journal article on this subject was ranked second in the 2013 IEEE Marconi paper awards in wireless communications. He was a recipient of the Faculty Award of Excellence in Research and the Faculty Award of Excellence in Teaching (at the university level), in May 2017. He is an Associate Editor of the *Frontiers in Communications and Networks*, and a Lead-Guest Editor of the Frontiers Special Issue on "Resource Allocation in Cloud-Radio Access Networks and Fog-Radio Access Networks for B5G Systems." His main research interests include 6G wireless systems, cloud-radio access networks, cross-layer optimization, cooperative networks, convex optimization, distributed algorithms, and optical communications.

HESHAM ELSAWY (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Manitoba, Canada, in 2014. He has joined the Electrical Engineering Department, King Fahd University of Petroleum and Minerals (KFUPM), Saudi Arabia, in Fall 2018, as an Assistant Professor. Prior to that, he was a Postdoctoral Fellow with King Abdullah University of Science and Technology (KAUST), Saudi Arabia, as a Research Assistant at TRTech, Winnipeg, MB, Canada, and as a Telecommunication Engineer at the National Telecommunication Institute, Egypt. His research interests include statistical modeling of wireless networks, stochastic geometry, and queueing analysis for wireless communication networks. During his Ph.D. at the University of Manitoba, he received several academic awards, including the NSERC Industrial Postgraduate Scholarship, from 2010–2013, and the TRTech Graduate Students Fellowship, from 2010–2014. He coauthored three award-winning apapers that are recognized by the IEEE ComSoc Best Survey Paper Award, the Best Scientific Contribution Award to the IEEE International Symposium on Wireless Systems 2017, and the Best Paper Award in Small Cell and 5G Networks (SmallNets) Workshop of the 2015 IEEE International Conference on Communications (ICC). He is recognized as an Exemplary Reviewer by the IEEE Transactions on Communications (2014–2016), the IEEE Transactions on Wireless Communications (2017-2018), and the IEEE Wireless Communications Letters (2018).



MOHAMED-SLIM ALOUINI (Fellow, IEEE) was born in Tunis, Tunisia. He received the Ph.D. degree in electrical engineering from the California Institute of Technology (Caltech), Pasadena, CA, USA, in 1998. He served as a Faculty Member of the University of Minnesota, Minneapolis, MN, USA, and of Texas A&M University at Qatar, Doha, Qatar, before joining the King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia, as a Professor of Electrical Engineering, in 2009. His current research interest includes the modeling, design, and performance analysis of wireless communication systems.

• • •