User-Centric Base-Station Wireless Access Virtualization for Future 5G Networks

Slim Zaidi, Oussama Ben Smida, Sofiène Affes, *Senior Member, IEEE*,
Usa Vilaipornsawai, Liqing Zhang, and Peiying Zhu, *Fellow, IEEE*

***Abstract*— User-centric wireless access virtualization (WAV) allows each user to be served by a set of carefully selected transmission points (TPs) forming a user-specific virtual base station (uVBS) adapted to its environment and quality-of-service (QoS) requirement. In this way, this new concept breaks away from the conventional cell-centric architecture to provide boundaryless communications in future fifth-generation (5G) networks. This fundamental structural 5G evolution and the ultra-dense multi-tier heterogeneous context foreseen in such networks require an inevitable rethinking of efficient scalable TP clustering. As such, this paper proposes three innovative low-cost clustering approaches that enable the user-centric WAV and provide dynamic, adaptive, and overlapping TP clusters while requiring not only negligible overhead cost but also minimum signaling changes at both network and user sides. Contrary to existing clustering techniques, the new ones we propose better leverage the 5G features such as extreme densification and massive connectivity as well as new concepts such as millimeter wave (mmWave) spectrum and massive multiple-input-multiple-output (MIMO). The simulations show that they may achieve until 154% and 282% of throughput and coverage gains, respectively. Furthermore, these approaches are flexible enough to be adapted to different network dimensions (i.e., space and time), thereby paving the way for achieving the dramatic performance improvements required by the 5G networks to cope with the upcoming mobile data deluge.**

***Index Terms*— Wireless/radio access virtualization, user-centric architecture, cloud-radio access network (C-RAN), dynamic adaptive clustering, mmWave, massive MIMO.**

I. Introduction

MOST academic researchers and industry scientists have agreed that the poor cell-edge user experience is the most limiting factor of the fourth-generation (4G) radio access network (RAN). Such an issue was even exacerbated with the recent trend of extreme densification which originally aimed to increase the network capacity by allowing an aggressive frequency reuse across large geographic areas [1]–[7]. Significant research endeavors have been devoted to developing some remedial solutions to this issue such as inter-cell interference coordination, coordinated beamforming [8]–[11], and fractional frequency reuse. Although these solutions offered some performance gains at the cost of increased complexity and overhead, they were unable to completely mitigate the cell-boundary effect.

Using wireless access virtualization (WAV), future fifth-generation (5G) mobile networks will capitalize, in contrast to their predecessors, on both the extreme densification and massive connectivity that will characterize them to provide boundaryless communications [2]–[7]. This would potentially lead to substantial improvements in terms of network's spectral and power efficiencies and, hence, to the fulfillment of 5G's pledge of ubiquitous user experience [12]–[16]. Indeed, with WAV, coverage is planned around the user,[1] making it the network's focal point rather than the cell as is the case in current cell-centric RANs. By adapting the communication link to both its quality-of-service (QoS) requirements and changing propagation environments, the network creates the illusion that each user is virtually followed by a moving cell [17]–[20]. In this way, we break away from the traditional cell-centric RAN to provide boundaryless communications where all users do not experience any cell-edge effects [21]. Practically, this will be done through enabling each user to be served by a set (i.e., cluster) of carefully and optimally selected transmission points (TPs) forming a user-specific virtual base-station (uVBS), making TPs' clustering (i.e., selection) crucial to any user-centric WAV strategy [22].

Nevertheless, this does not necessarily imply that conventional TP clustering approaches developed for 3G/4G networks could be automatically exploited in the virtualized 5G RAN of our concern. Indeed, the fundamental structural 5G evolution toward a user-centric architecture, along with the ultra-dense multi-tier heterogenous context foreseen in such networks, requires an inevitable rethinking of efficient scalable network partitioning into several user-specific virtual base-stations (uVBSs) [23]–[27]. This goal cannot actually be achieved without forsaking the conventional clustering approaches aiming to form TP sets using solely system information,

Manuscript received October 17, 2018; revised January 15, 2019 and March 30, 2019; accepted March 30, 2019. This work was supported by the NSERC/Huawei Canada/TELUS CRD Grant on 5G-WAVES (WAV Enabling Schemes), the DG and CREATE PERSWADE <www.create-perswade.ca> Programs of NSERC, and a Discovery Accelerator Supplement Award from NSERC. Parts of this work were presented in [4] and [19]. The associate editor coordinating the review of this paper and approving it for publication was S. Muhaidat. *(Corresponding author: Slim Zaidi.)*

S. Zaidi is with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON M5S 3G4, Canada (e-mail: slim.zaidi@utoronto.ca).

O. B. Smida and S. Affes are with the EMT Centre of INRS, Montreal, QC H5A 1K6, Canada (e-mail: oussama.ben.smida@emt.inrs.ca; affes@emt.inrs.ca).

U. Vilaipornsawai, L. Zhang, and P. Zhu are with Huawei Technologies Canada Co. Ltd., Ottawa, ON K2K 3J1, Canada (e-mail: usa.vilaipornsawai@huawei.com; liqing.zhang@huawei.com; peiying.zhu@huawei.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCOMM.2019.2910258

[1] User refers here to any user equipment such as wireless devices (i.e., smartphones, sensors, etc.), vehicles, or machines connected to the network.

i.e., TPs' positions and density, their available resources, etc. Although they do not incur significant costs in terms of complexity, overhead, latency, and power consumption, since their resulting uVBSs are predetermined and rarely updated (i.e., static), such approaches often achieve poor performance both in throughput and spectral efficiency [28]. This is mainly due to the fact that these sets are not adapted to the highly changing users' environments stemming from the lack of user-side information such as the user's channel state information (CSI), channel quality indicator (CQI), and signal-to-interference-plus-noise-ratio (SINR).

Many research groups have focused then on developing dynamic adaptive clustering approaches [29]–[36]. Exploiting the users' CSIs and/or received SINRs, these approaches dynamically adapt the TP sets forming the uVBSs to each user's environment and QoS requirements. As the user moves, its uVBS is updated by dropping some TPs and/or adding others so as to achieve much better performance. However, dynamic clustering usually requires that all users share their CSIs and/or SINRs with a central processor able to design and dynamically update the TP sets in order to comply with all users' environments and QoS requirements [37]–[39]. This obviously causes huge overhead, latency, and power costs which will certainly be exacerbated with the network densification and massive connectivity foreseen in future 5G networks. Moreover, the uVBSs are usually formed using highly-complex iterative greedy algorithms that explore all potential set constructions to ultimately settle on network partitions that are very often far from optimal. Besides, in order to completely remove the cell-edge effect, TP sets forming the uVBSs must overlap [36]. This may increase exponentially the number of possibilities and, hence, the clustering complexity. What also makes existing dynamic clustering approaches unsuitable for virtualized 5G RAN is that set construction possibilities may dramatically increase due to the extreme densification and massive connectivity in the ultra-dense multi-tier heterogenous context foreseen in such networks.

Besides their high complexity and cost, most dynamic clustering techniques suffer from another drawback that may also hinder their implementation in virtualized 5G RANs. Indeed, they often ignore key system information such as TP density and available resources, thereby causing substantial discrepancies between traffic loads at different TPs. Among the few attempts to overcome such an issue, we found the pioneering work of Zarifi et al. [36] which has developed a dynamic clustering approach able of balancing the traffic load among different TPs in a user-centric WAV context. By accounting for the TP loads when forming the users' serving uVBSs, the approach in [36] significantly improves dynamic clustering. This comes, however, again at the cost of increased complexity.

In summary, two different TP clustering approaches exist so far: i) static low-cost but inefficient clustering, and ii) dynamic adaptive efficient but highly-complex and expensive clustering. As both dynamic clustering's high efficiency and static clustering's low cost features are key to enabling user-centric WAV, this work aims at developing a *best-of-the-two-worlds* solution that combines both approaches' benefits while avoiding their drawbacks.

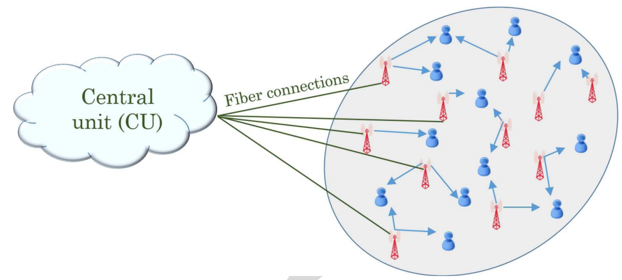


Fig. 1. System model.

In this paper, we propose three innovative low-cost clustering approaches that enable user-centric WAV of base stations and provide dynamic, adaptive, and overlapping TP clusters while requiring not only negligible overhead cost, but also minimum signaling changes at both network and user sides. In contrast to existing clustering techniques, the new ones better leverage 5G features such as extreme densification and massive connectivity as well as new concepts such as millimeter wave (mmWave) spectrum and massive multiple-input multiple-output (MIMO). Furthermore, these approaches are flexible enough to be adapted to different network dimensions (i.e., space, time, etc.), thereby paving the way for achieving the dramatic performance improvements required by 5G networks to cope with the upcoming mobile data deluge.

II. Network Model

The system of our concern consists, as illustrated in Fig. 1, of a cloud-RAN (C-RAN) comprised of M TPs connected through fiber to a central unit (CU) and N users. Each TP is equipped with K antennas while users are assumed, for the sole sake of simplicity, to have a single antenna. We also assume that all users are actively communicating with the network during TP clustering.

III. Proposed User-Centric WAV Approaches

In this section, we propose three innovative clustering approaches aiming to enable user-centric WAV of base stations.

A. Approach 1

In this approach, we propose to use the maximum reference signal received power (RSRP) as user-side information. Let $P^k_{\max}$ denote the maximum RSRP at the k-th user given by

$$P^k_{\max} = \max\{P_{i-k}, i = 1, \ldots, M\}, \quad (1)$$

where P_{i-k} is the RSRP of the i-th TP at the k-th user. Let us also consider a system parameter $\alpha \in [0, 1]$ which encompasses system information such as users' and TPs' densities, positions, and available resources. Using α along with (1), one could build from the M TPs in the C-RAN the following k-th user's serving cluster (SC) (i.e., serving uVBS):

$$\mathrm{SC}_k = \left\{\mathrm{TP}_{i=1,\ldots,M}/\text{s.t. } \alpha P^k_{\max} \leq P_{i-k} \leq P^k_{\max}\right\}. \quad (2)$$

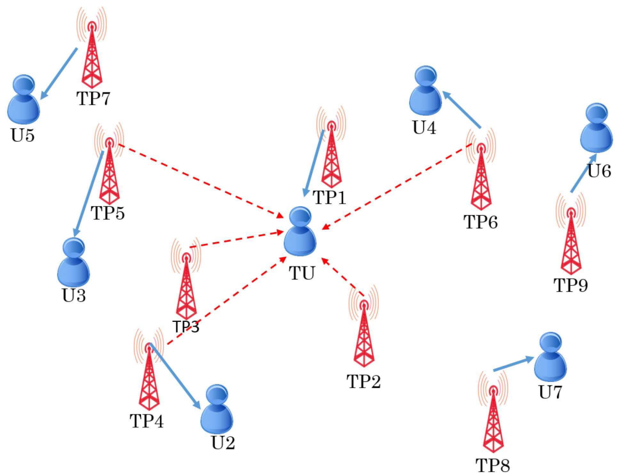


Fig. 2. Single-serving TP selection.

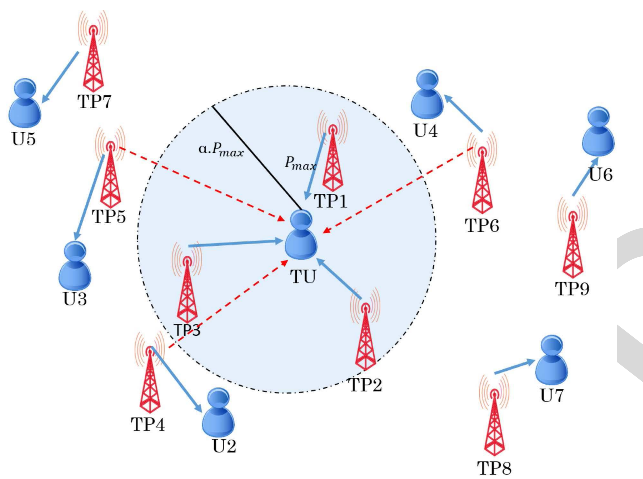


Fig. 3. Approach 1.

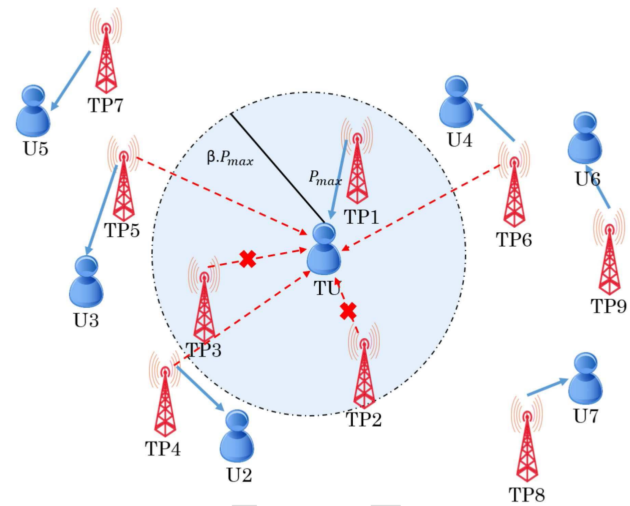


Fig. 4. Approach 2.

In other words, using Approach 1, any TP whose RSRP at the k-th user is large enough to be in the interval $\left[\alpha P_{\max}^k, P_{\max}^k\right]$ will serve it. Let us consider the conventional serving-cell TP selection illustrated in Fig. 2 where each user is served only by the TP with the highest RSRP. Solid blue and dotted red arrows refer to serving and interference links, respectively. From this figure, the target user (TU) is subject to many interference sources from neighboring TPs. However, when the proposed clustering approach is applied, as shown in Fig. 3, most of this interference will be turned into useful power, thereby improving the perceived QoS at this TU.

As α decreases, more TPs may join the SC (i.e., serving uVBS) and, hence, the TU's throughput could be improved by the joint transmission of its data through all TPs in its SC. However, a very small α may unfortunately have an opposite effect on not only the TU's throughput, but also the overall network performance. Indeed, in such a case, other users may have solicited a large number of TPs to jointly transmit their data, thereby decreasing the resources left for allocation to the TU and, hence, its throughput. Moreover, a small α means that more resources are dedicated to a smaller number of users. Since the network resources are limited, it becomes much more likely that an increasing number of TPs and users be in shortage of resources or outage of service, respectively. Consequently, α must be carefully optimized to guarantee a tradeoff between each user QoS and overall network performance, through optimal resources utilization. Computation of this parameter will be carefully discussed later in Section IV.

B. Approach 2

In this approach, we propose that the k-th user requests the TPs causing strong interference to perform interference nulling towards it instead of serving it as in Approach 1. The selected TPs form then the k-th user's nulling cluster (NC) (i.e., nulling uVBS) defined as

$$\mathrm{NC}_k = \left\{\mathrm{TP}_{i=1,\ldots,M}/\text{s.t. } \beta P_{\max}^k \leq P_{i-k} < P_{\max}^k\right\}, \quad (3)$$

where β is a system parameter broadcasted by the CU. Another major difference worth underlining here between Approaches 1 and 2 is that the CU broadcasts both the k-th user's data and CSIs to the TPs in SC_k in the first, but only the CSIs to the TPs in NC_k in the second. Hence, Approach 2 allows both overhead and latency saving. As could be observed from Fig. 4, using Approach 2, the strong interfering links are canceled by performing interference nulling toward TU, resulting thereby in substantial throughput improvement. As β decreases, more interference is canceled and better will be the performance. As in Approach 1, it is not possible to indefinitely decrease β due to the limited TPs' nulling capabilities. Indeed, each TP could perform simultaneous interference nulling toward at most $(K-1)$ users. As β decreases, the number of nulling requests received by a TP increases and may exceed $(K-1)$. At some point, this TP could no longer handle all the constantly-increasing number of nulling requests and, hence, some other users will no longer be able to equally benefit from the TP's nulling capabilities and will suffer instead helplessly from its interference. Accordingly, β must also be optimized to guarantee both optimal system performance and resource utilization. Again, computation of this parameter will be also carefully discussed later in Section IV.

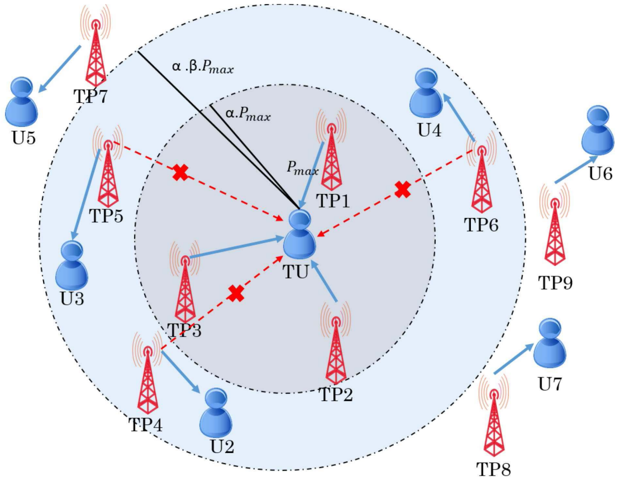


Fig. 5. Approach 3.

C. Approach 3

In this approach, we propose to combine the two previous approaches, as illustrated in Figure 5. Two different clusters are then associated at the same time to the k-th user:

$$\mathrm{SC}_k = \left\{\mathrm{TP}_{i=1,\ldots,M}/\mathrm{s.t.}\ \alpha P_{\max}^k \le P_{i-k} \le P_{\max}^k\right\}, \quad (4)$$

and

$$\mathrm{NC}_k = \left\{\mathrm{TP}_{i=1,\ldots,M}/\mathrm{s.t.}\ \beta\alpha P_{\max}^k \le P_{i-k} < \alpha P_{\max}^k\right\}. \quad (5)$$

This means that the k-th user will send serving requests to the TPs with high RSRPs (i.e., $P_{i-k} \in \left[\alpha P_{\max}^k, P_{\max}^k\right]$) and nulling ones to those with moderate RSRPs (i.e., $P_{i-k} \in \left[\beta\alpha P_{\max}^k, \alpha P_{\max}^k\right[$) yet strong enough to affect the TU's performance. Hence, joint optimization of both α and β is required in this approach.

As mentioned above, all the proposed approaches rely on clever choices of α and/or β that must be properly optimized to guarantee optimal network performance. One should then investigate the available methods able to compute such parameters. Some of them are listed and carefully discussed in the next section.

IV. Parameters Computation

The system parameters α and β may be actually computed online or offline using one of the following methods:

- **System-level simulations (i.e,. heuristic):** The parameters are obtained offline using a system-level simulator. A set of α and/or β value/s are first picked from the interval $[0, 1]$ with ideally a small step before running a simulation campaign for each them. The optimal parameters are those providing the best overall network performance. This process should be repeated for different network setups (i.e., different TP and user densities). Please note that we opt in this work for this heuristic method due to its simplicity and low cost.
- **Experimentation:** The parameters are obtained online by conducting several field-tests during network operation. This method obviously offers more accurate results but increases considerably the cost.
- **Calibration:** α and/or β could be randomly selected from the interval $[0, 1]$ or initialized by one of the two methods listed above and then broadcasted throughout the network. The CU then saves the resulting throughput before updating after each given time period (in minutes, hours, days, ... depending on traffic variations) α and β as $\alpha \pm \Delta\alpha$ and $\beta \pm \Delta\beta$ and then broadcasting them once again throughout the network. If the resulting throughput increases or decreases, the CU calibrates both parameters accordingly at their next broadcast. These steps can be repeated online very rapidly until stabilization, then at relatively much slower paste for regular updates as the need be.
- **Artificial Intelligence (AI) and Machine Learning (ML):** TPs could help the CU build the complex relationship between the optimal parameter values and the network and user information by applying AI and ML online over their data. The latter can be easily collected in a C-RAN deployment through the centralized fiber connections to the CU.

Please note that α and/or β could be computed for the whole network or locally (i.e. location-based parameters) for each subnetwork (i.e., group of TPs and users). This makes our new clustering approaches more adequate for deployment in different subnetwork conditions varying from one sub-area to another and, hence, capable of further enhancing the overall network performance. Subnetworks are not only allowed to adopt different parameters, but also different approaches, among the three proposed here. Besides the spatial dimension, one may also exploit the temporal one for even better adjusted service differentiation among subnetworks and obtain time-varying (i.e., period-based parameters) values of α and/or β that properly adjust to each subnetwork's traffic load variations using for instance the calibration method discussed above. Furthermore, α and/or β can be adapted to different network applications and services (i.e., *application-* and *service-based* parameters). Smaller and/or larger value(s) α and/or β, should be chosen to accommodate high data-rate or QoS applications and services to provide them with more payload and/or nulling resources, and vice-versa.

V. Enabling Mechanisms

In this section, we present and discuss the different mechanisms that may enable the implementation of the above developed approaches. We have actually three different options:

A. Option 1: User Recommends TP Cluster(s)

If this option is adopted, each user selects its own TP cluster(s) based on Approach 1, 2, or 3 and feedbacks only the RSRPs of the TPs in SC and/or NC to the network. Using this scheme, users do not transmit any non-selected TP's RSRPs, thereby reducing not only their power consumption, but also the system overhead cost. However, it requires that the network broadcast α and/or β. Obviously, these information of at most two reals incur negligible additional overhead

TABLE I

COMPLEXITY COMPARISON BETWEEN THE PROPOSED APPROACHES AND SOME BENCHMARKS AVAILABLE IN THE LITERATURE

	Offline processing	Infrastructure-side online processing	User-side online processing
Proposed approaches	Yes	0	$O(M)$
Approach in [30]	No	$O(KNM^2 + KM^2 + 4KNM + 4KM)$	$O(M)$
Approach in [40]	No	$O(141MK^2N^2 + 141MK^2N^2 - 39MKN^2 + 235KN^2 - 38MKN + 470KN - 65N)$	$O(M)$

and power costs and, further, do not require any additional computing capability at the network side. Option 1 requires then a minimal change at low-cost of the current RAN generation. Another advantage of this option is that it allows refinement or overwriting of each user's TP cluster(s). Indeed, the network may deny the access to some TPs for instance when their traffic load is extremely high or to serve users with higher priority or QoS requirement. In such a case, some selected TPs could be substituted or completely removed from SC and/or NC.

B. Option 2: User Decides on TP Cluster(s)

The decision on TP cluster(s) may be locally made by each user using the parameter(s) broadcasted by the network. In such a case, the user does not feedback any RSRP, thereby further reducing both the overhead and power costs. However, each user needs to inform the network of its selected TPs at the cost of a negligible overhead. Even the latter could actually be easily avoided if the user simply feedbacks the CSIs/CQIs of the selected TPs during the transmission phase that follows the clustering phase. The main drawback of Option 2 is that the network is unable to overwrite users' TP clusters to cope with certain conditions, users' priority, or QoS requirements. Nevertheless, this responsibility could be easily handled by the user itself at the cost of additional complexity at its side.

C. Option 3: Network Decides on TP Clusters

With Option 3, each user feedbacks all its RSRPs to the network which decides on TP clusters without broadcasting α and/or β. It is obvious that the main drawback of this option is the overhead and power costs it incurs. Such costs may certainly be exacerbated with the network densification and massive connectivity foreseen in future 5G networks. However, Option 3 is simple and does not require the least change at the user side, making it potentially an interesting candidate for early deployments of 5G networks.

Once again, we have flexibility in the choice of one of the above mechanisms. Indeed, different options could be used with different subnetworks, at different periods, and/or for different applications and services. Furthermore, the choice of Option 1, 2, or 3 may depend on the network or each subnetwork conditions. Option 3 could be preferred at high traffic loads to allow the network make some adjustments on TP clusters more easily while Option 1 or 2 could be adopted at low traffic loads. The choice among the above options could also depend on the users' priority requirements. For instance, privileged users or customers are allowed to use Option 2 while the rest of the subscribers are only entitled to Option 3. Moreover, the selected option could depend on the user equipment's capabilities. The smarter is the terminal, more suitable to it will be Option 2. And the higher is its power budget, more appropriate for it will be Option 3. Accordingly, Options 1 and 2 should find better use with future smart devices (smartphones, sensors, etc.) having limited power resources.

VI. ADVANTAGES OF THE PROPOSED APPROACHES

We summarize below the advantages of the proposed WAV approaches:

- **Low complexity:** Our approaches solely require the optimization of one or two parameters for utilization by multiple users in the same network or subnetwork. Such optimization could be easily achieved through offline simulations and/or calibration as discussed previously. In contrast to the clustering approaches existing thus far, we avoid the implementation of highly-complex iterative greedy, yet often sub-optimal algorithms. Table. I shows the complexity of the proposed approaches and the clustering algorithms developed in [30] and [40] at both infrastructure and user sides. In all clustering solutions, all user equipments are expected to forward the information they collect each on the TPs in their vicinity. Therefore, the user-side complexity is proportional to M. On the other hand, at the infrastructure side, whereas the proposed approaches require no extra processing since the parameters alpha and beta are computed offline, once for all, the conventional clustering techniques suffer from relatively huge complexity loads significantly increasing with the numbers of TPs, per TP antennas, and users. Consequently, in contrast to the latter, the proposed approaches may capitalize both on the extreme densification and massive connectivity foreseen in the upcoming 5G networks.
- **Dynamic, adaptive, user-centric:** With our approaches, the TP clusters are formed from overlapping sets whose cardinalities (i.e., the number of TPs in each set) are adapted dynamically to the users' operating conditions. As one example, using Approach 1, more (less) serving TPs are associated with a user when it is subject to high (low) interference. Such a feature is key to fulfill the 5G pledge of ubiquitous user experience.
- **Low overhead and latency:** Using our approaches alongside Option 1 or 2, the clustering decisions are made locally at the user side. This is in contrast with most existing clustering approaches which require that the CU be aware of all users' CSIs/SINRs to be able to form the TP sets [30], [36], [40]. Hence, overhead and

latency can be significantly reduced with the developed approaches. Indeed, the overhead incurred by the conventional clustering approaches could be expressed as $B_{\rm oh} = R_r K Q_l \sum_{i=1}^{N} M_i$ where $M_i \in \{1, \ldots, M\}$ is the number of TPs in the i-th user vicinity, Q_l is the quantization level of CSI/SINR, and R_r^{CSI} is the clusters formation refreshment rate. On the other hand, the overhead incurred by Approach 3, which requires the broadcast of both α and β, is $B_{\rm oh}^{Prop} = 2R_r^{\alpha,\beta}\acute{Q}_l$ where $R_r^{\alpha,\beta}$ is the refreshment rate of α and β and $\acute{Q}_l$ is their quantization level. Assuming for extreme simplification in favor of the conventional clustering techniques that $Q_l = \acute{Q}_l$,[2] we have $\Omega = B_{\rm oh}/B_{\rm oh}^{Prop} = (R_r^{CSI}\sum_{i=1}^{N} M_i)/2R_r^{\alpha,\beta})$. Therefore, Ω substantially increases not only with the users', TPs', and antennas' numbers, but also with $R_r^{CSI}/R_r^{\alpha,\beta}$. Note here that the CSI's refreshment rate is usually in the range of milliseconds, i.e., in the TTI (transmission time interval) duration scale in LTE, while that of α and β is in the range of minutes or even hours, since they depend on the numbers of TPs and users. This is actually a fundamental difference that drastically reduces the overhead cost. Assuming for simplicity, again in favor of conventional clustering techniques, that $R_r^{CSI}/R_r^{\alpha,\beta} = 10^3$, we measure $\Omega = 27.3\,10^3$ and $\Omega = 35.7\,10^3$ when $\rho = 0.31$ and $\rho = 0.44$, respectively, with the simulation setup described in Section VII. This means that the proposed approaches, under the most unfavorable assumptions to them (i.e., equal quantization level and much smaller than expected refreshment rate ratio), still incur as much as 10^3 times less overhead, and consequently much less latency as well (following the same rationale) than their conventional counterparts, making them unambiguously more suitable for ultra-reliable and low latency (URLLC) 5G services.

- **Scalability:** It is obvious that the performance gain achieved by the proposed approaches increases with the available network resources. Therefore, they may capitalize on multi-user strategies that allow users to share the same resources as well as on new concepts envisioned in 5G such as mmWave spectrum and massive MIMO which offer abundant spectrum and huge degrees of freedoms, respectively. For instance, Approach 1 may take advantage of the mmWave spectrum while Approach 2 may capitalize on massive MIMO. As far as Approach 3 is concerned, it may take advantage of both concepts. This is in contrast with existing clustering techniques whose complexity increases exponentially with such technologies.
- **Flexibility:** By associating different parameters to different network dimensions, our approaches pave the way towards dramatic improvements in both spectral and power efficiencies. Indeed, the definition of **user-class**-, **service**-, and **application**-based parameters allows adequate adaptation of the allocated resources to different classes of subscribers and network services and applications. Furthermore, **period**- and **location**-based parameters that properly adjust to the network conditions at different places and periods would further enhance the throughput of each user. This is again a key feature to fulfill the 5G pledge of ubiquitous user experience.

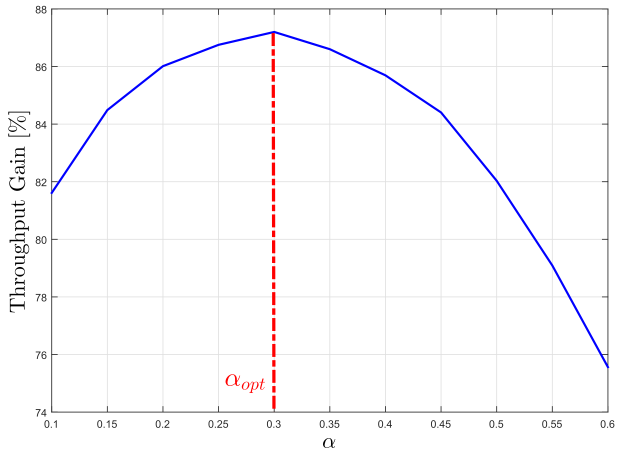
[2]Please note that this assumption is made only for the sake of simplicity. We will show later in Section VII that α and β requires much less accuracy and, hence, much smaller quantization level than CSIs/SINRs.

VII. Simulations Results

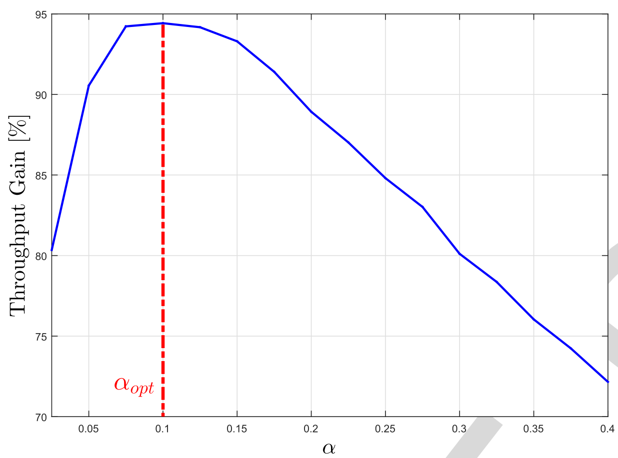
In this section, system-level simulations are conducted to analyze the performance of the proposed approaches and compare them with the conventional single-serving TP selection and a static clustering solutions. The static clustering technique partitions the network into three adjacent TPs set wherein the user is served by one TP while the others perform interference nulling towards it. The heuristic method described in Section III is adopted here to optimize the parameters α and β. In order to highlight the gains provided by Approaches 1, 2 and 3, we remove any form of multi-user MIMO (MU-MIMO) from our LTE standard-compliant simulator. This means that only one user is associated with each single resource in the spectral and spatial domains. In all simulations, we consider 7 macro-TPs and 10 femto-TPs in each macro with transmit powers of 46 dBm and 20 dBm, respectively, ITU-R channel models of bandwidth 10 MHz, and a full buffer traffic model. We also consider that users are initially (i.e., at $t = 0$) uniformly distributed in the target area. All TPs are assumed to be equipped with two antennas (i.e., $K = 2$) while users are equipped with a single antenna. A proportional fair (PF) scheduling is adopted locally at each TP. TP clustering is updated at each subframe at the same rate of dynamic point selection (DPS) introduced in LTE release 11 [46]. In this work, maximum ratio transmission (MRT) is employed by SC TPs (i.e., serving uVBSs) to jointly transmit the user's data while zero-forcing beamforming is implemented by NC TPs (i.e., nulling uVBSs) to avoid interfering on it. Please note that we have opted for these particular signal combining techniques only for the sole sake of simplicity. Our new approaches can, however, support any other advanced signal combining and/or nulling techniques [41]–[45].

A. Approach 1

Fig. 6 plots the achieved network throughput gains of Approach 1 over single-serving TP selection versus α for different values of the TP-user densities ratio $\rho = M/N$. We consider in Figs. 6(a) and 6(b) 35 and 25 users per macro-TP, respectively. From these figures, we confirm the existence of an optimum value $\alpha_{\rm opt}$ for the parameter α. We also observe that $\alpha_{\rm opt}$ depends on ρ. Indeed, it increases when the ρ decreases and vice-versa. This is hardly surprising since the available resources per user increase with ρ and, hence, more serving requests could be accepted by the TPs. In such a case, more TPs may join each user's SC (i.e., serving uVBS), thereby decreasing $\alpha_{\rm opt}$. For instance, we find that $\alpha_{\rm opt} = 0.3$ when $\rho = 0.31$ whereas $\alpha_{\rm opt} = 0.1$ when $\rho = 0.44$. In these cases, Approach 1 achieves throughput gains as high as 49% and 83%, respectively. On the other



(a) $\rho = 0.31$



(b) $\rho = 0.44$

Fig. 6. Network throughput gain of Approach 1 over single-serving TP selection versus α for different values of TP-user densities ratio ρ.

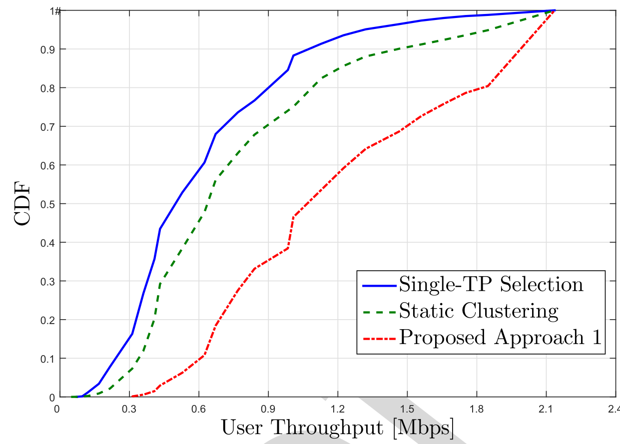


Fig. 7. CDFs of the user throughput achieved by Approach 1, single-serving TP selection, and static clustering when $\alpha_{\text{opt}} = 0.3$ and $\rho = 0.31$.

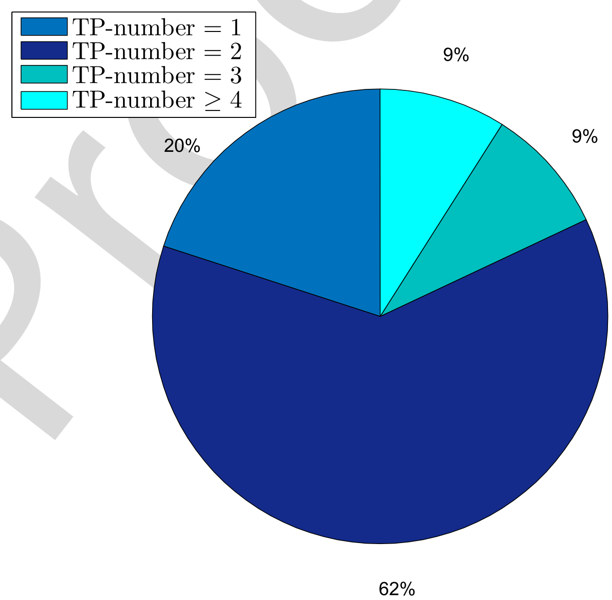
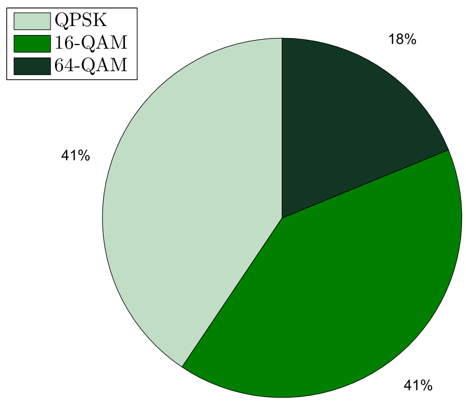


Fig. 8. Pie chart of probabilities for the number of serving TPs in each user's SC with Approach 1 for $\alpha_{\text{opt}} = 0.3$ and $\rho = 0.31$.

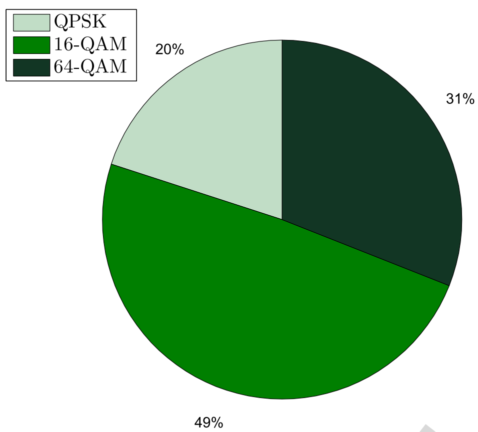
hand, from Fig. 6, a deviation of until 10% from the optimal value of α results in at most 2% loss in throughput gains. This very important feature makes α_{opt} robust against quantization errors. Therefore, with a low quantization level turning out to be acceptable, the overhead incurred when broadcasting α can be further reduced significantly.

Fig. 7 plots the CDFs of the user throughput achieved by Approach 1, single-serving TP selection, and the static clustering solution. With Approach 1, the throughput achieved by 42% of the users exceeds 1.2 Mbits/s while only 6% and 15% of users reach the same throughput level with single-serving TP selection and static clustering, respectively. This proves the efficiency of the proposed approach and highlights the dramatic performance improvements it may provide at low complexity, latency, and overhead costs, making it an interesting candidate for future 5G networks.

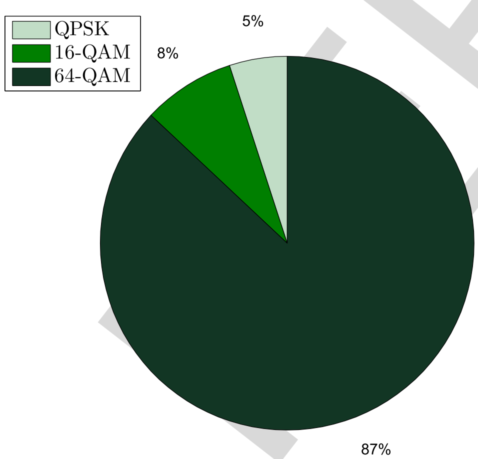
Fig. 8 illustrates the pie chart of probabilities for the number of serving TPs in each user's SC with $\alpha_{\text{opt}} = 0.3$ and $\rho = 0.31$. We observe that 20% of the users are served by a single TP whereas 62% of them are simultaneously served by two TPs, 9% by three, and the rest by four or more. Hence, in 90% and 97% of the cases, the user's SC cardinality does not exceed two or three, respectively, and as such does not burden the network virtualization cost. Furthermore, by relying on the collaboration of several TPs, Approach 1 exploits a form of MIMO commonly known as distributed MIMO. The latter reduces the need for deploying or soliciting a relatively costly network infrastructure components such as massive MIMO TPs with very large co-located antennas (inherently handled as well by all three approaches of the proposed BS WAV scheme, yet not considered here due to lack of space) by much more efficient use of the network resources already available. From a broader perspective, the more TPs whether distributed or co-located are available in the network, the larger would be the TP sets cardinality, thereby paving the way toward massive and even ultra massive (UM)-MIMO. Such very desirable feature makes once again the proposed WAV scheme an interesting candidate for future 5G networks.



(a) Single-serving TP selection

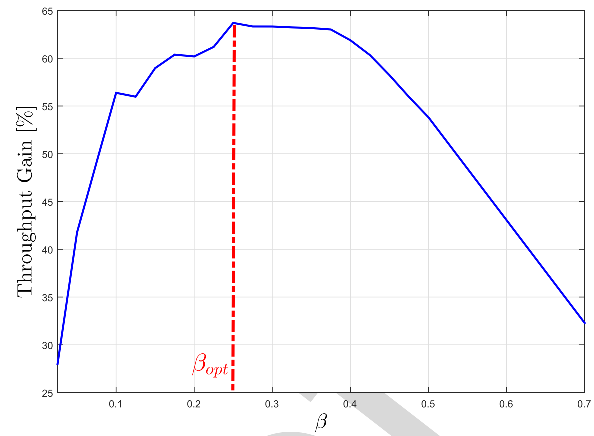


(b) Static selection

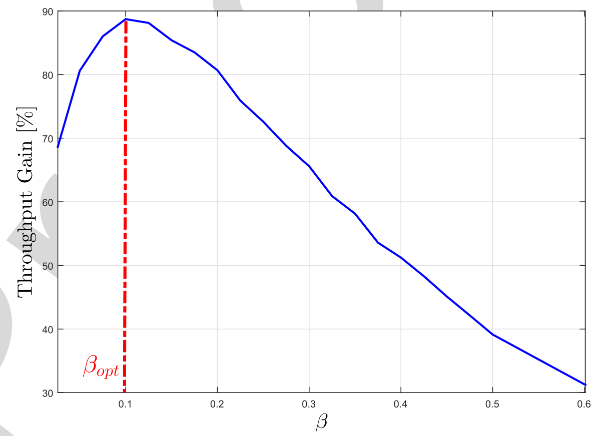


(c) Approach 1 with $\alpha_{\rm opt} = 0.3$ and $\rho = 0.31$

Fig. 9. Occurrence probabilities of the QPSK, 16-QAM, and 64-QAM modulations.



(a) $\rho = 0.31$



(b) $\rho = 0.44$

Fig. 10. Network throughput gains of Approach 2 over single-serving TP selection versus β for different values of ρ.

Fig. 9 shows the occurrence probabilities of QPSK, 16-QAM, and 64-QAM obtained with Approach 1, single-serving TP selection, and static clustering. With Approach 1, 64-QAM occurs 87% of the time against 18% and 31% with single-serving TP selection and static clustering, respectively. This is expected since Approach 1 offers a dramatic SINR improvement by turning the most powerful interference experienced by each user into a useful one, thereby increasing drastically its link capacity. Consequently, the proposed WAV approach enables the adoption of higher-order modulations in 5G networks to ensure higher rates that better cope with the unprecedented mobile data deluge foreseen in the near future.

B. Approach 2

In Figs. 10(a) and 10(b), we plot the throughput gain achieved by Approach 2 over single-serving TP selection versus β when $\rho = 0.31$ and $\rho = 0.44$, respectively. These figures confirm the existence of an optimum value $\beta_{\rm opt}$ for the parameters β that depends on ρ. They also confirm the robustness of $\beta_{\rm opt}$ against quantization errors. On the other hand, the user throughput CDFs in Fig. 11 confirm the significant superiority of Approach 2 over single-serving TP selection and static clustering. Fig. 12 suggests that the optimal throughput gain of Approach 2 can be achieved with 78% of the users requesting only a single nulling TP.

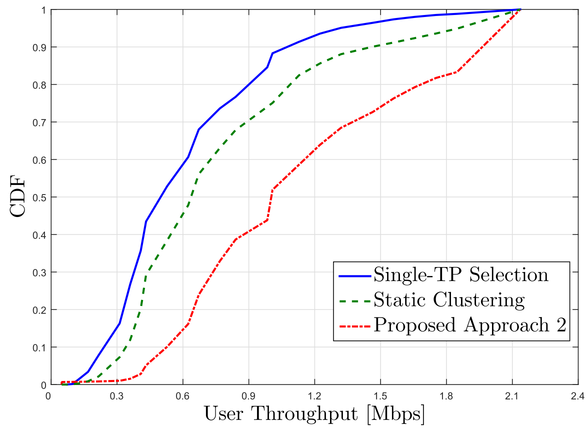


Fig. 11. CDFs of the user throughput achieved by Approach 2, single-serving TP selection, and static clustering when $\beta_{\text{opt}} = 0.25$ and $\rho = 0.31$.

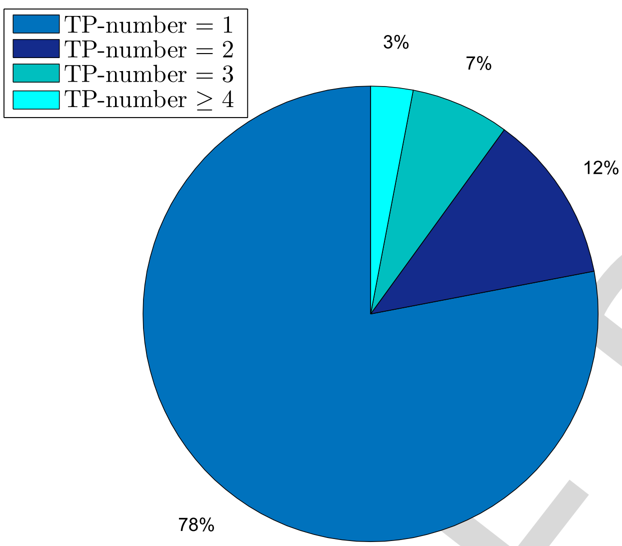


Fig. 12. Pie chart of probabilities for the number of nulling TPs in each user's NC with Approach 2 for $\beta_{\text{opt}} = 0.25$ and $\rho = 0.31$.

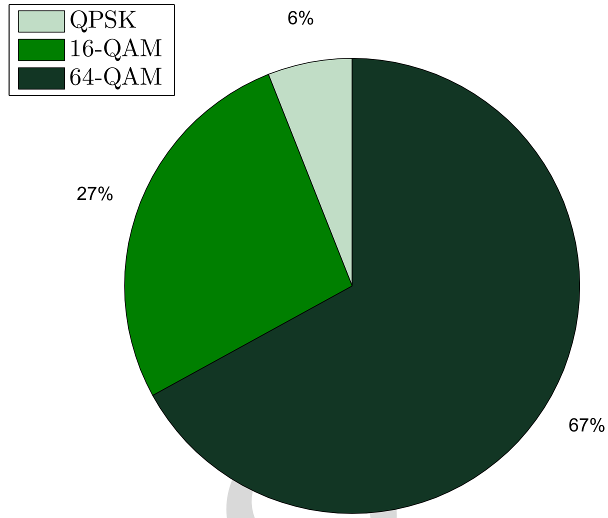


Fig. 13. Occurrence probabilities of the QPSK, 16-QAM, and 64-QAM modulations with Approach 2 for $\beta_{\text{opt}} = 0.25$ and $\rho = 0.31$.

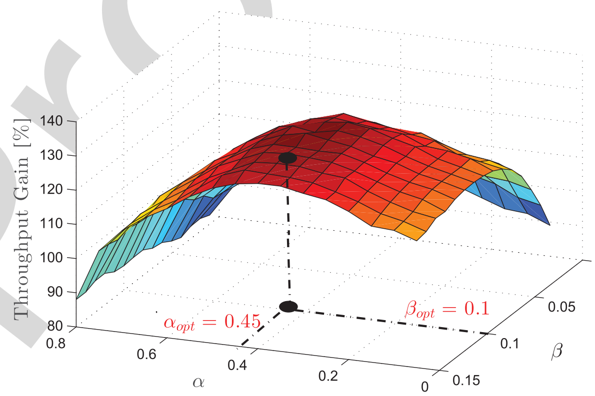


Fig. 14. Network throughput gains of Approach 3 over single-serving TP selection versus α and β for $\rho = 0.31$.

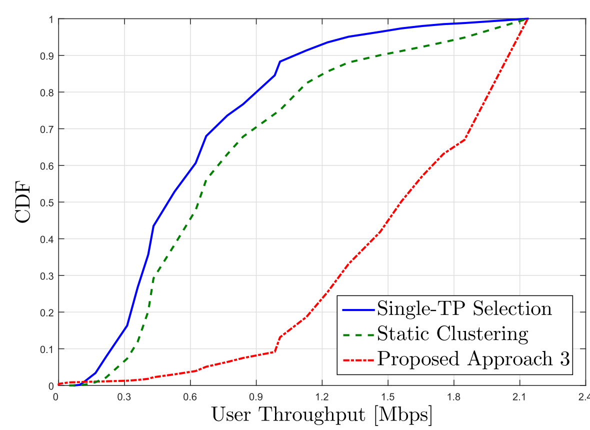


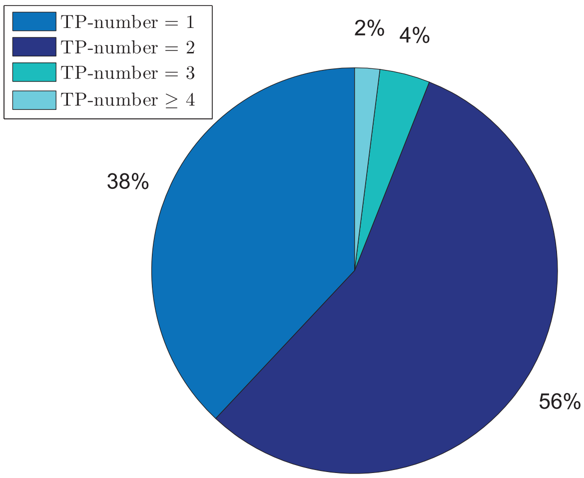
Fig. 15. CDFs of the user throughput achieved by Approach 3, single-serving TP selection, and static clustering when $(\alpha_{\text{opt}}, \beta_{\text{opt}}) = (0.45, 0.1)$ and $\rho = 0.31$.

Fig. 13 shows the occurrence probabilities of QPSK, 16-QAM, and 64-QAM modulations with Approach 2 and suggests that 64-QAM occurs 67% of the time. This is once again hardly surprising since Approach 2, like Approach 1, also offers a dramatic SINR improvement, although relatively lower due to nulling instead of combining. Hence, the lower occurrence of 64-QAM with Approach 2 instead of Approach 1. However, the former has the merit of avoiding user data broadcast to the NC TPs (i.e., nulling uVBS) during the transmission phase. All these results underline once again the great potential of the proposed WAV approaches in enabling future 5G networks.

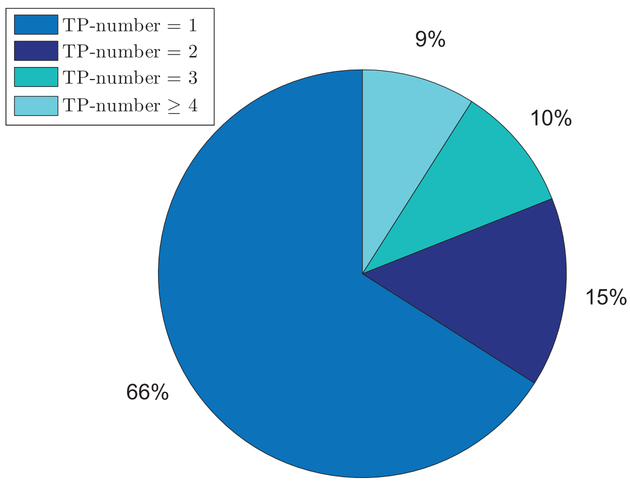
C. Approach 3

Fig. 14 plots the achieved network throughput gain of Approach 3 over single-serving TP selection versus α and β for $\rho = 0.31$. From this figure, we confirm the existence of optimum values $(\alpha_{\text{opt}}, \beta_{\text{opt}})$ for the parameters (α, β) that depend once again on ρ. We find that $(\alpha_{\text{opt}}, \beta_{\text{opt}}) = (0.45, 0.1)$ when $\rho = 0.31$. In such a case, the proposed approach achieves a throughput gain as high as 120%.

Fig. 15 plots the CDFs of the user throughput achieved by the proposed approach, single-serving TP selection, and



(a) Serving TPs in SC



(b) Nulling TPs in NC

Fig. 16. Pie charts of probabilities for the number of serving and nulling TPs in each user's SC and NC, respectively, with Approach 3 for $(\alpha_{\rm opt}, \beta_{\rm opt}) = (0.45, 0.1)$ and $\rho = 0.31$.

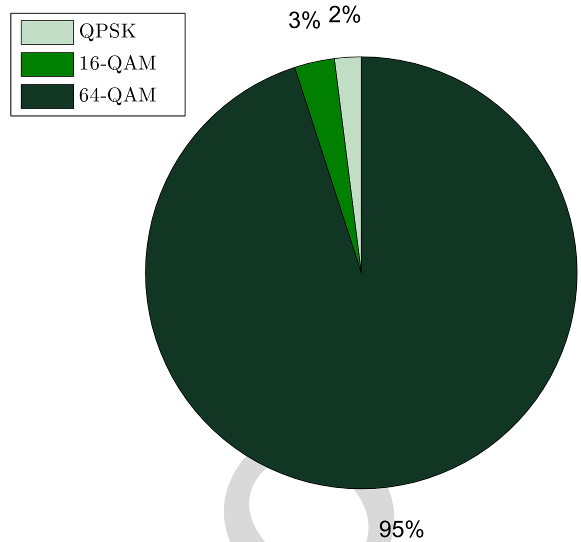


Fig. 17. Occurrence probabilities of the QPSK, 16-QAM, and 64-QAM modulations with Approach 3 for $(\alpha_{\rm opt}, \beta_{\rm opt}) = (0.45, 0.1)$ and $\rho = 0.31$.

static clustering. We observe that the throughput achieved by 55% of the users exceeds 1.5 Mbits/s with Approach 3 whereas only 3% and 10% of users reach the same throughput level with single-serving TP selection and static clustering, respectively. Besides these throughput gains, Approach 3 achieves significant coverage gains against its counterpart, thereby reducing (if not suppressing) the cell-edge effect.

Figs. 16(a) and 16(b) illustrate the pie charts of probabilities for the number of serving and nulling TPs in each user's SC and NC, respectively. In Fig. 16(a), 38% of the users are served by a single TP whereas 56% of them are simultaneously served by two TPs, 4% by three, and the rest (about 2%) by four or more. In Fig. 16(b), only one TP cancels its interference towards 66% of the users whereas two TPs simultaneously cancel their interference towards 15% of them, three are required for 10% of users, and four or more for the rest (about 9%). Hence, in most cases, each user's SC (i.e., serving uVBS) or NC (i.e., nulling uVBS) cardinality does not exceed two and as such Approach 3 does not burden the network virtualization cost.

Fig. 17 shows the occurrence probabilities of QPSK, 16-QAM, and 64-QAM when Approach 3 is employed. We observe that 64-QAM occurs 95% of the time against 18% with single-serving TP. It is noteworthy here that Approach 3 offers a higher occurrence of 64-QAM than Approach 1 and 2. This is hardly surprising since Approach 3 offers a dramatic SINR improvement by turning the strongest interference links into useful ones and by canceling the moderate yet still problematic ones, thereby increasing drastically its link capacity. Consequently, Approach 3 enables the adoption of higher-order modulations in 5G networks to ensure higher rates that better address the unprecedented demand for mobile data expected in the near future.

Tab. II summarizes the performance of the proposed approaches and compares them with single-serving TP selection and static clustering. We show that the proposed approaches dramatically outperform single-serving TP selection in terms of both throughput and coverage. Indeed, Approaches 1, 2, and 3 achieve an average throughput of 1.469, 1.256, and 1.785 Mbps while single-serving TP and static clustering do not exceed 0.703 and 0.924 Mbps, respectively. This represents a throughput gain of up to 108.9%, 78.6%, and 153.9%, respectively, against the single-serving TP and 58.9%, 35.9%, and 93.1%, respectively, against static clustering. Furthermore, according to Tab. II, Approaches 1, 2, and 3 achieve a coverage gain over single-serving TP of 197.7%, 138.2%, and 282.2%, respectively, and of 139%, 91.3%, and 206.9%, respectively, over the static one. These huge performance gains highlight the efficiency of the proposed approaches and their net superiority over their conventional benchmarks.

Tab. III shows the performance of the proposed approaches and their counterparts using high-order modulations (i.e., 256-QAM and 1024-QAM). We show that the average throughput and coverage of all techniques improve with respect to the previous setup of Tab. II that uses only QPSK, 16-QAM, and 64-QAM. However, Approaches 1, 2, and 3

TABLE II
PERFORMANCE OF THE PROPOSED APPROACHES AND THEIR COUNTERPARTS

	QPSK	16-QAM	64-QAM	Average Throughput [Mbps]	5-th Percentile Coverage [Mbps]
Single-serving TP	41%	41%	18%	0.703	0.175
Static approach	20%	49%	31%	0.924	0.218
Approach 1	3%	6%	91%	1.469	0.521
Approach 2	8%	23%	69%	1.256	0.417
Approach 3	2%	3%	95%	1.785	0.669

TABLE III
PERFORMANCE OF THE PROPOSED APPROACHES AND THEIR COUNTERPARTS USING VERY-HIGH-ORDER MODULATIONS

	QPSK	16-QAM	64-QAM	256-QAM	1024-QAM	Average Throughput [Mbps]	5-th Percentile Coverage [Mbps]
Single-serving TP	38%	41%	15%	4%	2%	0.771	0.197
Static approach	16%	42%	35%	4%	3%	1.135	0.287
Approach 1	1%	3%	56%	26%	14%	2.470	0.739
Approach 2	4%	15%	61%	14%	6%	2.004	0.624
Approach 3	1%	0%	41%	37%	21%	3.093	0.922

relatively benefit a lot more from very-high-order modulations that will most likely characterize future 5G and beyond RANs. Indeed, they increase average throughput by up to 68.1%, 59.1%, and 73.2%, respectively, whereas the single-serving and static approaches merely offer 9.6% and 22.8% gains, respectively. As far as coverage is concerned, it improves by 41.8%, 49.6%, and 37.9% with Approaches 1, 2, and 3, respectively, while it increases only by 12.5% and 31.6% with the single-serving and static approaches, respectively. Indeed, the proposed WAV approaches provide much better SINR performance than their counterparts and, hence, much more high-order-modulation opportunities since they adapt the uVBSs (i.e., cooperative TP sets) to the users' environments. Indeed, according to Tab. III, the occurrence probabilities of 256-QAM and 1024-QAM substantially increase with the proposed approaches and may reach until 37% and 21%, respectively. All these prove again that the new WAV approaches better capitalize on any extra resources made available in any dimension (i.e., temporal, spectral, spatial, modulation/signaling order, etc.) that should characterize 5G and beyond RANs, and that is - a major asset - regardless of the specific radio access technologies to be adopted.

VIII. CONCLUSION

In this paper, we proposed three innovative low-cost clustering approaches that enable user-centric WAV of future 5G network base stations and that provide dynamic, adaptive, and overlapping TP clusters while requiring not only negligible overhead, but also minimum signaling changes at both network and user sides. In contrast to existing clustering techniques, the new ones better leverage 5G features such as extreme densification and massive connectivity as well as new concepts such as mmWave spectrum and massive MIMO. Simulations showed that they may achieve until 154% and 282% of throughput and coverage gains, respectively. Furthermore, these approaches are flexible enough to be adapted to different network dimensions (i.e., space, time, etc.), thereby paving the way for achieving the dramatic performance improvements required by 5G networks to cope with the upcoming mobile data deluge.

REFERENCES

[1] J. G. Andrews *et al.*, "What will 5G be?" *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065–1082, Jun. 2014.

[2] Z. Chang, Z. Zhou, S. Zhou, T. Chen, and T. Ristaniemi, "Towards service-oriented 5G: Virtualizing the networks for everything-as-a-service," *IEEE Access*, vol. 6, pp. 1480–1489, 2017.

[3] I. Ahmad, T. Kumar, M. Liyanage, J. Okwuibe, M. Ylianttila, and A. Gurtov, "Overview of 5G security challenges and solutions," *IEEE Commun. Mag.*, vol. 2, no. 1, pp. 36–43, Mar. 2018.

[4] S. Zaidi, S. Affes, U. Vilaipornsawai, L. Zhang, and P. Zhu, "Wireless access virtualization strategies for future user-centric 5G networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Washington, DC, USA, Dec. 2016, pp. 1–7.

[5] D. C. Mur, P. Flegkas, D. Syrivelis, Q. Wei, and J. Gutiérrez, "5G-XHaul: Enabling scalable virtualization for future 5G transport networks," in *Proc. IEEE 15th Int. Conf. Ubiquitous Comput. Commun. Symp. Cyberspace Secur. (IUCC-CSS)*, Granada, Spain, Dec. 2016, pp. 173–180.

[6] S. Abdelwahab, B. Hamdaoui, M. Guizani, and T. Znati, "Network function virtualization in 5G," *IEEE Commun. Mag.*, vol. 54, no. 4, pp. 84–91, Apr. 2016.

[7] E. J. Kitindi, S. Fu, Y. Jia, A. Kabir, and Y. Wang, "Wireless network virtualization with SDN and C-RAN for 5G networks: Requirements, opportunities, and challenges," *IEEE Access*, vol. 5, pp. 19099–19115, 2017.

[8] S. Zaidi and S. Affes, "Distributed collaborative beamforming in the presence of angular scattering," *IEEE Trans. Commun.*, vol. 62, no. 5, pp. 1668–1680, May 2014.

[9] O. Ben Smida, S. Zaidi, S. Affes, and S. Valaee, "Low-cost robust distributed collaborative beamforming against implementation impairments," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Dec. 2018, pp. 1–7.

[10] O. Ben Smida, S. Zaidi, S. Affes, and S. Valaee, "Robust distributed collaborative beamforming for wireless sensor networks with channel estimation impairments," *Sensors*, vol. 19, p. 1061, Mar. 2019.

[11] S. Zaidi, O. Ben Smida, S. Affes, and S. Valaee, "Distributed zero-forcing AF beamforming for energy-efficient communications in networked smart cities," in *Proc. IEEE PIMRC*, Montreal, QC, Canada, Oct. 2017, pp. 1–7.

[12] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *IEEE Netw.*, vol. 29, no. 3, pp. 68–74, May 2015.

[13] X. Wang *et al.*, "Virtualized cloud radio access network for 5G transport," *IEEE Commun. Mag.*, vol. 55, no. 9, pp. 202–209, Sep. 2017.

[14] K. Liang, L. Zhao, X. Chu, and H. H. Chen, "An integrated architecture for software defined and virtualized radio access networks with fog computing," *IEEE Netw.*, vol. 31, no. 1, pp. 80–87, Jan. 2017.

[15] M. Kalil, A. Al-Dweik, M. F. A. Sharkh, A. Shami, and A. Refaey, "A framework for joint wireless network virtualization and cloud radio access networks for next generation wireless networks," *IEEE Access*, vol. 5, pp. 20814–20827, 2017.

[16] Huawei Technologies, Co. Ltd. Shenzhen, China. (Nov. 2013). *5G: A Technology Vision*. [Online]. Available: www.huawei.com/5Gwhitepaper/

[17] M. M. Rahman, C. Despins, and S. Affes, "Green wireless access virtualization implementation: Cost vs. QoS trade-offs," in *Proc. IEEE ICECCS*, Mangalore, India, Dec. 2014, pp. 79–84.

[18] M. M. Rahman, C. Despins, and S. Affes, "Design optimization of wireless access virtualization based on cost & QoS trade-off utility maximization," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6146–6162, Sep. 2016.

[19] S. Zaidi, B. Hmidet, S. Affes, U. Vilaipornsawai, and L. Zhang, "User-centric wireless access virtualization strategies for future 5G networks," in *Proc. IEEE ICUWB*, Nanjing, China, Oct. 2016, pp. 1–4.

[20] S. Zaidi, O. Ben Smida, S. Affes, U. Vilaipornsawai, L. Zhang, and P. Zhu, "QoS-based virtualization of user equipment in 5G networks," in *Proc. IEEE IWCMC*, Limassol, Cyprus, Jun. 2018, pp. 180–187.

[21] M. Liu, Y. Mao, S. Leng, and S. Mao, "Full-duplex aided user virtualization for mobile edge computing in 5G networks," *IEEE Access*, vol. 6, pp. 2996–3007, 2017.

[22] M. S. Carmo, S. Jardim, A. V. Neto, R. Aguiar, and D. Corujo, "Towards fog-based slice-defined WLAN infrastructures to cope with future 5G use cases," in *Proc. IEEE NCA*, Cambridge, MA, USA, Oct. 2017, pp. 1–5.

[23] J. V. de Belt, H. Ahmadi, and L. E. Doyle, "Defining and surveying wireless link virtualization and wireless network virtualization," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 3, pp. 1603–1627, 3rd Quart., 2017.

[24] J. Zeng, X. Su, J. Gong, L. Rong, and J. Wang, "5G virtualized radio access network approach based on NO stack framework," in *Proc. IEEE ICC*, Paris, France, May 2017, pp. 1–5.

[25] R. Ferrus, O. Sallent, J. P. Romero, and R. Agusti, "On 5G radio access network slicing: Radio interface protocol features and configuration," *IEEE Commun. Mag.*, vol. 56, no. 5, pp. 184–192, May 2018.

[26] X. Wang, C. Xu, G. Zhao, and S. Yu, "Tuna: An efficient and practical scheme for wireless access point in 5G networks virtualization," *IEEE Commun. Lett.*, vol. 22, no. 4, pp. 748–751, Apr. 2018.

[27] O. Krasko, H. Al-Zayadi, V. Pashkevych, H. Kopets, and B. Humeniuk, "Network functions virtualization for flexible deployment of converged optical-wireless access infrastructure," in *Proc. IEEE TCSET*, Slavske, Ukraine, Feb. 2018, pp. 1135–1138.

[28] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, Jr., "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1910–1921, Apr. 2009.

[29] W. Saad, Z. Han, M. Debbah, and A. Hjorungnes, "A distributed coalition formation framework for fair user cooperation in wireless networks," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4580–4593, Sep. 2009.

[30] A. Maaref, J. Ma, M. Salem, H. Baligh, and K. Zarin, "Device-centric radio access virtualization for 5G networks," in *Proc. IEEE Globecom Workshops (GC Wkshps)*, Austin, TX, USA, Dec. 2014, pp. 887–893.

[31] S. Zaidi, S. Affes, M. Azzakhmam, C. Despins, K. Zarifi, and P. Zhu, "Progressive hybrid greyfield wireless access virtualization with leveraged combining of cloud, fog, and legacy RANs," in *Proc. IEEE PIMRC*, Montreal, QC, Canada, Oct. 2017, pp. 1–7.

[32] M. Richart, J. Baliosian, J. Serrat, and J.-L. Gorricho, "Resource slicing in virtual wireless networks: A survey," *IEEE Trans. Netw. Service Manag.*, vol. 13, no. 3, pp. 462–476, Sep. 2016.

[33] P. Han, L. Guo, and Y. Liu, "Green virtual network embedding framework based on zooming small cells in fiber-wireless access network for 5G," in *Proc. IEEE ICTON*, Girona, Spain, Jul. 2017, pp. 1–4.

[34] W. Chen, X. Xu, C. Yuan, J. Liu, and X. Tao, "Virtualized radio resource pre-allocation for QoS based resource efficiency in mobile networks," in *Proc. IEEE GLOBECOM*, Singapore, Dec. 2017, pp. 1–6.

[35] F. Zhou, Y. Wu, R. Q. Hu, Y. Wang, and K. K. Wong, "Energy-efficient NOMA enabled heterogeneous cloud radio access networks," *IEEE Netw.*, vol. 32, no. 2, pp. 152–160, Mar. 2018.

[36] K. Zarifi, H. Baligh, J. Ma, M. Salem, and A. Maaref, "Radio access virtualization: Cell follows user," in *Proc. IEEE PIMRC*, Washington, DC, USA, Sep. 2014, pp. 1381–1385.

[37] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," in *Proc. IEEE Int. Conf. Commun.*, Beijing, China, May 2008, pp. 4033–4037.

[38] X. Chen, Z. Han, H. Zhang, G. Xue, Y. Xiao, and M. Bennis, "Wireless resource scheduling in virtualized radio access networks using stochastic learning," *IEEE Trans. Mobile Comput.*, vol. 17, no. 4, pp. 961–974, Apr. 2018.

[39] J. Gong, S. Zhou, Z. Niu, L. Geng, and M. Zheng, "Joint scheduling and dynamic clustering in downlink cellular networks," in *Proc. IEEE GLOBECOM*, Kathmandu, Nepal, Dec. 2011, pp. 1–5.

[40] S. Fu, H. Wen, and B. Wu, "Power-fractionizing mechanism: Achieving joint user scheduling and power allocation via geometric programming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 3, pp. 2025–2034, Mar. 2018.

[41] S. Zaidi and S. Affes, "Distributed collaborative beamforming design for maximized throughput in interfered and scattered environments," *IEEE Trans. Commun.*, vol. 63, no. 12, pp. 4905–4919, Dec. 2015.

[42] S. Zaidi, B. Hmidet, and S. Affes, "Distributed collaborative beamforming design in highly-scattered environments," in *Proc. IEEE Wireless Commun. Netw. Conf.*, Doha, Qatar, Apr. 2016, pp. 1–7.

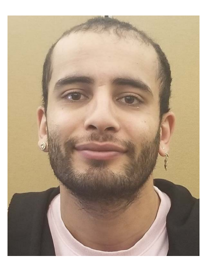
[43] S. Zaidi, O. B. Smida, S. Affes, and S. Valaee, "Distributed zero-forcing amplify-and-forward beamforming for WSN operation in interfered and highly-scattered environments," *IEEE Trans. Commun.*, to be published. doi: 10.1109/TCOMM.2019.2906609.

[44] S. Zaidi and S. Affes, "SNR and throughput analysis of distributed collaborative beamforming in locally-scattered environments," *Wireless Commun. Mobile Comput.*, vol. 12, no. 18, pp. 1620–1633, Dec. 2012.

[45] S. Zaidi, B. Hmidet, and S. Affes, "Power-constrained distributed implementation of SNR-optimal collaborative beamforming in highly-scattered environments," *IEEE Wireless Commun. Lett.*, vol. 4, no. 5, pp. 457–460, Oct. 2015.

[46] R. Ratasuk, D. Tolli, and A. Ghosh, "Carrier aggregation in LTE-advanced," in *Proc. IEEE 71st Veh. Technol. Conf.*, May 2010, pp. 1–5.

Slim Zaidi received the B.Eng. degree (Hons.) in telecommunications from the National Engineering School of Tunis, Tunis, Tunisia, in 2008 and the M.Sc. and Ph.D. degrees (Hons.) in electrical and computer engineering from INRS-EMT, Université du Québec, Montreal, QC, Canada, in 2011 and 2015, respectively. He is currently a Post-Doctoral Fellow of the University of Toronto, Toronto, ON, Canada. He has published over 50 peer-reviewed papers in major IEEE journals and conferences. His research interests include wireless systems (5G NR, LTE/LTE-A, LoRaWAN, NB-IoT, WSNs, and WiFi), radio access virtualization, software-defined networking, machine learning and artificial intelligence, statistical signal and array processing, MIMO, cooperative communications, and millimeter wave communications. He received twice the National Grant of Excellence from the Tunisian Government at both the M.Sc. (2009–2010) and the Ph.D. (2011–2013) programs. He also received the Top-Tier Graduate Ph.D. Scholarships from the Natural Sciences and Engineering Research Council (NSERC) of Canada (2013–2015) and the Fonds de Recherche du Québec Nature et Technologies (FRQNT) (2013–2015). Furthermore, he received two Prestigious Highly Selective Postdoctoral Fellowships from FRQNT (2016–2018) and NSERC (2017–2019). He also regularly serves as a TPC member for top-Notch conferences, including the IEEE PIMRC, IEEE WCNC, and IEEE Globecom and as a Reviewer for IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE ACCESS, and IEEE WIRELESS COMMUNICATIONS LETTERS.



Oussama Ben Smida received the B.Eng. degree (Hons.) in telecommunications from the National Engineering School of Tunis, Tunis, Tunisia, in 2015 and the M.Sc. degree (Hons.) from the Centre Énergie Matériaux Télécommunications, Institut National de la Recherche Scientifique, Université du Québec, Montreal, QC, Canada, in 2017, where he is currently pursuing the Ph.D. degree. His research interests include collaborative beamforming, wireless ad hoc networks, and beyond 4G systems. He received twice the National Grant of Excellence from the Tunisian Government at both the M.Sc. (2015–2017) and the Ph.D. (2017–ongoing) programs. He is also a Laureate of the scholarship of the INRS Academic Foundation Armand-Frappier for francophone students for the M.Sc. degree (2015).

Sofiène Affes (S'95–SM'05) received the Diplôme d'Ingénieur in telecommunications and the Ph.D. degree (Hons.) in signal processing from Télécom ParisTech (ENST), Paris, France, in 1992 and 1995, respectively. He was a Research Associate with INRS, Montreal, QC, Canada, until 1997; an Assistant Professor until 2000; and an Associate Professor until 2009. He is currently a Full Professor and the Director at INRS of PERWADE, a unique M$4 million research-training program on wireless in Canada involving 27 partners from eight universities and 10 industrial organizations. He has been twice a recipient of the Discovery Accelerator Supplement Award from NSERC, from 2008 to 2011 and from 2013 to 2016. From 2003 to 2013, he was the Canada Research Chair in Wireless Communications. Since 2017, he holds the Cyrille-Duquet Research Chair in telecommunications. In 2006, 2015, and 2017, he has served as the General Co-Chair or Chair of the 64th IEEE VTC'2006-Fall, the 15th IEEE ICUWB'2015, and the 28th IEEE PIMRC'2017 jointly with the co-located 28th IEEE 5G Summit, respectively, all held in Montreal, QC, Canada. In 2008 and 2015, he received the IEEE VTC Chair Recognition Award from the IEEE VTS and IEEE ICUWB Chair Recognition Certificate from the IEEE MTT-S for the exemplary contributions to the success of both events, respectively. He previously served as an Associate Editor for the IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS, IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE TRANSACTIONS ON SIGNAL PROCESSING, the *Journal of Electrical and Computer Engineering* (Hindawi), and the *Journal of Wireless Communications and Mobile Computing* (Wiley). He currently serves as a member of the Editorial Board of the *Sensors* (MDPI) and the Advisory Board of the multidisciplinary journal *Sci* (MDPI).

Usa Vilaipornsawai received the B.Eng. degree from the King Mongkut University of Technology Thonburi, Thailand, in 1997, the M.Sc. degree from Concordia University, in 2001, and the Ph.D. degree from McGill University, Montreal, QC, Canada, in 2009.

From 2009 to 2011, she was a Research Fellow of the Institute for Systems and Robotics, University of Algarve, Faro, Portugal, working on underwater acoustic communications, for the European FP7-ICT funded Underwater Acoustic Network (UAN) Project. Since 2012, she has been with Huawei Ottawa Research and Development Centre, where she was involved in 5G wireless system research and standard specification development. Her research interests include multi-TRP cooperation, ultra-reliable low-latency communication, non-orthogonal multiple access, spectrally-contained waveforms, channel coding, receiver design for MIMO systems, and adaptive channel equalization.

Liqing Zhang received the B.E. and M.E. degrees from the Department of Electrical Engineering, University of Science and Technology of China (USTC), China, in 1986 and 1989, respectively, and the Ph.D. degree from the Department of Electrical Engineering, McGill University, Canada, in 2003.

He was a Lecturer with USTC, from 1989 to 1994; a visiting Research Fellow of McGill University, in 1995 and a Research Associate with McGill University, from 1996 to 1999. He was with Nortel Network, Ottawa, Canada, from 2000 to 2007 and with SOMA network, Ottawa, Canada, from 2008 to 2009, as a Technical Lead and System Architect for radio access system design on UMTS/CDMA2000, WiMAX, LTE (2.5G/3G/4G) networks. Since 2010, he has been with the Canada Research Centre (CRC) of Huawei Technologies, Ottawa, Canada, where he currently holds the position of Senior Staff Engineer and LTE/NR Design Expert, focusing on the 5G radio access research and standardization. His research interests lie within the realm of wireless communications, including communication and information theories, signal processing, PHY/MAC cross-layer design and optimization, grant-free and non-orthogonal multiple access, cooperative communications, interference cancellation, and co-ordination in heterogeneous networks. For the 5G research, he has coauthored the chapter of Grant-Free Multiple Access Scheme in the book *Multiple Access Techniques for 5G Wireless Networks and Beyond* (Springer, 2018) (Editors: Mojtaba Vaezi, Zhiguo Ding, and H. Vincent Poor) and holds over 50 U.S. awarded patents and pending patent applications. He is a recipient of the Outstanding Inventor Award at Huawei CRC in 2017.

Peiying Zhu (F'18) received the M.Sc. degree from Southeast University in 1985 and the Ph.D. degree from Concordia University in 1993.

Prior to joining Huawei in 2009, she was a Nortel Fellow and the Director of the Advanced Wireless Access Technology at the Nortel Wireless Technology Lab. She led the team and pioneered research and prototyping on MIMO-OFDM and Multi-hop relay. Many of these technologies developed by the team have been adopted into the LTE standards and 4G products. She is currently leading the 5G Wireless System Research in Huawei. The focus of her research is advanced wireless access technologies with over 200 granted patents. She has been regularly giving talks and panel discussions on 5G vision and enabling technologies. She is actively involved in 3GPP and IEEE 802 Standards Development.

Dr. Zhu is an Huawei Fellow. She is currently a WiFi Alliance Board Member. She has served as a Guest Editor for the IEEE Signal Processing Magazine Special Issue on the 5G revolution and the IEEE JSAC on Deployment Issues and Performance Challenges for 5G. She has co-chaired various 5G workshops in the IEEE GLOBECOM.