Wireless Access Virtualization Strategies for Future User-Centric 5G Networks

Slim Zaidi*, Sofiène Affes*, Usa Vilaipornsawai†, Liqing Zhang†, and Peiying Zhu†
*INRS-EMT, Université du Québec, Canada, Emails: {zaidi,affes}@emt.inrs.ca
†Huawei Technologies Canada Co. Ltd., Canada, Emails: {usa.vilaipornsawai,liqing.zhang,peiying.zhu}@huawei.com

***Abstract*—User-centric wireless access virtualization (WAV) allows each user to be served by a set of carefully selected transmission points (TP)s 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 5G networks. This fundamental structural 5G evolution alongside the ultra-dense multi-tier heterogenous context foreseen in such networks require an inevitable rethinking of efficient scalable TPs clustering. As such, this paper proposes three innovative low-cost clustering approaches that enable user-centric WAV and provide dynamic, adaptive, and overlapping TPs clusters while requiring not only negligible overhead and power costs, but also minimum signaling changes at both network and user sides. In contrast to existing clustering techniques, the new ones we propose better leverage 5G features such as extreme densification and massive connectivity as well as new concepts such as mmWave spectrum and massive 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.**

***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 4G radio access networks (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]. Significant research endeavors have been devoted to developing some remedial solutions to this cell-edge effect issue such as inter-cell interference coordination, coordinated beamforming, and fractional frequency reuse. Although the latter offered some performance gains at the cost of increased complexity and overhead, they were unable to completely remove the cell-boundary effect.

Using wireless access virtualization (WAV), future 5G networks will capitalize, in contrast to their predecessors, on both the extreme densification and massive connectivity to provide boundaryless communications. This would potentially leads 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 [2]. Indeed, with WAV, the coverage is built around user, making it the network's focal point rather than the cell as is the case in current cell-centric RANs. The network will then adapt the data transmission to each user's environment and quality of service (QoS) requirements, thereby creating the illusion of a moving virtual cell following the latter. 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. Practically, this will be done through enabling each user to be served by a set of carefully and optimally selected transmission points (TP)s forming a user-specific virtual base-station (uVBS), making TPs' clustering crucial to any user-centric WAV strategy.

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 uVBSs. This goal cannot actually be achieved without forsaking the conventional clustering approaches aiming to form TP sets using solely system information, i.e., TPs' positions and density, their available resources, etc. Although it does not incur significant costs, in terms of complexity, overhead, latency, and power consumption, since its resulting TP sets are predetermined and rarely updated (i.e., static), such an approach often achieves poor performance in terms of both throughput and spectral efficiency [3]. This is mainly due to the fact that these sets are not adapted to the highly changing users' environments owing to the lack of user-side information such as the user's channel state information (CSI), channel quality indicator (CQI), signal-to-interference-plus-noise-ratio (SINR), etc.

Many research groups have focused then on developing dynamic adaptive clustering approaches [4]-[8]. Exploiting the users' CSIs and/or received SINRs, the latter dynamically adapt the TPs sets to each user's environment and QoS requirements. As the user moves, its serving set is updated

Work 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.

User refers here to devices (i.e., smartphones, sensors, etc.), vehicles, or machines connected to the network.

978-1-5090-2482-7/16/$31.00 ©2016 IEEE

by dropping some TPs and/or adding others and, hence, much better performance is achieved. 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. 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 TP sets are usually formed using highly-complex iterative greedy algorithms that explore all potential set constructions to ultimately settle on network partitions that very often far from optimal. Besides, in order to completely remove the cell-edge effect, TP sets must overlap [8]. 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 sets 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 to their high complexity and cost, most dynamic clustering techniques suffer from another drawback that may also hinder their implementation in virtualized 5G RAN. Indeed, they often cause a considerable discrepancy between traffic loads at different TPs should they lack key system information such as TPs' density, available resources, etc. during their design. Among the few attempts to overcome such an issue, we found the pioneering work of Zarifi et al. [8] which has developed a dynamic clustering approach able of balancing the traffic load among different TPs in the user-centric WAV context. By accounting for the TPs' loads when forming the users' serving TP sets, the approach in [8] significantly improves dynamic clustering. This comes, however, again at the cost of increased complexity.

To summarize, so far, two different TP clustering approaches exist: 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 enable user-centric WAV, this work aims to develop a *best-of-the-two-worlds* clustering technique that combines these approaches' benefits while avoiding their drawbacks.

In this paper, we propose three innovative low-cost clustering approaches that enable user-centric WAV and provide dynamic, adaptive, and overlapping TPs clusters while requiring not only negligible overhead and power costs, but also minimum signaling changes at both network and user sides. In contrast to existing clustering techniques, the new ones we propose better leverage 5G features such as extreme densification and massive connectivity as well as new concepts such as mmWave spectrum and massive 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. TPs are equipped each with K antennas while users are assumed, for the sole sake of simplicity, to have a single antenna. We assume that all users are actively communicating with the network during TP clustering.

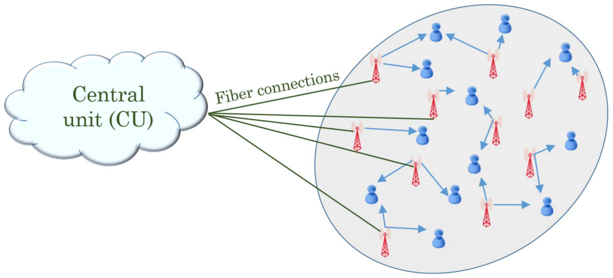


Fig. 1. System model.

III. Proposed user-centric WAV approaches

In this section, we propose three innovative clustering approaches aiming to enable WAV.

A. Approach 1

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

$$P_{\max}^k = \max\left\{P_{i-k}, i = 1, \ldots, M\right\}, \qquad (1)$$

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, etc. 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):

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

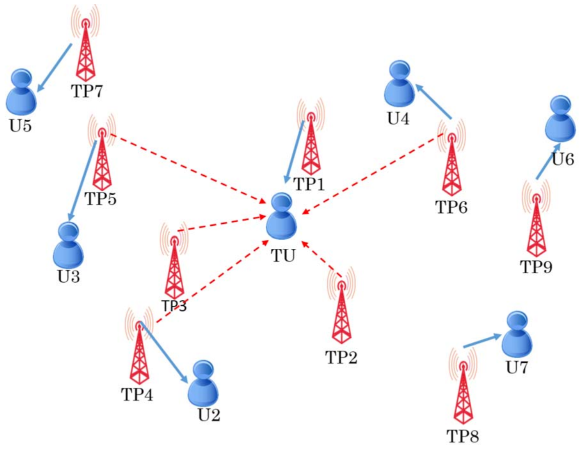


Fig. 2. Single-serving TP selection.

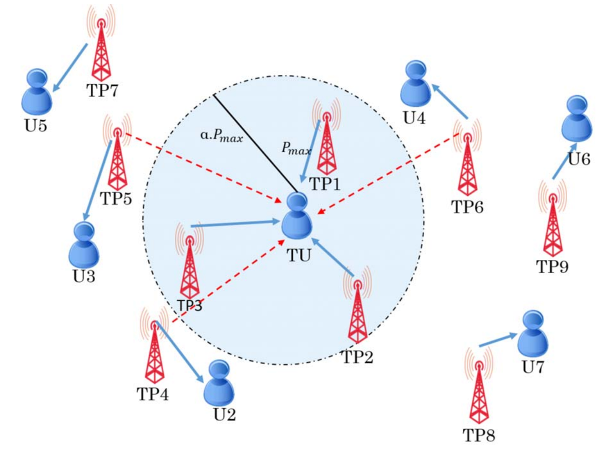


Fig. 3. Approach 1.

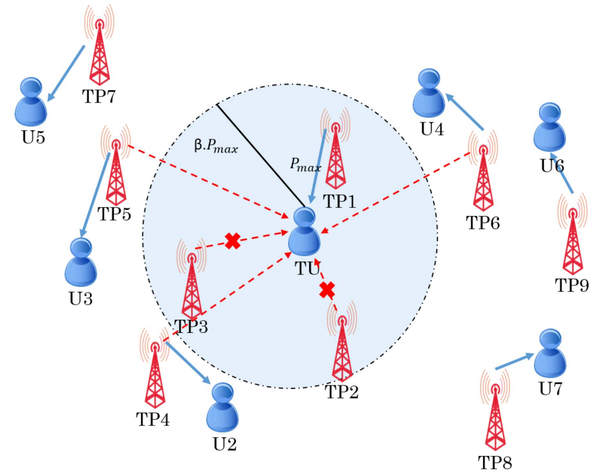


Fig. 4. Approach 2.

In other words, using Approach 1, any TP whose RSRP at the k-th user is large enough to be in the interval $\left[\alpha P^k_{\max}, P^k_{\max}\right]$ will serve this user. 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 TU's SC and better will be its throughput. Nevertheless, it is not possible to indefinitely decrease α without affecting the performance of other users. Indeed, when a number of TPs are solicited by another user to jointly transmit its data, they would have increasingly limited resources left for allocation to the latter. As α decreases, more TPs are solicited and, hence, more resources are dedicated to a smaller number of users. It becomes much more likely then 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 not only optimal system performance, but also optimal resource utilization. Computation of this parameter will be further discussed 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) defined as

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

where β is the system parameter broadcasted by the CU. The other major difference worth underlining here between Approaches 1 and 2 is the CU broadcasts both the k-th user's data and CSIs to the TPs in SC_k, in the first, while, in the second, it broadcasts only the CSIs to the TPs in NC_k. 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. Computation of this parameter will be further discussed in Section IV.

C. Approach 3

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

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

and

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

This means that the k-th user will send serving requests to the TPs with the high RSRPs (i.e., $P_{i-k} \in \left[\alpha P^k_{\max}, P^k_{\max}\right]$) and nulling ones to those with moderate RSRPs (i.e., $P_{i-k} \in \left[\beta\alpha P^k_{\max}, \alpha P^k_{\max}\right[$) 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

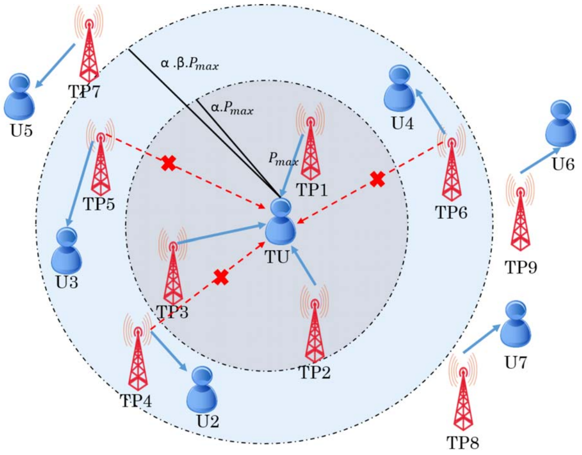


Fig. 5. Approach 3.

investigate the available methods able to compute such parameters. Some of them are listed and 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**: The parameters are obtained offline by optimizing them heuristically for different network setups (i.e., different TPs and user's densities).
- **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 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 place 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. Besides the spatial dimension, one may also exploit temporal dimension 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)

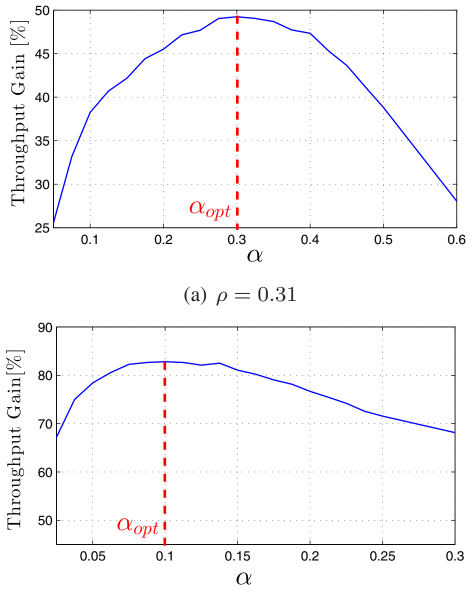
If this option is adopted, each user selects its own TP cluster(s) based on Approach 1, 2, or 3 and feedbacks only the RSRPs of the TPs in SC and/or NC to the network. Option 1 avoids then the feedback of the non-selected TPs' RSRPs and, hence, substantially reduces the overhead and power costs with respect to the clustering strategies so far existing. However, it requires that the network broadcasts α and/or β. This obviously neither burdens the complexity of the proposed WAV strategy nor the network power and overhead costs. Another advantage of Option 1 is that it allows refinement or overwriting of each user's TP cluster(s) based on the traffic load, certain network conditions, users' priority, and Qos requirements, etc.

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 reducing even more 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 not able to overwrite users' TP clusters to cope with certain conditions and/or users' priority, and 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



(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 TPs-users densities ratio ρ.

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 to the first versions of 5G networks.

Once again, we have flexibility in the choice of one the above mechanisms. Indeed, different options could be used with different subnetworks, at different periods, and/or for different applications and services, etc. 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 latter, 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 simulations and/or calibration as discussed previously. In contrast to the clustering approaches thus far existing, we avoid the implementation of highly-complex iterative greedy, yet often sub-optimal algorithms that incur prohibitive latency, overhead, and power costs.
- **Dynamic, adaptive**: With our approaches, the TP clusters are formed from overlapping sets whose cardinalities (i.e., the number of TPs in each set) are adapted to the users' situations and environments. As one example, using Approach 1, more (less) serving TPs are associated with a user when it is subject to high (low) interference.
- **Low overhead, latency, and power costs**: 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/CQIs to be able to form the TPs sets. Hence, significant overhead, latency, and power saving can be obtained with the developed approaches.
- **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 allows users to share the same resources as well as on new concepts 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 advantages of both concepts. This is in contrast with existing clustering techniques whose complexity increases exponentially with such technologies.
- **Flexibility**: By associating different parameters to the different network dimensions, our approaches pave the way towards dramatic improvements in both spectral and power efficiencies. Indeed, **user-class**-, **service**-, and **application**-based parameters allows adequate adaptation of the allocated resources to the 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.

VII. Simulations Results

In this section, system-level simulations are conducted to verify the efficiency of the proposed approaches. In order to highlight the gains they provide, 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 and vice-versa. In such a case, the peak data rate is approximately 712 Kbits/s according to [9]. 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, and a channel bandwidth of 10 MHz. We also consider that users, initially (i.e., at $t = 0$), are 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. In this work, maximum ratio transmission (MRT) is employed by SC TPs to jointly transmit the user's data while zero-forcing beamforming is implemented by NC TPs 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 techniques [10].

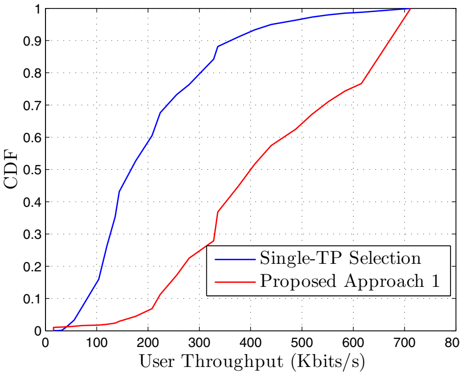


Fig. 7. CDFs of the user throughput achieved by Approach 1 and single-serving TP selection when $\alpha_{\text{opt}} = 0.1$ and $\rho = 0.44$.

Fig. 6 plots the achieved network throughput gain of Approach 1 over single-serving TP selection versus α for different values of the TPs-users densities ratio ρ. 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 level α_{opt} of the parameter α. We also observe that α_{opt} depends on ρ. Indeed, it increases when the latter decreases and vice-versa. This is hardly surprising since the available resources per user increases 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, thereby decreasing α_{opt}. For instance, we find that $\alpha_{\text{opt}} = 0.3$ when $\rho = 0.31$ whereas $\alpha_{\text{opt}} = 0.1$ when $\rho = 0.44$. In these cases, Approach 1 achieves throughput gains as high as 49% and 83%, respectively.

Fig. 7 plots the CDFs of the user throughput achieved by Approach 1 and single-serving TP selection. With Approach 1, the throughput achieved by 40% of the users exceeds 450 Kbits/s (i.e., approximatively 65% of the peak data rate) while only 5% of users reach the same throughput level with single-serving TP selection. This proves the efficiency of the proposed approach and highlights the dramatic performance improvements it may provide at low complexity, latency, overhead, and power costs, making it an interesting candidate for future 5G networks.

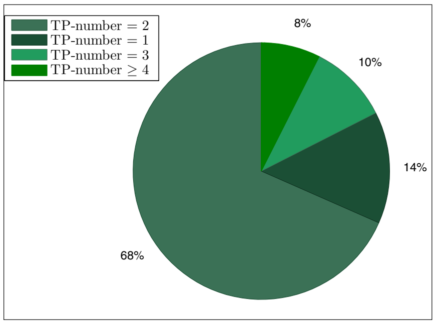
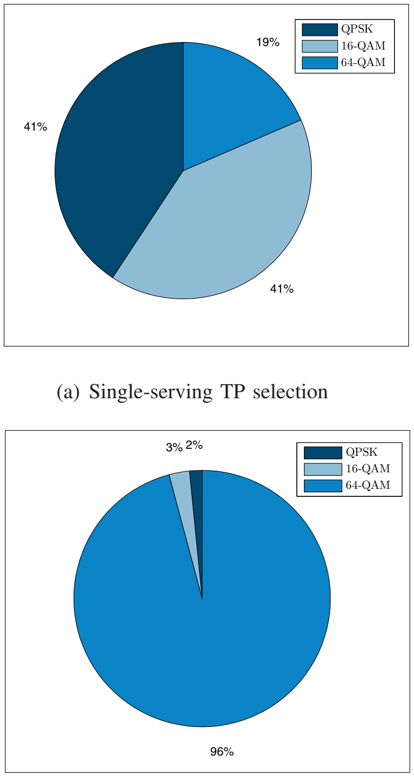


Fig. 8. Pie chart of user's serving TPs number for $\alpha_{\text{opt}} = 0.1$ and $\rho = 0.44$.



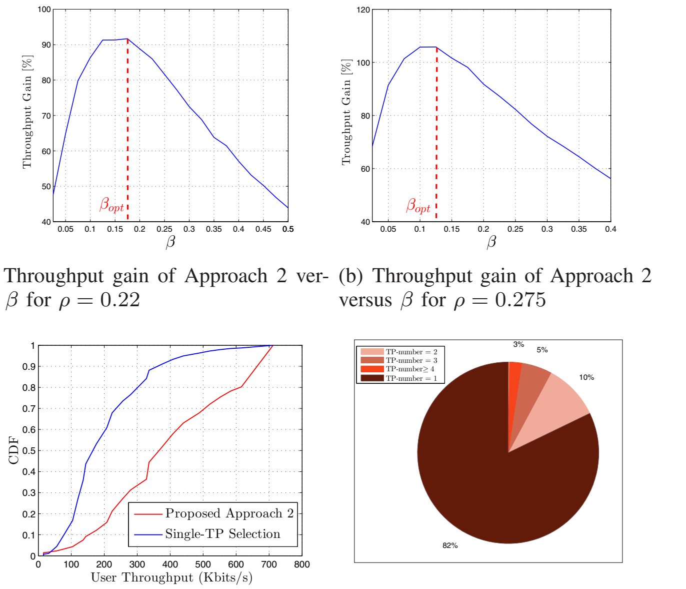
(a) Single-serving TP selection

(b) Approach 1 with $\alpha_{\text{opt}} = 0.1$ and $\rho = 0.44$

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

Fig. 8 illustrates the pie chart of the user's serving TPs number with $\alpha_{\text{opt}} = 0.1$ and $\rho = 0.44$. From this figure, 14% of the users are served by a single TP whereas 68% of them are simultaneously served by two TPs, 10% by three TPs, and the rest (about 8%) by four TPs or more. Hence, in most cases, the user's SC cardinality does not exceed three and as such does not burden the network virtualization cost. Again, this very desirable feature makes the proposed WAV approach an interesting candidate for future 5G networks.

Fig. 9 shows the occurrence probabilities of QPSK, 16-QAM, and 64-QAM when Approach 1 and single-serving TP selection are used. With Approach 1, 64-QAM occurs 96% of the time against 19% with single-serving TP selection. This is expected since Approach 1 offers a dramatic SINR



(a) Throughput gain of Approach 2 versus β for $\rho = 0.22$

(b) Throughput gain of Approach 2 versus β for $\rho = 0.275$

(c) User throughput CDFs for $\beta_{\text{opt}} = 0.175$ and $\rho = 0.22$

(d) Pie chart of user's nulling TPs number for $\beta_{\text{opt}} = 0.175$ and $\rho = 0.22$

Fig. 10. The performance of Approach 2.

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.

Fig. 10 evaluates the performance of Approach 2. In Figs. 10(a) and 10(b), we plot the throughput gain achieved by this approach over single-serving TP selection versus β when the number of users per macro-TP is 50 and 40, respectively. These figures confirm the existence of an optimum value β_{opt} that depends on ρ. The user throughput CDFs in Fig. 10(c) confirms the significant superiority of Approach 2 over single-serving TP selection. Fig. 10(d) suggests that the optimal throughput gain of Approach 2 can be achieved with 82% of the users requesting only a single nulling TP. All these results underline once again the great potential of the proposed WAV approaches in enabling future 5G networks.

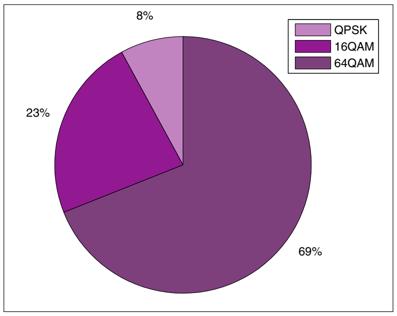


Fig. 11. Occurrence probabilities of the QPSK, 16-QAM, and 64-QAM modulations with $\beta_{\text{opt}} = 0.175$ and $\rho = 0.22$.

Fig. 11 shows the occurrence probabilities of QPSK, 16-QAM, and 64-QAM modulations with Approach 2 and suggests that 64-QAM occurs as often as 69% 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 during the transmission phase.

Due to the lack of space, we are obviously unable to include the numerical results of Approach 3 which potentially outperforms both Approaches 1 and 2. However, we will report on them in future publications.

VIII. Conclusion

In this paper, we proposed three innovative low-cost clustering approaches that enable user-centric WAV and provide dynamic, adaptive, and overlapping TPs clusters while requiring not only negligible overhead and power costs, but also minimum signaling changes at both network and user sides. In contrast to existing clustering techniques, the new ones we propose better leverage 5G features such as extreme densification and massive connectivity as well as new concepts such as mmWave spectrum and massive 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.

References

[1] J. G. Andrews, S. Buzzi, W. Choi, S. Hanly, A. Lozano, A. C. K. Soong, and J. C. Zhang, "What will 5G be?," *IEEE J. Select. Areas Commun.*, vol. 32, pp. 1065-1082, June 2014.

[2] "5G: A technology vision," White Paper, Huawei Technologies, Co. Ltd., Nov. 2013. [Online]. Available: www.huawei.com/5Gwhitepaper/

[3] J. Zhang, R. Chen, J. G. Andrews, A. Ghosh, and R. W. Heath, "Networked MIMO with clustered linear precoding," *IEEE Trans. Wireless Commun.*, vol. 8, pp. 1910-1921, Apr. 2009.

[4] A. Papadogiannis, D. Gesbert, and E. Hardouin, "A dynamic clustering approach in wireless networks with multi-cell cooperative processing," *Proc. IEEE ICC'2008,* Beijing, China, May 19-23, 2008.

[5] J. Gong, S. Zhou, Z. Niu, L. Geng, and M. Zheng, "Joint scheduling and dynamic clustering in downlink cellular networks," *Proc. IEEE GLOBECOM'2011,* Houston, TX, USA, Dec. 5-9, 2011.

[6] 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, pp. 4580-4593, Sep. 2009.

[7] A. Maaref, J. Ma, M. Salem, H. Baligh, and K. Zarifi, "Device-centric radio access virtualization for 5G Networks," *Proc. IEEE GLOBECOM'2014,* Austin, TX, USA, Dec. 8-12, 2014.

[8] K. Zarifi, H. Baligh, J. Ma, M. Salem, and A. Maaref, "Radio access virtualization: cell follows user," *Proc. IEEE PIMRC'2014,* Washington DC, USA, Sep. 2-5, 2014.

[9] 3GPP, "3GPP TS 36.213 V9.2.0: Evolved universal terrestrial radio access network (E-UTRA); physical layer procedures," June 2010.

[10] S. Zaidi and S. Affes, "Distributed collaborative beamforming design for maximized throughput in interfered and scattered environments,“ *IEEE Trans. Commun.,* vol. 63, pp. 4905-4919, Dec. 2015.