Next-Generation MmWave Small Cell Networks: Multiple Access, Caching and Resource Management

Jingjing Cui, *Member, IEEE,* Yuanwei Liu, *Member, IEEE,*
Zhiguo Ding, *Senior Member, IEEE,* Pingzhi Fan, *Fellow, IEEE*,
Arumugam Nallanathan, *Fellow, IEEE* and Lajos Hanzo, *Fellow, IEEE,*

***Abstract*—Millimeter wave (mmWave) small cells have been considered as an effective technique of significantly improving the data rates of future networks. More particularly, this article investigates the potential benefits of mmWave small cell networks from the perspective of non-orthogonal multiple access (NOMA) and wireless caching. We highlight a range of innovative resource management solutions conceived for mmWave small cell networks by invoking adaptive learning. Finally, several promising future research directions of mmWave small cell networks are identified.**

I. Introduction

In order to meet the explosive increase in the volume of mobile traffic over the coming decade, new solutions have to be conceived for addressing future challenges. Given the availability of large bandwidths, millimeter wave (mmWave) solutions may find their way into next generation networks. To support the ever-growing mobile traffic demand and massive connectivity required, the combination of mmWave techniques and network densification has been considered as a potential future candidate [1].

MmWave small cell networks are generally different from the systems used at lower frequencies owing to their advanced radio technologies. One reason is that the short millimeter wavelength allows large numbers of antennas to be packed into compact form factors [2], which supports the much needed highly directional transmission to compensate for the high path loss. Moreover, the unique propagation conditions at mmWave frequencies impose fundamental challenges on mmWave small cell systems. As a result, new system concepts and architectures are required for efficiently exploiting these characteristics. The goal of this article is to provide a comprehensive overview of mmWave small cell networks in terms of their multiple access, resource management and caching, which is motivated by the exploration of emerging technologies for improving the spectral efficiency. For instance, non-orthogonal multiple access (NOMA)-aided mmWave small cell networks are capable of providing multiplexing gains by encouraging multiple users to share the same resource block. In addition, another application of cache-enabled mmWave small cell networks is to exploit the benefits of memory for reducing the network's tele-traffic. Moreover, advanced resource management techniques are capable of enhancing the network capacity and the fairness by efficient algorithmic designs. By jointly designing the resource allocation in terms of subchannel assignments, user scheduling and power allocation with the aid of machine learning tools [3], one can achieve a near-optimal resource management [4]. Indeed the application of mmWave small cells attains potential benefits, but there are still substantial research challenges, which motivates us to contribute this article.

J. Cui and P. Fan are with the Institute of Mobile Communications, Southwest Jiaotong University, Chengdu 610031, P. R. China. (email: cuijingj@foxmail.com, p.fan@ieee.org). J. Cui is also with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K

Y. Liu and A. Nallanathan are with the School of Electronic Engineering and Computer Science, Queen Mary University of London, London E1 4NS, U.K. (email: {yuanwei.liu, arumugam.nallanathan}@qmul.ac.uk).

Z. Ding is with the School of Electrical and Electronic Engineering, The University of Manchester, Manchester, M13 9PL, UK. (e-mail: zhiguo.ding@manchester.ac.uk).

L. Hanzo is with the School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K. (e-mail: lh@ecs.soton.ac.uk).

L. Hanzo would like to acknowledge the ERC's financial support of the Advanced Fellow Award.

II. Key Features of MmWave Small Cell Networks

In conventional small cell networks, substantial inter-cell interference is encountered. The configuration of mmWave small cell networks is more challenging than that of the classic sub-2GHz networks owing to their pre-dominantly Line-of-Sight (LOS) wideband transmissions. The key features of future mmWave small cell networks have to be harmonized with their propagation model and network architecture, which will be discussed in the following subsections.

A. From Sub-6 GHz to MmWave Band

In contrast to the propagation encountered in traditional wireless communication in the sub-6 GHz band, propagation in the mmWave band takes place between 30 and 300 GHz [5]. As a consequence, the mmWave solutions are expected to have: 1) highly directional transmission to compensate for the path loss, which has the benefit of increasing the number of users served in small cells, 2) low wall-penetration and consequently high signal attenuation, hence reducing the inter-cell interference, 3) high bandwidth resulting in a high data rate for the users.

The channel impulse response (CIR) of mmWave channels is expected to exhibit limited scattering, thus it tends to have a sparse angular domain response [5]. Based on the well-established mmWave transmission model, the channel vector associated with a uniform linear array (ULA) between a multiple-antenna aided BS and a single-antenna aided user can be expressed as $\mathbf{h} = \sum_{l=0}^{L} \mathbf{a}(\theta_l) \frac{\alpha_l}{\sqrt{L(1+d^{\eta_l})}}$, where $l = 0$ denotes the line-of-sight (LoS) path and L is the total number of non-line-of-sight (NLoS) paths, while d denotes the distance between the BS and the user. Furthermore, η_0 and η_l, $l = 1, \cdots, L$, denotes the path loss exponents corresponding to the LoS and NLoS paths, respectively. Moreover, α_l is the complex-valued gain of the l-th path of the user, and $\mathbf{a}(\theta_l)$ is the antenna array response vector of the BS with $\theta_l \in [-1, 1]$ being the normalized direction of the l-th path. Finally, θ_l is a function of the physical angle of departure, the signal wavelength and the distance between antenna elements.

B. Architecture of MmWave Small Cell Networks

Again, mmWave small cell networks will operate in a different manner from traditional sub-6 GHz networks, since they rely on hybrid analog/digital beamforming. Recent channel measurements have shown that due to the high path loss, mmWave transmission is only capable of achieving a range of about 150-200 meters by using highly directional beamforming [5]. As a consequence, it is desired that mmWave transmissions co-exist with a traditional sub-6 GHz cellular network capable of providing a wide area coverage for avoiding excessively frequent handovers.

Fig. 1 illustrates the system architecture of mmWave small cell networks relying both on NOMA and caching, where the macro BS operates at sub-6 GHz frequencies and the small cell BSs operate in the mmWave frequency band. As illustrated by Fig. 1, NOMA increases the number of users supported beyond the number of resource slots available, while caching offloads the core network's traffic. In such networks, mmWave small cells are deployed in ultra-dense scenarios to support a large number of connections, where resource management becomes a key issue of improving the network's utility.

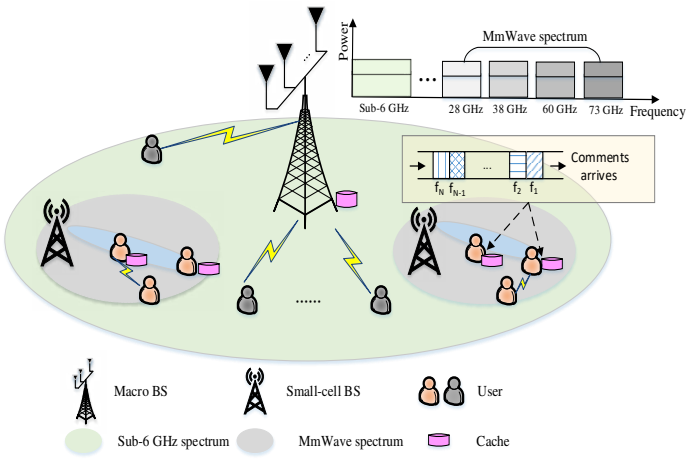


Figure 1: MmWave small cell network that can support the NOMA transmission and has the capability of caching.

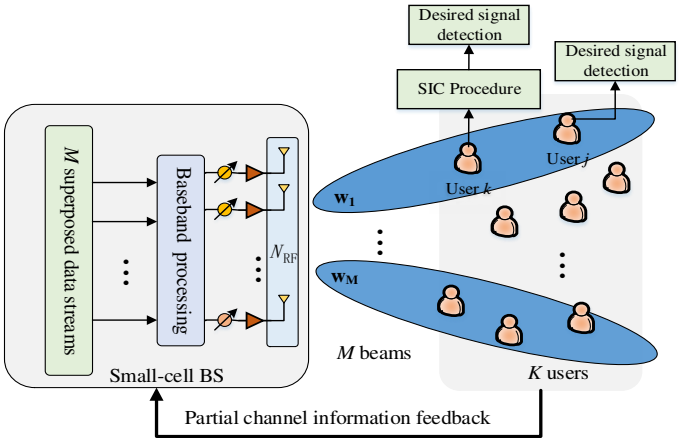


Figure 2: System model for mmWave-NOMA transmission in downlink multiple input and single output (MISO) scenarios.

III. New Multiple Access Techniques for MmWave Small Cell Networks

Next-generation mmWave small cell networks are expected to support massive connectivity of wireless devices. Hence the NOMA principle may be exploited, which supports multiple users in each time/frequency resource-slot by distinguishing them in the power domain [6], [7]. As a consequence, NOMA-aided mmWave small cells are eminently suitable for supporting massive connectivity and for meeting the users' diverse service requirements. In contrast to the conventional orthogonal multiple access (OMA)-aided mmWave small cells, multiple users can also share the same beam simultaneously in NOMA-assisted mmWave small cells. In order to better illustrate the structure of NOMA-aided mmWave small cell systems, we consider a NOMA-aided mmWave downlink (DL) scenario as our example. As shown in Fig. 2, the BS performs multiple input and multiple output (MIMO) transmission relying on M beams, where multiple users can be served by a single beam at different power levels. Specifically, the BS will simultaneously send M superposed data streams for the K users relying on M clusters. As for the users in the same cluster, those having better channel conditions (better users) shall perform SIC for removing the intra-cluster interference from the user having weaker channel conditions (weaker users). Depending on the power domain multiplexing principle in NOMA, the signals from the weaker users is remodulated and subtracted from the composite multi-user signal at the better users, hence automatically leaving behind the signal of the users having better channel conditions.

It is worth pointing out that the hybrid analog/digital beamforming design becomes more challenging in NOMA-aided mmWave transmissions due to the massive scale of connectivity. As a result, inter-beam interference will contaminate the signal of each clusters, which makes the structure of the

NOMA-aided mmWave small cells advocated rather different from that of the conventional NOMA-aided single input and single output (SISO) or multi-carrier systems [6], [8]. Let us consider the beamformer weights $\mathbf{w}_1$ supporting User k and User j as an example in Fig. 2. To reduce the beamformer's feedback overhead, we assume that only partial channel state information (CSI) feedback is adopted for the mmWave-NOMA system of Fig. 2. Note that depending on the specific communication scenarios encountered, the terminology of partial CSI may represent partial knowledge of the channel gains, of the distance of the users or the angle of arrival of the users' channels, as discussed for example in [9]. Let us denote the channel's column vectors by $\mathbf{h}_k$ and $\mathbf{h}_j$ valid for User k and User j, respectively. Let us assume furthermore that the equivalent channel gains of User k and User j satisfy the condition of $\frac{|\mathbf{h}_k\mathbf{w}_1|}{\sum_{n=2}^{M}|\mathbf{h}_k\mathbf{w}_n|p_n+\sigma^2} \geq \frac{|\mathbf{h}_j\mathbf{w}_1|}{\sum_{n=2}^{M}|\mathbf{h}_j\mathbf{w}_n|p_n+\sigma^2}$, where $p_n, n = 1, \cdots, M$, denotes the power allocated to the n-th beam and σ^2 denotes the noise power. As shown in Fig. 2, to obtain a potential gain for NOMA in mmWave small cells, User k expects to decode the signal of User j and to subtract it before decoding its desired signal.

A remarkable advantage of this mmWave-NOMA design is that the number of users can be much higher than that of the radio frequency (RF) chains. This is beneficial, since the number of RF chains is limited due to their high hardware costs. Nevertheless, conceiving sophisticated hybrid beamforming designs for mmWave-NOMA systems still requires further research efforts, accounting for the coupling between hybrid beamformers and the SIC decoding orders of NOMA users [10].

IV. Cache-Enabled MmWave Small Cell Communications

MmWave small cell networks naturally support ultra-dense deployments and ultra-high data rates, given rich spectral resources. However, the tele-traffic of small BSs is often limited by the capacity of backhaul links, which represent the connection between the small BSs and the core network. Wireless caching is capable of alleviating the backhaul burden as well as reducing both the delay and the energy consumption [11]. For maximizing the offloading benefits of cache-enabled small cell mmWave communications, it should be judiciously decided on what to cache, where to cache and how to cache. Motivated by this, we will discuss these three issues in the following subsections.

A. What to Cache

A cache-enabled mmWave cellular network relying on D2D communications is shown in Fig. 1. Each user has a local cache and can invoke short-range communications to share cached files. During off-peak time, the users prefetch popular files for storing in their local caches. This is capable of substantially reducing the average delay, while mitigating the network's traffic during the peak time. Due to the limited cache size, it may become difficult for a single user to cache all popular files in their storage. In order to improve the caching performance attained, each user has to decide what contents to cache. Generally, the long-term popularity of contents provides an accurate reflection of the users' requests. The impact of the caching memory size on the density of small cell BSs was investigated in [12] by assuming that the BSs only store the most popular contents. For a static content catalogue, the Zipf distribution provides a good model to capture the asymptotic properties of the content popularity. In practice, due to the dynamically fluctuating features of the popularity of contents, an accurate prediction model is helpful for designing adaptive caching policies. Big data processing and machine learning techniques are capable of capturing the dynamics of the popularity by relying on historical data, which can enhance the accuracy of the prediction model. A machine learning aided caching policy was developed in [13] by learning the user's preference. Moreover, the contents associated with similar applications often tend to be correlated, as exemplified by still pictures found by image recognition and articles on topical subjects, which encourage the system to cache contents, where possible. A key problem in the context of enhancing the caching efficiency is how to characterize the contents. For instance, for the contents in a specific virtual reality game, the specific image semantics becomes important for image retrieval. Due to the limited time available for caching and owing to the stringent requirements of low-latency transmissions, the specific types of content the system should cache depends on the practical application.

B. Where to Cache

Naturally, it is preferable to aim for a high data rate by exploiting the substantial bandwidth reserves of the mmWave frequency bands coupled with the high area-spectral efficiency (ASE) of small cells. Therefore, adopting mmWave carriers for cache-enabled networks is appealing for further enhancing the throughput. When the content is stored in small BSs, this so-called femto-caching circumvents the backhaul constraints of small cells. In contrast to femto-caching, caching at the user devices allows the subscribers to exchange their cached contents through device to device (D2D) communications, which has the potential of significantly enhancing both the connectivity and the spectral efficiency by relieving the BSs. To elaborate on cache-enabled D2D mmWave communications, we consider the simple downlink transmission scenario of Fig. 3 as an example. At the commencement of the cache-enabled D2D mmWave transmission, some users may volunteer to store files and hence may act as a helper to share files via D2D links. Depending on whether a user can find the requested file stored by one of its neighbour, the users can be classified into two types:

- D2D users (DUs): If a user requests one of the files stored in its neighbours' caches, the D2D transmitter will handle the request locally through D2D communication.
- Cellular users (CUs): If the file requested by a user is not cached by its neighbours, the user fetches the file from the BS as a regular cellular user.

C. How to Cache

To maximize the benefits of traffic offloading for the BSs, it is conducive to adopt incentives for promoting D2D transmis-

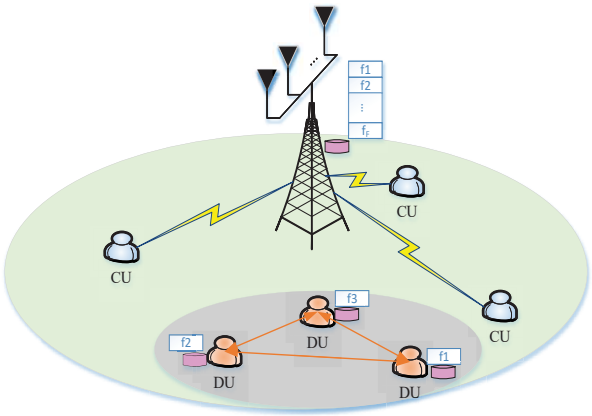
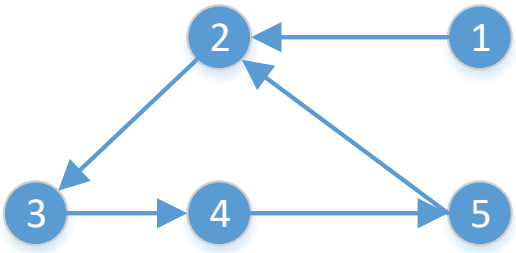
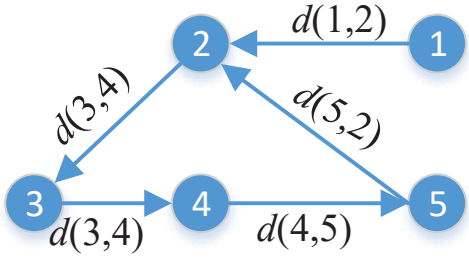


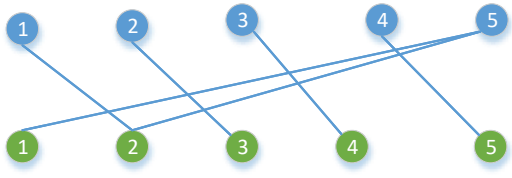
Figure 3: An illustration of network model for cache-enabled mmWave D2D communications with three potential D2D users. We have illustrated all possible D2D communication links in the figure.



(a) Illustration of directed graph



(b) Illustration of directed weighted graph



(c) Illustration of bipartite graph.

Figure 4: An example of potential D2D communication links.

sions by maximizing the total number of D2D links. To model the process of caching delivery, we may invoke the classic graph theory for illustrating the potential ways of obtaining the requested file for each user with the aid of D2D mmWave communications as seen in Fig. 4, where the vertices and the edges denote the potential D2D users and the corresponding potential D2D links, respectively. With the goal of maximizing the total number of D2D links, we can use a directed graph having 5 vertices and 5 edges, as seen in Fig. 4(a), for example. Again, with the goal of minimizing the transmission distances in mind, a directed graph is employed in Fig. 4(b), where the weights $d(i,j)$ reflect the distances between user i and user j, $i,j \in \{1,2,3,4,5\}$.

Bearing in mind that all users can play the role of the transmitter and the receiver in the D2D links, the resultant graph $\vec{G} = (\mathcal{V}, \vec{E})$ is non-bipartite. However, based on the specific roles of the users, the vertices $\mathcal{V}$ can be partitioned into two sets $\mathcal{T}$ and $\mathcal{R}$, where the vertices in $\mathcal{T}$ represent all of the transmitters in G and the vertices in $\mathcal{R}$ represent all of the receivers in G. Note that the vertex set $\mathcal{T}$ and $\mathcal{R}$ are disjoint, since every edge connects a vertex in $\mathcal{T}$ to one in $\mathcal{R}$. Hence, the directed graph $\vec{G}$ can be transformed into a bipartite graph, denoted by $G = (\mathcal{T}, \mathcal{R}, E)$. As shown in Fig. 4(c), let us consider the simple bipartite graph transformation of Fig. 4(a) and Fig. 4(b) as an example. With the advent of low-cost storage techniques, low-delay and highly-quality caching strategies are expected to emerge. These might rely on the radically new concept of caching multiple contents simultaneously in the spirit of the popular non-orthogonal resource-allocation principles. Additionally, the users' requests tend to change with the elapse of time, depending on the specific task they are engaged in. In this case, graph based machine learning provides a powerful tool that incorporates adaptive learning algorithms.

V. Resource Management for MmWave Small Cell Networks

Motivated by the reduced interference encountered in these systems, improved resource utilization, energy efficiency and load balancing may be attained. Specifically, the resource management involves the optimization of power allocation, user scheduling and hybrid beamforming designs in mmWave small cells. In this section, we will first touch upon the challenges of the resource management problem of mmWave small cells from the perspective of spectral efficiency and the diverse user requirements. Then we will contrast them to the conventional resource management strategies. Finally, we will discuss the potential of machine learning in tackling these challenging problems.

A. Conventional Resource Management

In typical next-generation applications and scenarios the demands of users will be quite diverse. Therefore, it is essential to seek an optimal resource management strategy with the ultimate aim of maximizing the benefits for the users in mmWave small cells. Note that maximizing the sum rate of a single mmWave cell relying on hybrid beamforming, configuring both the power allocation and user scheduling is challenging owing to the coupling amongst the analog/digital beamforming vectors, the power allocation coefficients and the

user scheduling. To elaborate, hybrid beamforming complicates the channel estimation and the training signal design, especially in dense networks. One of the particular challenges is that in small cells the angles-of-arrival (AoA) changes rapidly, which increases the call-dropping probability, unless frequent AoA updates are used. This in turn would increase the pilot overhead. Therefore, using partial CSI based resource management strategies is desirable for mmWave small cells. Random beamforming strikes a compromise between reducing the overhead of channel estimation as well as the feedback requirement and performance gain [14], [15]. In an effort to improve the performance of random beamforming, an efficient user scheduling method is required. Moreover, sophisticated power sharing strategies among the beams are capable of further enhancing the performance of mmWave-NOMA networks, which is also a pivotal issue of mmWave small cells.

Not withstanding the above discussions, another challenging task in the resource management is that the objective function and the related constraints tend to be nonlinear functions of their parameters. Specifically, the related resource-management performance optimization tends to exhibit combinatorial features such as those routinely encountered both in user-scheduling and in subchannel assignment as well as resulting in non-convex data-rate and energy-efficiency problems. Depending on the number of parameters and on the specific nature of the particular resource management problem, the task of finding the globally optimal solution can be rather challenging. Hence substantial research efforts have been invested in developing efficient algorithms for tackling these challenges, which can be broadly classified as follows: 1) deterministic approaches, 2) heuristic approaches, and 3) learning based approaches. Most of the above deterministic methods rely on strategies that infer the structures from the problem of interest [2], [5]. By contrast, the heuristic approaches invoke various of random guided search methodologies for finding the globally optimal solutions despite only visiting a fraction of the entire search space [8], [14]. The third class of methods attempts to harness probabilistic techniques for spotting the globally optimal solution, which is particularly suitable for scenarios associated with the unknown dynamics of the model considered [4], [13]. An overview of resource management in mmWave small cells is provided at a glance in Table I. In this treatise, we propose to exploit the powerful family of branch and bound (BB) techniques for global optimization by invoking the idea of monotonic transformation of the problem at hand.

Specially, we consider the NOMA-aided mmWave small cell shown in Fig. 2, where the BS uses random analog beamforming and the NOMA principle is invoked within each beam. For this system, we developed an optimal user scheduling and power allocation strategy in [10]. For the mmWave-NOMA system, the problem of maximizing the sum rate can be formulated as a multi-dimensional rectangular constrained optimization problem. Then BB techniques can be employed for finding both the optimal user scheduling and power allocation solution. Beneficial user-scheduling and power-allocation schemes may be designed based on classic matching theory and on successive convex approximation (SCA) approaches, respectively. Accordingly, in Fig. 5 we compare the performance of our low-complexity solution [10] based on matching theory and SCA to the excessive-complexity optimal solution based on BB both in NOMA-aided and in OMA-assisted mmWave networks. Our low-complexity algorithm achieves a sum rate close to that attained by the optimal algorithm for the specific parameters illustrated. The results also reveal that NOMA-aided mmWave small cell transmissions achieve a beneficial performance gain over OMA-based solutions, despite requiring a lower feedback overhead. Hence, NOMA-aided mmWave small cells constitute a promising architecture for the evolution of mmWave small cell networks.

B. Machine Learning Aided Resource Management

Traditional resource management techniques are generally designed for specific network configurations relying on scenario-oriented algorithms without any learning capability. By contrast, conceiving mmWave small cell networks relying on intelligent adaptive learning and decision making is capable of making resource management schemes more efficient. More explicitly, the main motivation of invoking machine learning for resource management is to enable the mmWave small cell networks to infer and harvest the diverse features of both the networks' architecture and the users' specific scenarios for autonomously determining the optimal system configurations. This enables the mmWave small cell networks to become self-organized, and to serve users at increased rates.

Machine learning enables devices in mmWave small cell networks such as the BS and user terminals to rely on human-like thinking, in order to utilize intelligent algorithms for resource managements. Cluster analysis constitutes a popular unsupervised learning method relying on sophisticated cognitive capabilities [3]. However, the choice of the clustering algorithm has to take into account the structure of the data set in the feature space. Having said that, it is challenging to beneficially adopt machine learning algorithms to the user clustering problems of mmWave-NOMA systems, since the attainable performance directly depends both on the specific properties of the features to be optimized and on the measurements. To facilitate the application of machine learning aided clustering techniques, both a quantitative feature set and a measurement function characterizing the objects of interest are required. For example, the channel correlations provide beneficial measurements for user clustering in mmWave-NOMA systems. To illustrate the correlation characteristics of mmWave channels, let us consider the two-user downlink MISO scenario illustrated in Fig. 6 as an example in conjunction with a pair of single-path mmWave channel vectors $\mathbf{h}_i$ and $\mathbf{h}_j$. The similarity between $\mathbf{h}_i$ and $\mathbf{h}_j$ can be qualitified by $\cos(\mathbf{h}_i, \mathbf{h}_j) = F_M(\pi[\theta_i - \theta_j])$, where $F_M(x)$ denotes the Fejér kernel. This means that the cosine similarity of two users' channel vectors can be characterized by their normalized directions determined by θ_i and θ_j. Fig. 6 shows the correlations versus the AoD difference between two channels. Note that in Fig. 6 the two users' channels are strongly correlated, when the difference of the two AoDs tends to zero. This is attributed to the fact that the beamforming-aided highly directional

Table I: Overview of resource management in mmWave small cells

Problems		**Solutions**	**Mathematical tools**
Performance metrics	**Designing parameters**		
Network utility and diverse user requirements such as sum rate and energy efficiency.	Power allocation User scheduling BS association Hybrid beamforming	Deterministic approaches	Convex optimization Nonlinear Programming Random search methodologies.
		Heuristic approaches	
		Learning based approaches	Probability theory

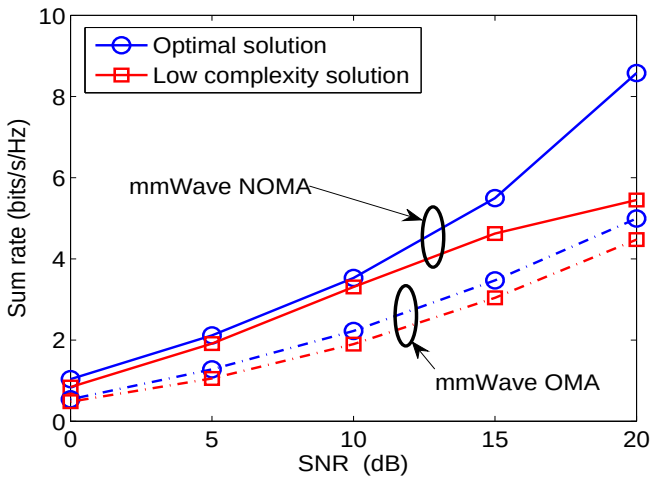


Figure 5: The sum rate of OMA and NOMA algorithms in mmWave networks at 28 GHz. The number of users is 6 with maximum 2 users sharing one beam. The minimum rate for each user is $R_{th} = 0.1$ bits/s/Hz. The radius of the mmWave small cell is 10 m.

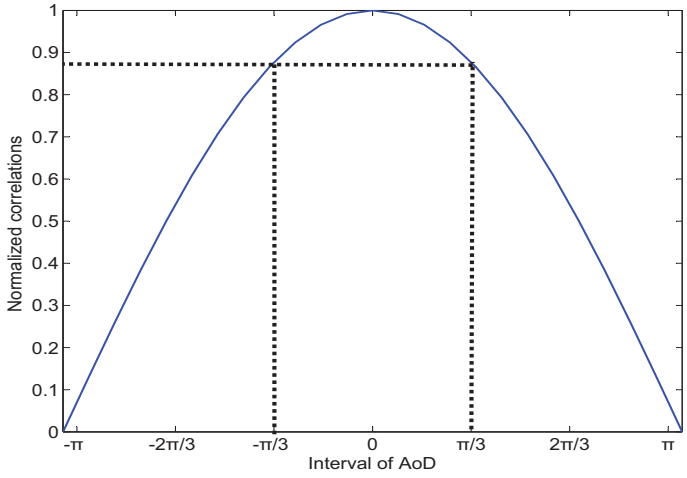


Figure 6: Illustrations of the channel correlation between two users under a single-path mmWave channel model.

transmission routinely used at mmWave frequencies makes the users channel strongly correlated, when their locations are close. In practice, mmWave small cells can be often found in the areas of high user density, such as coffee shops and airport terminals, etc [2], where users are more likely to be close to each other, hence forming hotspots. Therefore, a correlated setup based on spatial clustering models has become a preferred choice for the accurate modelling and analysis of these networks. In this spirit, a machine learning framework was proposed for user clustering in NOMA-aided mmWave small cell networks in [4].

VI. Conclusions and Future Challenges

In this article, the design challenges of NOMA-aided and cache-enabled mmWave communications have been highlighted. We first showed the key features of mmWave small cell networks, which provide new opportunities for ultra-dense deployments and massive connectivity. Then the benefits of NOMA-aided mmWave small cell networks have been demonstrated, followed by the design aspects of cache-enabled mmWave small cell networks. Finally, the challenges of resource management problems as well as the corresponding potential solutions suitable for mmWave small cell networks have been identified. There are still numerous open problems in the design of mmWave small cell networks, some of which are listed below.

- **Caching for mmWave small cell networks handling big data**: Some initial big data analysis based investigations in terms of cache-enabled mmWave small cell networks have been conducted by inferring and then exploiting the popularity of the content files. By relying on sophisticated data analytical tools such as stochastic modelling, data mining and machine learning, we can discover useful patterns from historical data, which constitutes a compelling research direction.
- **Unified OMA and NOMA mmWave small cell networks:** The family of advanced NOMA-aided mmWave transmission solutions is expected to co-exist both with conventional OMA and with sub-6 GHz transmission. The coexistence of these multiple access techniques raises challenging design issues, which have to be carefully considered.
- **Intelligent resource management paradigms:** Due to the heterogeneous network architecture of mmWave small cell networks, their resource allocation and interference coordination becomes a key challenge in the face of the ever-growing mobile traffic demand. Machine learning techniques are expected to further enhance the system performance of mmWave small cell networks by intelligent learning and decision making. However, given the dynamically fluctuating features and node mobilities, online and reinforcement learning algorithms have to be investigated.

- **Implementation cost:** In many proposals for mmWave small cell networks, most of the calculations are performed by the core network, which imposes a heavy feedback overhead and a high computational complexity. As a remedy, distributed computing algorithms are capable of alleviating the computational load imposed on the core network, while reducing the backhaul traffic. Given the recent advances in self-organizing network (SON) enabled mmWave network infrastructures, distributed computing provides beneficial insights into sophisticated algorithmic designs exhibiting improved flexibility at a reduced cost. However, the distributed computing algorithms are often constrained by the internal structure of distributed systems, such as their inherent coupling constraints and local information. These impediments have to be eliminated by further research.

REFERENCES

[1] H. Elshaer, M. N. Kulkarni, F. Boccardi, J. G. Andrews, and M. Dohler, "Downlink and uplink cell association with traditional macrocells and millimeter wave small cells," *IEEE Trans. Wireless Commun.*, vol. 15, no. 9, pp. 6244–6258, Sep. 2016.

[2] A. Ghosh, T. A. Thomas, M. C. Cudak, R. Ratasuk, P. Moorut, F. W. Vook, T. S. Rappaport, G. R. MacCartney, S. Sun, and S. Nie, "Millimeter-wave enhanced local area systems: A high-data-rate approach for future wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1152–1163, Jun. 2014.

[3] Y. Baştanlar and M. Özuysal, *Introduction to Machine Learning*. Totowa, NJ: Humana Press, 2014, pp. 105–128.

[4] J. Cui, Z. Ding, and P. Fan, "Machine learning based user clustering in mmWave-NOMA systems (invited)," in *IEEE Proc. of Veh. Technol. Conf. (VTC)*, Jun. 2018.

[5] A. Alkhateeb, J. Mo, N. Gonzalez-Prelcic, and R. W. Heath, "MIMO precoding and combining solutions for millimeter-wave systems," *IEEE Commun. Mag.*, vol. 52, no. 12, pp. 122–131, Dec. 2014.

[6] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, C. L. I, and H. V. Poor, "Application of non-orthogonal multiple access in LTE and 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 185–191, Feb. 2017.

[7] Y. Sun, Z. Ding, and X. Dai, "On the performance of downlink NOMA in multi-cell mmwave networks," *IEEE Commun. Lett.*, vol. 22, no. 11, pp. 2366–2369, Nov. 2018.

[8] B. Di, L. Song, and Y. Li, "Sub-channel assignment, power allocation, and user scheduling for non-orthogonal multiple access networks," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7686–7698, Nov. 2016.

[9] N. Rupasinghe, Y. Yapici, I. Guvenc, M. Ghosh, and Y. Kakishima, "Angle feedback for NOMA transmission in mmwave drone networks," *IEEE J. Sel. Topics Signal Process.*, pp. 1–1, 2019.

[10] J. Cui, Y. Liu, Z. Ding, P. Fan, and A. Nallanathan, "Optimal user scheduling and power allocation for millimeter wave NOMA systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1502–1517, Mar. 2018.

[11] M. Gregori, J. Gómez-Vilardebó, J. Matamoros, and D. Gündüz, "Wireless content caching for small cell and D2D networks," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 5, pp. 1222–1234, May 2016.

[12] E. Bastug, M. Bennis, and M. Debbah, "Cache-enabled small cell networks: Modeling and tradeoffs," in *International Symposium on Wireless Communications Systems (ISWCS)*, Aug. 2014, pp. 649–653.

[13] B. Chen and C. Yang, "Caching policy for cache-enabled D2D communications by learning user preference," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6586–6601, Dec. 2018.

[14] G. Lee, Y. Sung, and M. Kountouris, "On the performance of random beamforming in sparse millimeter wave channels," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 560–575, Apr. 2016.

[15] Z. Ding, P. Fan, and H. V. Poor, "Random beamforming in millimeter-wave NOMA networks," *IEEE Access*, vol. 5, pp. 7667–7681, 2017.

Jingjing Cui (M'18) received the Ph.D from Southwest Jiaotong University, Chengdu, China 2018. She is currently a research fellow with the School of Electronics and Computer Science, University of Southampton, UK. She was a research assistant with the School of Electronic Engineering and Computer Science, Queen Mary University of London, UK, from May 2018 to May 2019 and a Visiting Ph.D. Student at the School of Computing and Communications, Lancaster University, U.K., from November 2016 to November 2017. Her research interests include optimization theory and algorithm design, machine learning for wireless networks, and quantum communications.

Yuanwei Liu (S'13–M'16–SM'19) has been a Lecturer (Assistant Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since 2017. His research interests include 5G and beyond wireless networks, Internet of Things, machine learning, and stochastic geometry. He is in the editorial board of serving as an Editor of the IEEE Transactions on Communications, IEEE communication letters and the IEEE access. He has served as the Publicity Co-Chairs for VTC2019-Fall. He has served as a TPC Member for many IEEE conferences, such as GLOBECOM and ICC.

Zhiguo Ding (S'03-M'05) received his Ph.D from Imperial College London in 2005. He is currently a Professor in Communications at the University of Manchester. From Sept. 2012 to Sept. 2019, he has also been an academic visitor in Princeton University. Dr Ding' research interests are 5G networks, game theory, cooperative and energy harvesting networks and statistical signal processing. He has been serving as an Editor for IEEE TCOM, IEEE TVT, and served as an editor for IEEE WCL and IEEE CL. He received the best paper award in ICWMC-2009 and WCSP-2015, IEEE Communication Letter Exemplary Reviewer 2012, the EU Marie Curie Fellowship 2012-2014, IEEE TVT Top Editor 2017, 2018 IEEE Communication Society Heinrich Hertz Award, 2018 IEEE Vehicular Technology Society Jack Neubauer Memorial Award, and 2018 IEEE Signal Processing Society Best Signal Processing Letter Award.

Pingzhi Fan (F'15) is a distinguished professor and director of the institute of mobile communications, Southwest Jiaotong University, China, and visiting professor of Leeds University, UK, and Shanghai Jiaotong University, China. He is a recipient of the 1992 UK ORS Award, 1998 NSFC Outstanding Young Scientist Award, 2018 IEEE VTS Jack Neubauer Memorial Award, and 2018 IEEE SPS Signal Processing Letters Best Paper Award. He is an IEEE VTS Distinguished Lecturer, a fellow of IEEE, IET, CIE and CIC. His research interests include vehicular communications, massive multiple access and coding techniques, etc.

Arumugam Nallanathan (M'00–SM'05–F'17) is Professor of Wireless Communications and Head of the Communication Systems Research group in the School of Electronic Engineering and Computer Science at Queen Mary University of London since September 2017. Previously, he held academic positions with the Department of Informatics at King's College London, UK and the Department of Electrical and Computer Engineering, National University of Singapore. His research interests include 5G Wireless Networks, Internet of Things and Molecular Communications. He is a co-recipient of the Best Paper Awards presented at the IEEE International Conference on Communications 2016 and the IEEE Global Communications Conference 2017. He is an IEEE Distinguished Lecturer. He has been selected as a Web of Science Highly Cited Researcher in 2016. He has been an editor for various IEEE journals and served as the chair and member of numerous IEEE conferences. He received the IEEE Communications Society SPCE outstanding service award 2012 and IEEE Communications Society RCC outstanding service award 2014.

Lajos Hanzo (FREng, F'04, FIET and Fellow of EURASIP) received his 5-year degree in electronics in 1976 and his doctorate in 1983 from the Technical University of Budapest. In 2009 he was awarded an honorary doctorate by the Technical University of Budapest and in 2015 by the University of Edinburgh. In 2016 he was admitted to the Hungarian Academy of Science. During his 40-year career in telecommunications he has held various research and academic posts in Hungary, Germany and the UK. Since 1986 he has been with the School of Electronics and Computer Science, University of Southampton, UK, where he holds the chair in telecommunications. He has successfully supervised 119 PhD students, co-authored 18 John Wiley/IEEE Press books on mobile radio communications totalling in excess of 10 000 pages, published 1800+ research contributions at IEEE Xplore, acted both as TPC and General Chair of IEEE conferences, presented keynote lectures and has been awarded a number of distinctions. Currently he is directing a 60-strong academic research team, working on a range of research projects in the field of wireless multimedia communications sponsored by industry, the Engineering and Physical Sciences Research Council (EPSRC) UK, the European Research Council's Advanced Fellow Grant and the Royal Society's Wolfson Research Merit Award. He is an enthusiastic supporter of industrial and academic liaison and he offers a range of industrial courses. He is also a Governor of the IEEE ComSoc and VTS. He is a former Editor-in-Chief of the IEEE Press and a former Chaired Professor also at Tsinghua University, Beijing. For further information on research in progress and associated publications please refer to http://www-mobile.ecs.soton.ac.uk.