Embracing Non-Orthogonal Multiple Access in Future Wireless Networks

Zhiguo Ding, *Senior Member, IEEE*, Mai Xu, *Senior Member, IEEE*, Yan Chen, *Senior Member, IEEE*, Mugen Peng, *Senior Member, IEEE*, and H. Vincent Poor, *Fellow, IEEE*

***Abstract*—This paper is to provide a comprehensive survey for the impact of the emerging communication technique, non-orthogonal multiple access (NOMA), on future wireless networks. Particularly, how the NOMA principle affects the design of the next generation multiple access techniques is introduced first. Then, the applications of NOMA to other advanced communication techniques, such as wireless caching, multiple-input multiple-output (MIMO) techniques, millimeter-wave communications, and cooperative relaying, are discussed. The impact of NOMA on communication systems beyond cellular networks is also illustrated, by using digital TV, satellite communications, vehicular networks and visible light communications, as examples. Finally, the paper is concluded with some detailed discussions about important research challenges and promising future directions in NOMA.**

***Index Terms*—Non-orthogonal multiple access (NOMA), wireless caching, MIMO-NOMA, cooperative NOMA, millimeter-wave networks, VLC**

I. Introduction

Unlike wireline communications, the broadcasting nature of wireless communications means that wireless transmission is particularly prone to interference [1]. As a result, the use of the orthogonality principle, which provides a simple way to avoid co-channel interference, has been a dominant approach for those multiple access techniques used by the previous generations of mobile networks. For example, for the first generation of mobile networks, frequency division multiple access (FDMA) was used, by dividing the frequency domain into a lot of orthogonal small parts, which are termed frequency channels. These orthogonal frequency channels are then exclusively allocated to users, which effectively avoids the multiple access interference, i.e., one user solely occupies a frequency channel and one user's signal does not cause co-channel interference to others. Similar to the first generation, the following generations of mobile systems have also employed multiple access techniques based on the same idea that orthogonal resource blocks are obtained by using frequency/time/code domains and then allocated to users separately.

However, from the information theoretic perspective, it is well known that the use of orthogonal multiple access (OMA) approaches is not optimal in terms of the spectral efficiency [2], [3]. Take multi-user uplink transmission, which is termed multiple access channels in the information theory, as an example. As pointed out in [2], the rate region achieved by an orthogonal multiple access approach is only a part of the capacity region of multiple access channels, and this capacity region can be achievable if users are allowed to transmit at the same time/frequency/code. While this performance loss of OMA has been known for more than 50 years, the OMA approaches have been continuously used during the past, which is due to the fact that the implementation of those multiple access techniques based on non-orthogonality rely on the use of sophisticated transceiver designs. These designs typically result in high computational complexity as well as implementation costs, and hence cannot be supported in the previous generations of mobile systems.

Starting from 2013, the telecommunication industry has started the discussions for removing the orthogonality in the design of multiple access techniques for the next generation of mobile networks [4]–[6]. Meanwhile, various academic efforts have also been devoted to design new types of multiple access techniques based on the idea of spectrum sharing and serving multiple users at the same orthogonal resource blocks, which have been generally termed non-orthogonal multiple access (NOMA) [7]–[9]. These tremendous industrial and academic interests in NOMA are mainly due to the following three reasons. Firstly, thanks to Moore's Law, the computation power of devices in mobile networks has been significantly improved during the recent years, e.g., smart phones we are using now days are as powerful as computers and are capable of high performance computing. This increase of processing power is crucial for the implementation of NOMA. For example, many forms of NOMA require the receivers to carry out successive interference cancellation (SIC), a step which has been conventionally believed not feasible at the user side. Recently a NOMA chipset-embedded device has been developed to implement SIC at smartphones [10].

Secondly, NOMA was proposed at a time when the fifth generation (5G) networks are envisioned to not only support conventional voice and data services, but also provide the Internet of Things (IoT) functionalities. Recall that a key feature of IoT is that the number of devices to be connected can be massive, and hence realizing massive connectivity is important to support IoT in 5G. However, conventional OMA schemes cannot realize massive connectivity straightforwardly. Take time division multiple access (TDMA) as an example. If TDMA is used to support massive connectivity, a short time duration, e.g., one millisecond, needs to be further divided into a huge number of time slots, and hence the duration

Z. Ding and H. V. Poor are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544, USA. Z. Ding is also with the School of Computing and Communications, Lancaster University, Lancaster, UK (email: z.ding@lancaster.ac.uk, poor@princeton.edu).

M. Xu is with School of Electronic and Information Engineering, Beihang University, Beijing, China (email: MaiXu@buaa.edu.cn).

Y. Chen is with Huawei Technologies Co., Ltd., Shanghai, China (email: bigbird.chenyan@huawei.com.)

M. Peng is with the Institute of Telecommunications, Beijing University of Posts and Telecommunications, Beijing, China. (email: pmg@bupt.edu.cn).

of each time slot will be very small, which increases the implementation costs. Note that the use of FDMA for massive connectivity also results in a situation that adjacent frequency channels are too close, which can cause severe inter-channel interference. The use of the NOMA principle provides a more flexible way to support massive connectivity. Thirdly, future wireless networks face a situation that devices and users to be connected have diverse quality of service (QoS) requirements, to which the use of OMA is not appropriate [11]. For example, consider a scenario in which there are ten sensors which need to be served with low data rates only, and one broadband user. The use of OMA, such as orthogonal frequency division multiple access (OFDMA), means that each sensor is allowed to solely occupy one resource block, such as one OFDM subcarrier with 10 MHz. This is a waste to the valuable spectrum since sensors are given more bandwidth than what they need, but the broadband user might not have enough. A more spectrally efficient way is to encourage spectrum sharing, by implementing NOMA and integrating these sensors and the broadband user into a single subcarrier.

This paper is to provide a survey for the impact of the NOMA principle on wireless communications, from the following four perspectives. Firstly, how the NOMA principle is used to affect the design of multiple access techniques for future networks is focused. In particular, the general principle of NOMA is first discussed, then practical forms of NOMA using a single resource block are introduced, and various designs of NOMA schemes using multiple resource blocks are also described. It is important to point out that no multiple access technique, including NOMA, is perfect, where each multiple access scheme has its own advantages and disadvantages. This is the reason why the bandwidth resource blocks obtained from other types of OMA are used for the implementation of NOMA. Secondly, the impact of the NOMA principle to various advanced communication technologies, such as millimeter-wave (mmWave) transmission, multiple-input multiple-output (MIMO) techniques, cooperative communications, etc, are discussed. As shown in the paper, the spectral efficiency of these advanced communication technologies can be significantly improved with the application of NOMA. Furthermore, many features of these advanced communication techniques can be efficiently utilized to facilitate the implementation of NOMA for improving the system performance.

Thirdly, the NOMA principle is shown to be useful to many communication scenarios beyond cellular networks, although the concept of NOMA was originally designed for cellular networks. For example, in addition to radio frequency communication networks, the NOMA principle has been shown particularly useful to the design of visible light communication (VLC) networks. Another example is that the NOMA principle can be straightforwardly applied to those scenarios beyond telecommunications, such as terrestrial TV broadcasting and satellite communications. These discussions will illustrate that NOMA not only brings the chances to the design of future multiple access techniques, but also has the capability to shape the future communication networks. Fourthly, important directions for future research about NOMA are outlined and discussed. In particular, the challenges for the implementation of NOMA with imperfect channel state information (CSI) are described. The potentials for the applications of NOMA to physical layer security, full duplex communication systems, as well as radio frequency and VLC based energy harvesting, are also illustrated and discussed.

II. A Paradigm Shift in Designing Multiple Access Techniques

A. General Principles of NOMA

The essential principle of NOMA is to encourage spectrum sharing among multiple users, instead of allowing them to solely occupy orthogonal resource blocks. The basic idea of NOMA can be clearly illustrated by using two-user downlink power-domain NOMA as an example [12]–[14]. As its name suggests, power-domain NOMA is to use the power domain for multiple access. Without loss of generality, a two-user downlink scenario is used as an example, where the users are to receive different messages from their base station. If power-domain NOMA is used, the base station first superimposes the users' signals and broadcast this mixture to all the users. As a result, all the users are served at the same time/frequency/code, but with different power levels. These power levels are decided by the superposition coefficients, also termed power allocation coefficients. It is worth pointing out that power allocation of power-domain NOMA is different from conventional power allocation. Particularly, the user with poorer channel conditions gets more power allocated, compared to the user with stronger channel conditions. The reason for this type of power allocation is to ensure user fairness, since NOMA is a multiple access technique and needs to ensure that all the users are served. Assigned more power to the user with stronger channel conditions might improve the throughput, but can cause the user with poorer channel conditions disconnected.

The receivers of power-domain NOMA have different detection strategies, according to their channel conditions. Particularly, the user with poorer channel conditions treats its partner's information as noise, and directly decodes its own information, which is feasible since its own message was put on a higher power level than its partner's message. On the other hand, the user with stronger channel conditions will have to decode its parter's information first, before decoding its own, a procedure known as SIC [15]. The reason to use SIC at the user with stronger channel conditions is due to the use of power-domain NOMA power allocation, i.e., its own message was buried underneath its partner's information. The benefit of NOMA can be easily illustrated by considering an extreme case that the user with poorer channel conditions experiences deep fading. In this case, the use of conventional OMA, such as OFDMA, is very inefficient, since the subcarrier allocated to the weak user is wasted. By using NOMA, the bandwidth solely occupied by the weak user in the OMA mode can be released and used by other users, which significantly improves the spectral efficiency.

B. Implementing NOMA at a Single Bandwidth Resource Block

When multiple users are to share a single bandwidth resource block, NOMA can be implemented by simply using

power-domain NOMA, as explained in the previous subsection. Recall that the key idea of power-domain NOMA is to allocate more power to users with weaker channel conditions. However, how much power should be allocated to these users is not rigorously defined, which leads to an issue that power-domain NOMA cannot strictly guarantee the users' QoS requirements. In addition, power-domain NOMA cannot be applied to the scenario in which users have similar channel conditions. These become the motivations for another form of NOMA, termed cognitive radio (CR) inspired NOMA [16]–[18]. CR-NOMA is to treat NOMA as a special case of cognitive radio networks. Again take the two-user downlink case as an example. The weak user can be viewed as a primary user in cognitive radio networks, and the use of OMA is equivalent to a situation without any spectrum sharing, i.e., the weak user solely occupies the bandwidth. The use of NOMA is to introduce spectrum sharing, where a strong user, viewed as a secondary user in cognitive radio networks, is introduced to the system. Since the secondary user has a strong connection to the base station, it can significantly improve the overall system throughput. By using this synergy between NOMA and cognitive radio networks, a new form of NOMA, CR inspired NOMA, can be developed [16], [18].

The key difference between power-domain NOMA and CR-NOMA lies in two aspects: how the users are ordered and how the transmission power is allocated among the users. Particularly, CR-NOMA is to order users according to their QoS requirements, instead of their channel conditions. The CR-NOMA power allocation policy is to first provide the sufficient power to the users with strict QoS requirements, and the remaining power, if there is any left, is allocated to the users which can be served opportunistically. The performance gain of CR-NOMA over OMA and power-domain NOMA can be illustrated by a simple two-user downlink example, with the following assumptions:

- User 1 needs to be served timely, but its targeted data rate is low. Without loss of generality, assume the target rate is 1 bit/Hz/s. In practice, examples of this type of users can be healthcare devices connected wirelessly or wireless sensors for disaster management.
- User 2 is delay tolerant and can be served opportunistically, e.g., a data downloading task in the background for system updates.
- Both the users have the same channel gains, which are assumed to be 1 for the illustration purpose.

Since both the channel gains are the same, it is easy to show that, for this considered scenario, the sum rate offered by power-domain NOMA is the same as OMA, i.e., power-domain NOMA is not applicable to this scenario. When the transmit signal-to-noise ratio (SNR), denoted by ρ, is high, i.e., $\rho \to \infty$, the sum rates achieved by OMA and CR-NOMA can be approximated as follows:

- Each user's rate in OMA can be approximated as $\frac{1}{2}\log\rho$ bits/Hz/s. But user 1 only needs to be served with a rate of 1 bit/s/Hz. So the sum rate of OMA can be approximated as $(1+\frac{1}{2}\log\rho)$ bit/Hz/s.
- At high SNR, a very small amount of power needs to be consumed to guarantee the small targeted data rate of user 1. Therefore, the sum rate of CR-NOMA can be approximated as $(1+\log\rho)$ bit/Hz/s, which is much larger than that of OMA.

CR-NOMA also has other features which are different from power-domain NOMA. For example, the outage probability of a user in power-domain NOMA is only determined by its own channel condition, not by the other users' channels. However, in CR-NOMA, the outage performance of the users which are served opportunistically is not just related to their own channel conditions, but also determined by the channel quality of the other users. This is because CR-NOMA first serves those users with strict QoS requirements, which means that how much power available to those opportunistic users is determined by the channel conditions of the users with strict QoS requirements.

C. Implementing NOMA with Multiple Bandwidth Resource Blocks

1) Hybrid NOMA: Hybrid NOMA refers to a type of NOMA implementation, where each user is allowed to use multiple bandwidth resource blocks simultaneously and each resource block is to accommodate multiple users [16]. The key motivation for hybrid NOMA is to reduce the complexity for the implementation of NOMA. For example, consider a scenario in which there are 100 users in a cell. If all the users are grouped into a single group for the implementation of NOMA, the best user has to decode the rest 99 users' signals before decoding its own, which is obviously not feasible. Hybrid NOMA provides a low-complexity alternative for the implementation of NOMA. To be consistent to the exiting literature about hybrid NOMA, we use OFDMA subcarriers as examples of bandwidth resource blocks, given the fact that OFDMA will be used in 5G. Again take the 100-user case as an example. Hybrid NOMA can divide these users into 20 groups with 5 users in each group. Different OFDMA subcarriers are allocated to different groups, in order to avoid inter-group interference. Within each group, NOMA can be applied to serve 5 users at the same subcarrier, which significantly reduces the system complexity.

It is worth pointing out that hybrid multiple access techniques have already been used in the previous generations of mobile networks. For example, in GSM systems, 8 time slots created by TDMA are not sufficient to support a system with a large number of users, which motivates the combination between TDMA and FDMA in GSM. In the third generation (3G) mobile system, frequency division duplex is combined with CDMA to provide sufficient connections with reasonable reception reliability to multiple users. The fourth generation (4G) mobile network is another example of hybrid multiple access based mobile networks, where TDMA and OFDMA are efficiently combined together. Following the same rationale, it is expected that NOMA is also to be implemented in this hybrid manner in future wireless networks.

2) User grouping: A key step for the design of hybrid NOMA is user grouping, since the overall system performance is depending on which user is grouped with whom at

which subcarrier [19]–[21]. Initial studies about user grouping have drawn some interesting conclusions, as discussed in the following [16]. Provided that power-domain NOMA is implemented and users are grouped according to their channel conditions, one important conclusion is that users with different channel conditions can have completely different experiences. Particularly, a user with strong channel conditions benefits the implementation of NOMA, since this user's data rate in NOMA is very likely to be larger than that in OMA. On the other hand, a user with poor channel conditions may suffer some data rate loss, compared to the case with OMA, as it experiences strong co-channel interference caused by its partner. Another important conclusion is that, if CR-NOMA is used , the QoS requirements of the primary users can be strictly guaranteed, but the performance achieved by those secondary users can be largely depending on the channel conditions of the primary users, as discussed in the previous subsection for CR-NOMA.

With these insightful understandings, various user grouping algorithms have been developed in hybrid NOMA networks. It is worth pointing out that finding optimal user pairing for hybrid NOMA is not a trivial problem to solve, as it is essentially an integer programming problem. Furthermore, the user pairing issue is coupled with other optimization problems, such as power allocation and subcarrier allocation, which makes the overall system optimization very challenging. In [19], the monotonic optimization tool has been applied to hybrid NOMA for joint user grouping and power allocation. The benefit to use this tool is to ensure that an optimal solution for the non-convex mixed integer optimization problem can be found. While the computational complexity of the monotonic optimization tool is high, the use of this tool is still important, as it provides a useful benchmark for those developed low complexity sub-optimal solutions. It is also worth pointing out that other optimization tools other than monotonic optimization, such as branch-and-bound algorithms and machine learning methods, can also be applied to the addressed optimization problem [22].

3) Practical forms of hybrid NOMA: Because of its low complexity and superior spectral efficiency, the industry has developed various forms of hybrid NOMA. One of the most well-known hybrid NOMA is sparse code multiple access (SCMA) [23], [24]. The key advantage of SCMA is overloading, where the number of subcarriers is smaller than the number of the supported users, which is important to realize massive connectivity. The sparsity feature of SCMA is due to the requirement that each user is allowed to use a very small number of subcarriers. This sparsity feature is important to reduce the system complexity since the number of users occupying the same subcarrier becomes small. Compared to power-domain NOMA, SCMA exhibits two differences, one at the transmitter and the other at the receiver, as explained in the following.

- Unlike power-domain NOMA, SCMA requires the use of multi-dimensional coding at the transmitter, and the reason to use this coding is explained in the following. In SCMA, each user can use multiple subcarriers to transmit a single data stream, a feature also termed low-density spreading, and how subcarriers are allocated to a user is determined by the factor graph matrix [25], [26]. One option to use the multiple subcarriers is to generate multiple identical copies of the user's data stream and sends them over the multiple subcarriers, as done in low-density spreading. But SCMA adopts a more efficient way which is to generate correlated copies of the data stream and send these copies over the subcarriers.
- At the receiver, SCMA uses the message passing algorithm (MPA) instead of SIC, which is due to the following reason. If a user's information spread over multiple subcarriers is independently coded, SIC can be applied to decode the user's information at each subcarrier individually and then maximum radio combining can be used to combine the decoded information from different subcarriers. However, due to the use of multi-dimensional coding, one user's transmitted messages over different subcarriers are correlated, to which the MPA yields better performance than SIC [27], [28].

It is worth pointing out that there are other types of hybrid NOMA. For example, patten division multiple access (PDMA) is another example of hybrid NOMA [29], where a user's information is spread over multiple subcarriers, similar to SCMA, but the sparsity constraint of SCMA is removed, i.e., one user might use a lot of subcarriers in PDMA. Since there are less constraints for subcarrier allocation in PDMA, there are more degrees of freedom for the system design, which can be used to improve the system performance but at a price of increased complexity.

III. Applying NOMA to Other Advanced Communication Technologies

The NOMA principle not only brings changes for the design of the next generation multiple access techniques, but also has an important impact on the design of other advanced communication technologies, as illustrated in the following subsections.

A. NOMA Assisted Wireless Caching

The key idea of wireless caching is to proactively push popular content files to local caching infrastructure, e.g., local content servers or other users in the device-to-device (D2D) caching case [30], [31]. As a result, when the users request these files, they do not need to directly communicate with the base station, but simply fetch the files from their local content servers or D2D helpers. The benefit of wireless caching can be illustrated by the following example. Consider that there are 100 users which request different files. Without wireless caching, 100 resource blocks need to be consumed to accommodate these users' requests. However, provided that these files have been previously cached by the local content servers, only one resource block is needed to serve these 100 users. The reason for this is due to the fact that the content servers can help their associated users locally and the use of short range communications ensures that all the transmission by the content servers can be carried out simultaneously.

Conventional wireless caching assumes that content pushing is carried out by using off peak hours, during which a lot of the spectrum is idle and can be used for content pushing [31]–[34]. This assumption is valid if the popularity of the content files varies slowly. Typical examples for this type of content are software updates, popular movies and TV streaming, etc. However, many other types of content, such as up-to-date sport event news and sale pricing information, exhibit a fast time-varying feature, and need to be updated frequently. To these types of content, the assumption that using off peak hours for caching is not applicable, since the files cached during off peak hours might become outdated during peak hours. The application of the NOMA principle can bring some fundamental changes for the design of wireless caching, as illustrated in the following [35].

Since off peak hours cannot be used, content pushing has to be carried out during peak hours. In order to keep the files cached at the local content servers are frequently updated, a short time duration needs to be periodically used for content pushing. This periodically used duration has to be short since not all the pushed files are useful for users and spending a large amount of time for content pushing will reduce the spectral efficiency of wireless caching. If OMA based content pushing is used, this duration will be further divided into small time slots, and the base station will push one file to a single content server during each time slot. If the number of the content servers is large, some content servers might not get any file pushed from the base station, which is the drawback of OMA based content pushing. If the NOMA principle is applied, the base station can superimpose multiple content files which are intended to different content servers, and uses one time slot to serve multiple content servers. As a result, the use of the NOMA based caching scheme is more suitable to meet the constraint that limited bandwidth resources are reserved for content pushing. Similarly, the concept of NOMA can also be applied to the content delivery stage. Recall that the purpose of the content delivery stage is to ask the content servers to serve their associated users, if these users' quested files can be found locally. The drawback of OMA based content delivery is that at each time, a content server can serve one user only. However, for many high-density wireless networks used in airports or stadiums, it is very likely that one content server has more than one user to serve. The use of NOMA can ensure that multiple users can be connected to the same content server, which improves the latency of wireless caching, since users do not have to wait for a long time to be served.

Another NOMA assisted wireless caching scheme is to opportunistically carry out content pushing during the content delivery stage. In conventional wireless caching, the stages for content pushing and content delivery are strictly separated, i.e., time slots during the content delivery stage cannot be used for content pushing. However, if the time duration between two adjacent content pushing stages is large, the content files at the local content servers cannot be frequently updated. By using the NOMA principle, this drawback of conventional wireless caching can be avoided. Particularly, some time slots during the content delivery stage can be identified as opportunities for content pushing. For example, during some time slots in the content delivery stage, users make the requests to be served, but their requested files cannot be found in the caches of the local content servers. Conventionally this type of events are viewed as non-ideal since the base station has to serve these users directly and hence the spectral efficiency of wireless caching is reduced. With the application of NOMA, the base station can superimpose two types of signals, one to be delivered to the users directly and the other to be pushed to the content servers. As a result, the base station does not have to wait until the next content pushing stage to push files to the content servers, and the files stored in the local caches can be frequently updated.

B. MIMO-NOMA

The NOMA principle has also a significant impact on the design of MIMO technologies. Particularly, spatial directions can also be viewed as a type of bandwidth resource blocks. Conventional MIMO techniques, such as zero forcing, prefer to serve a single user at one of orthogonal spatial directions, whereas the use of NOMA ensures that more users can be connected at a single spatial direction [36]. In the following, general principles of MIMO-NOMA are discussed first, and then some practical designs are introduced.

1) General principles: Unlike single-input single-output (SISO) NOMA, it is very challenging to identify the optimality of MIMO-NOMA. Without loss of general, we mainly focus on downlink NOMA in the following. In [37], the relationship between the rate region achieved by SISO-NOMA and the capacity region of broadcast channels has been clearly illustrated. But little is known about how optimal MIMO-NOMA is, partially because the capacity region for general broadcast channels is still unknown. Note that dirty paper coding (DPC) has been well accepted as a reasonable benchmark given its capability to approach an upper bound on the capacity region. Therefore, it is of interest to study the comparison between NOMA and DPC.

In [38], [39], a condition for NOMA to realize the same performance as DPC, termed the quasi-degradation criterion, is established, for the multi-input single-output (MISO) scenario in which the base station has multiple antennas and each user has a single antenna. Provided that users' channels satisfy the quasi-degradation criterion, the use of NOMA yields the same performance as DPC, but it is important to point out that the complexity of NOMA is linearly proportional to the number of users, much smaller than that of DPC. The following two examples are provided to illustrate the key idea of the quasi-degradation criterion:

- When users' vectors have the same directions but different magnitudes, the quasi-degradation criterion is satisfied. The optimality of NOMA in this scenario is intuitive, since a beamforming vector good to one user is also good to the others, i.e., users are located at the same spatial direction and hence can be served by using a single beam.
- The quasi-degradation criterion cannot be satisfied if users have orthogonal channels. The conclusion that NOMA cannot be applied to this scenario is also intuitive since one user's beam is useless to the others due to the orthogonality of the users' channels.

However, the quasi-degradation criterion can be applied to MISO-NOMA only, and its extension to general MIMO-NOMA is still unknown.

2) Practical designs of MIMO-NOMA: Even though the optimality of MIMO-NOMA is still unknown, it is worth to developing practical MIMO-NOMA designs, with the aim that they can outperform MIMO-OMA. One popular way for designing MIMO-NOMA is to ask the base station to generate many non-orthogonal beams, where a single user is accommodated by one beam [40], [41]. Since the generated beams are non-orthogonal, overloading can be supported by this type of MIMO-NOMA, i.e., the number of the supported users is larger than the number of the antennas at the base station. A key challenge for this type of MIMO-NOMA is how to order users according to their channel conditions, since channels are in forms of vectors or matrix. The existing studies in [40], [42] have shown that the use of path loss for user ordering can ensure a reasonable performance gain over OMA.

Another way for designing MIMO-NOMA is to decompose MIMO-NOMA into SISO-NOMA, by carefully designing precoding and detection matrices [43], [44]. Particularly, the spatial degrees of freedom are first used to create some orthogonal beams by using conventional MIMO techniques, and then the NOMA principle is applied to ensure that multiple users can be served by each of the generated beams. The benefit of this type of MIMO-NOMA is that there is no need to directly order users' channel vectors/matrices, since after converting MIMO-NOMA to SISO-NOMA, the effective channel gains are in forms of scales, instead of vectors or matrices. In addition, this type of NOMA facilitates the implementation of hybrid NOMA, and it is also applicable to both uplink and downlink transmission. Furthermore, this type of MIMO-NOMA designs is particularly suitable to massive MIMO scenarios, where the users sharing the same channel correlation matrix can be grouped together and served by the same beam [45], [46].

C. MmWave-NOMA

With the rapid growth of traffic demand, the frequency below 6 GHz used by conventional wireless networks becomes too crowded, which motivates the recent industrial and academic interests in mmWave communications by using the less occupied mmWave spectrum [47], [48]. It is interesting to point out that the motivation to use NOMA is exactly the same as mmWave, but the solution provided by NOMA is to improve the efficiency for using the available bandwidth. Obviously mmWave communications and NOMA are not conflicting but complementary to each other. On one hand, the mmWave bands are not free of charge, but can be very expensive according to the lessons learned from 3G/4G spectrum auctions, which motivates the use of NOMA in mmWave communications as an cost-effective measure. On the other hand, even if mmWave bands turn out to be much less expensive than the lower-frequency ones, the tremendous increase in the number of mobile devices and the types of bandwidth demanding services, such as ultra-high definition video streaming and online interactive games, will soon place a strict requirement on how efficiently the mmWave bands are used, which also motivates the use NOMA in mmWave networks.

In addition to the aforementioned motivations, the application of NOMA can efficiently use some features of mmWave transmission, and hence significantly improve the spectral efficiency of mmWave communications. For example, one of the key features of mmWave communications is that mmWave transmission is highly directional. In conventional wireless communications using the frequency lower than 6 GHz, the channels of two receivers which are spaced more than half of the wavelength can be assumed to be independent, due to multi-path fading, i.e., the number of the paths between a transmitter and a receiver can approach the infinity in a rich scattering environment. However, in mmWave transmission, the number of paths is very small, and the path of line-of-sight is dominant, which means that two users' channels can be highly correlated, even if the distance between the two users is large. According to the quasi-degradation criterion [39], the situation in which users' channels are correlated is ideal for the application of NOMA, where a single beam generated by the base station can accommodate both users. The benefit for this type of mmWave-NOMA can be explained by using the following example [49], [50]. Consider that there are 8 single-antenna users and the base station has 4 antennas only. The use of conventional zero forcing can only ensure that 4 users are simultaneously served by the base station. By applying mmWave-NOMA and exploring the channel correlation, all the users might be supported at the same time. Furthermore, the use of conventional zero forcing results in poor reception reliability if users' channels are correlated, since it tries to create two orthogonal beams for these users. But spatial degrees of freedom can be more efficiently used in mmWave-NOMA, by accommodating users with correlated channels at a single orthogonal direction and relying on NOMA for handling the intra-beam interference.

Another example for the features of mmWave communications to facilitate the application of NOMA is the use of finite-resolution analog beamforming (FRAB) [51], [52]. In particular, FRAB is a special case of analog beamforming, where the phases of the transmitted signals are changed, but their amplitudes are kept the same. The reason to use analog beamforming is mainly due to the high cost of radio frequency chains, where changing the signal amplitudes can be much more expensive than that of changing the signal phases. In practice, the signal phases are changed by using phase shifters, and the number of phase shifts supported by practical circuits is limited. For example, if a perfect analog beamformer requires a shift of 1×10^{-5} degrees, it most likely cannot be supported in practice. It is worth pointing out that FRAB is not only applicable to mmWave communications, but also commonly used in massive MIMO systems. While the use of FRAB can significantly reduce the cost of hardwares, it is well known that this type of imperfect beamforming causes performance degradation, since these generated beamforming vectors are not perfectly aligned with the users' channels. However, the feature of FRAB that beams are not aligned with users' channels can be used to facilitate the implementation of NOMA, as illustrated in the following example [53]. Consider

that there are 2 single-antenna users and the base station has 2 antennas. Furthermore, assume that the users have orthogonal channels. Using conventional zero forcing techniques, the base station can serve the 2 users simultaneously, which consumes all the degrees of freedom at the base station. It is preferable to apply NOMA to this scenario, so the 2 users can be grouped and served by one beam, which saves the degrees of freedom at the base station and provides the possibility to serve additional users. According to the quasi-degradation criterion, the application of NOMA to this scenario is not possible, since the users have orthogonal channels. However, if FRAB is used to generate beams, it is possible that the two users with the orthogonal channels prefer the same FRAB vector, particularly when the resolution of FRAB is low. As a result, the base station can use a single beam to serve the two users and hence additional beams can be generated to serve more users, which is not possible if perfect beamforming is used.

D. Cooperative NOMA

The existing cooperative NOMA schemes can be divided into two types. The first one is to seek the opportunity for cooperation by asking one NOMA user to help the others [54]–[56]. This type of cooperative NOMA is motivated by the fact that the implementation of NOMA can degrade some NOMA users' performance. As discussed in the subsection for CR-NOMA, the far user can be viewed as a primary user, and the use of NOMA can potentially reduce its reception reliability since an additional user is introduced to the system. The key idea of the first type of cooperative NOMA is to recruit the users with strong channel conditions as relays and help those users with poor channel conditions. It is worth pointing out that the SIC feature of the NOMA receivers facilitates the cooperation among NOMA users. In particular, those users with strong channel conditions have to first decode the signals to the users with poor channel conditions. As a result, the information for the users with poor channel conditions becomes available to the strong users after carrying out SIC. Therefore, it is natural to use these strong NOMA users as relays, where no extra time slot is needed to deliver the weak users' information to the strong users.

One drawback of the first type of cooperative NOMA is its limited diversity gain, since it relies on the cooperation among the active users but the number of active users in practice might be small. The second type of cooperative NOMA is to avoid this drawback and seek the help from dedicated relays which assist a base station to deliver the information to its users [57], [58]. Because the number of inactive users in a network is much larger than that of the active ones, a higher diversity gain can be achieved by using these inactive users as dedicated relays, compared to the case that only the cooperation among the NOMA users is employed. In addition, using dedicated relays can be particularly useful if the base station does not have direct links with the NOMA users, i.e., the NOMA users are located close to the cell edge. In this case, employing cooperative NOMA can ensure that the users' information can be delivered from the relay to the users more spectrally efficiently than cooperative OMA, since one NOMA broadcasting by a relay can help multiple users.

It is also worth pointing out that the network topology has a great impact on the design of efficient cooperative NOMA. For example, when the cooperation among NOMA users is used, some users which have strong connections to the base station but not to the weak users should not be employed as relays, as the use of short communications for relay transmission becomes not possible and hence extra bandwidth resources are needed by these strong users to reach the weak users. When dedicated relays are used, different designs of cooperative NOMA can be developed depending on whether the users have direct links with the base station and which users need the help from the relay. Furthermore, many research efforts have also been devoted to a particular network topology, in which multiple relays are available for coopveative NOMA. While distributed beamforming can be used to exploit all the available relays, the coordination among these relays, such as time and phase synchronization, can consume a lot of system overhead. As a result, relay selection, i.e., selecting a single relay for the cooperation, is preferred in practice. Depending on whether amplify-and-forward or decode-and-forward is used at the relays, various relay selection schemes have been developed [59], [60]. It is worth pointing out that the max-min relay selection strategy which is proved to be optimal in conventional cooperative networks is no longer optimal in cooperative NOMA. The main reason for this is due to the fact that the max-min criterion is to select a relay whose incoming and outgoing channels are most balanced; however, these incoming and outgoing channels are not equally important in cooperative NOMA. As shown in [59], [60], various relay selection strategies have been developed and shown to outperform the max-min selection scheme.

IV. Applications of NOMA Beyond Cellular Networks

A. Vehicular Ad Hoc Networks

Vehicular ad hoc networks (VANETs) are envisioned to provide important applications related to road safety, data sharing among vehicles, intelligent transportation, etc, and also support the forthcoming connected and autonomous vehicles and systems [61]. Originally, only vehicle-to-infrastructure (V2I) and vehicle-to-vehicle (V2V) communications were considered in VANET, and recently other types of communications, such as vehicle-to-pedestrian (V2P), vehicle-to-device (V2D) and vehicle-to-grid (V2G), have also been considered, which leads to the general term, vehicle-to-everything (V2X) communications [62].

The key feature of V2X communications is the short duration for the communication connection, which can be illustrated by using V2I communications an example [63]. V2I communications refers to the scenario where the infrastructure, such as a roadside unit or base station, communicates with vehicles. For example, the infrastructure gathers the local information from vehicles, which is similar to uplink transmission in conventional cellular networks. In addition, via downlink, the infrastructure frequently disseminates global traffic and road information to the vehicles and may also need to provide a certain suggestions and instructions for real-time

motion planning to those autonomous cars. Because vehicles are moving at high speeds, the connection period between a vehicle and the infrastructure can be very short, which imposes a challenge for reliable communications over V2I channels. The fact that there are a large number of devices to be connected within this short duration makes the problem even more difficult.

Compared to OMA based transmission strategies, the use of the NOMA principle provides more flexibility to provide timely and massive connections, realize dynamic resource allocation and meet the users' diverse QoS requirements [11]. For example, consider that two vehicles need to be connected with the same infrastructure, where one needs to be connected to receive real-time content and the other can be served opportunistically as it requires non-real-time servies. The use of NOMA can ensure that both the users have access to all the bandwidth resources, such as the short connection duration and the spectrum, which is particularly important the VANET scenario. Furthermore, the transmission power allocated to the users can be flexibly designed to ensure that the QoS requirement for the user with the real-time service is strictly guaranteed, while the infrastructure opportunistically delivers the non-real-time files to the other user.

Handover is another key challenge for VANETs, since a vehicle with high mobility can travel through multiple cells covered by different roadside base stations in a short period. The use of network MIMO, such as coordinated multipoint (CoMP) and cloud radio access networks (CRAN), has been recognized as an efficient method to combat this handover issue, where one vehicle can be connected to multiple base stations and hence disconnection can be avoided. The use of NOMA can further improve the spectral efficiency of network MIMO. For example, while two base stations serve one user simultaneously, they cannot be accessed by other users in conventional network MIMO. However, by using NOMA, each base station can schedule additional users which are close to the base station. This strategy is particular important to VANETs, since more users can be connected during the short connection duration [64].

B. Visible Light Communications

Similarly to mmWave communications, visible light communications (VLC) is also motivated by the fact that there are not sufficient bandwidth resources below 6 GHz reserved for wireless communications, and it is preferable to use those less occupied high frequency bands [65]. Particularly, VLC is to use visible light whose frequency is between 400 and 800 THz for communications. Note that acquiring more bandwidth, i.e., using mmWave and VLC, is not conflicting with the goal of improving the spectral efficiency, i.e., using NOMA [66]. On the contrary, how to efficiently use the spectrum is important even if there are plenty of new bandwidth resources obtained, in order to support emerging broadband services, as discussed in the subsection for mmWave-NOMA.

Similar to mmWave transmission, VLC communications also exhibits some features which facilitate the implementation of NOMA [21]. For example, channels for the scenario using frequencies lower than 6 GHz can suffer fast-time-varying multi-path fading, which makes the design of NOMA challenging. This is because the important components of NOMA transceivers, such as SIC, MPA, or NOMA power allocation, require the perfect knowledge of CSI. The use of imperfect CSI can significantly degrade the performance of NOMA. However, in the context of VLC communications, the VLC channels can be viewed as static. This is due to the fact that VLC mainly relies on the line-of-sight path, and those effects important to conventional radio frequency systems, such as reflection and diffusion, can be ignored in VLC. With these static channels, the implementation of NOMA becomes more straightforward than the cases using the radio frequency.

In addition, it is well known that the performance gain of NOMA over OMA is particularly significant in the high SNR regime [67]. This phenomenon can be explained by using CR-NOMA as an example. Recall that CR-NOMA is to first provide sufficient power to ensure some users' QoS requirements guaranteed. In the high SNR regime, the users' predefined QoS requirements can be easily met, and there will be a lot of power left to serve additional users, which yields the significant gain of NOMA over OMA. In VLC communications, it is typical that there is a strong line-of-sight path between the transceivers and the distance between the transmit LED and the receive photo detector (PD) is short, which means that VLC mainly operates in the high SNR regime and hence it is ideal for the application of NOMA. It is also important to point out that VLC is typically applied to a cell with a small coverage [21]. Therefore, the number of devices to be connected within the small-size cell is also small, which is helpful to reduce the complexity for the implementation of NOMA. Furthermore, VLC channels are significantly affected by the transmission angles of the transmit LEDs and the field of views (FOVs) of the PD, which are new system parameters not presented in conventional radio frequency systems. By adjusting these system parameters, the channel conditions of the users can be more dynamically controlled to facilitate the implementation of NOMA.

C. Terrestrial TV Broadcasting and Terrestrial-Satellite Communications

Terrestrial digital TV broadcasting is surprisingly becoming one of the first practical systems to which NOMA has been applied. Conventionally orthogonal multiplexing techniques, such as frequency division multiplexing and time division multiplexing, have been used for TV broadcasting, due to their low system complexity and affordable costs. However, the spectral efficiency of these orthogonal multiplexing techniques is low, and cannot be used to meet users' diverse QoS requirements. Recently, the Advanced Television Systems Committee (ATSC) has proposed a new type of terrestrial TV broadcasting, where the corresponding physical layer protocol standard is known as ATSC 3.0 [68]. In this new generation of digital TV standards, a new type of multiplexing, which is termed layered division multiplexing (LDM) and based on the NOMA principle, has been used.

The key idea of LDM is very similar to power-domain NOMA, where multiple broadcasting services are integrated

at a single bandwidth resource block [69]. Particularly, the simplest form of LDM integrates two layers, namely Core Layer and Enhanced Layer, at the same time and frequency. The signals from the two layers are intended to destinations with different receive capabilities. These signals are encoded with different types of error correction codes and then are superimposed in the same manner as power-domain NOMA. The benefit of LDM can be explained by the following example. Consider that there are two types of TV broadcasting receivers. One type of receivers are static and have strong connections with the TV broadcasting station, and the other can be receivers with mobility, such as pedestrians, vehicles, etc. By using LDM, two types of broadcasting services, high-definition services and ultra-high-definition (UHD) streaming (such as 4K UHD and 8K UHD television), can be integrated at a single resource block and broadcasted to the users. Similar to the receivers for power-domain NOMA, the users with weak channel conditions can at least decode the signals in the Core Layer, by treating the information in the Enhanced Layer as noise. The users with strong channel conditions can receive UHD video streaming, by carrying out SIC and decoding the signals in the Core Layer before decoding UHD signals.

Terrestrial-satellite communications is another example to which the application of NOMA has been shown very useful [70], [71]. Conventional cellular networks can be viewed as a special case of terrestrial communications, which can support a large system throughput but suffer a coverage drawback. However, satellite communications can provide seamless global coverage, which motivates the joint design of terrestrial-satellite communications. To this new type of communications, existing studies have shown that the use of application of NOMA enables a heterogenous network architecture, where users are jointly served by a satellite as well as terrestrial base stations. Due to the large distances between the satellite and the users, the signals from two communication systems need to be carefully structured for the interference management purpose, where user clustering/grouping has a significant impact on the overall performance of the NOMA assisted terrestrial-satellite communication system.

V. Future Directions

A. NOMA with Imperfect CSI

In NOMA, users are allowed to use the same bandwidth resource block, which means that strong co-channel interference exists in NOMA systems and is suppressed by using advanced signal processing algorithms at the transceivers, such as SIC, power allocation, beamforming/precoding, etc. Typically the implementation of these signal processing algorithms requires perfect CSI at the transmitter, although it is important to point out that some NOMA schemes in [12], [43] do not have strong assumptions about CSI. For example, the simplest form of power domain NOMA requires the base station to know the order of users' channels only, instead of the exact information of the channels. Furthermore, in [43], the users' channel matrices are not required to be available at the base stations, and only scale effective channel gains are needed.

In practice, the transmitter can have access to imperfect CSI only, and imperfect CSI can be generally divided into three types. One is CSI with channel estimation error, which is due to the use of imperfect channel estimators. Note that a straightforward way for channel estimation in NOMA is to assign orthogonal pilots to users, in the same manner as OMA; however, a more spectrally efficient method is to superimpose the unknown data with predefined pilots signals, which reduces the system overhead for sending the training information [72]. The second type of imperfect CSI is statistical CSI, where instantaneous CSI is not available but the statistical information about CSI is known by the base station. This type of imperfect CSI is motivated by the fact that it is difficult for the base station to perfectly know its connections to the users. Take a downlink case with high-mobility users as an example. A huge amount of system overhead needs to be consumed in order to carry out frequent channel estimation at the users, and the fact that the CSI fed from the users back to the base station might become outdated makes the problem more challenging. As a result, asking the users to feed the statistical information of CSI, such as large scale path loss which varies slowly, back to the base station becomes preferable. In [73], NOMA has been shown robust to the two types of imperfect CSI, compared to OMA.

The third type of imperfect CSI is limited feedback, which is again to avoid using too much system overhead for channel feedback [53], [74]. Unlike the previous two types of imperfect CSI, limited feedback offers some degrees of freedom to realize a balanced tradeoff between system performance and complexity, which is the reason why this direction has attracted a lot of attention. Take one-bit feedback as an example, which is to ask the base station to broadcast a threshold and each user to feedback 1 (0) if its channel gain is above (below) the threshold. Obviously the threshold is an important system parameter which should be carefully designed. It is interesting to point out that an appropriate choice of the threshold can still ensure that the maximal multi-user diversity gain is achievable in NOMA systems, even with one-bit feedback. In practice, one-bit feedback is an extreme case, and multiple bits might be afforded for channel feedback, which means that the number of feedback bits is another system parameter to be optimized.

B. Combining NOMA with Full Duplexing

Full duplexing is another important technology to be used in 5G [75], and its key idea is to enable a communication node for transmitting and receiving at the same time. Intuitively full duplexing is more spectrally efficient than half duplexing, as a node can carry out the two functionalities, transmitting and receiving, by using a single resource block. While the performance gain of full duplexing over half duplexing is clear, most existing communication systems are still based on the half duplexing mode, since full duplexing suffers strong loop-back interference, e.g., the transmitted signals become cross-talk interference to the received signal. Thanks to the recent advance of loopback interference cancellation techniques, full duplexing has attracted a lot of attention during the past few years.

The most well known example for the combination between full duplexing and NOMA is the application of full duplexing

to cooperative NOMA [76]–[78]. Take cooperative NOMA without using dedicated relays as an example. As discussed in Section III-D, users with strong channel conditions are employed as relays to help those users with poor channel conditions. While this cooperative NOMA can improve the reception reliability of the weak users, the overall spectral efficiency of NOMA is reduced, since extra time slots are consumed for carrying out relay transmission. The application of full duplexing to cooperative NOMA means that those strong users receive signals from their base station while helping those weak users. As a result, no extra time slot is required for relay transmission, compared to non-cooperative NOMA, which demonstrates that full duplexing is particularly important cooperative NOMA. It is worth pointing out that the current full duplexing techniques cannot guarantee that the loopback interference is completely removed, and a lot of research efforts are currently devoted to fully identify the impact of the residual loopback interference on cooperative NOMA with full duplexing.

In addition to cooperative NOMA, full duplexing can be applied to other types of NOMA scenarios. For example, full duplexing can be used to realize simultaneous uplink and downlink transmission [19], [79], [80]. In particular, a base station with the full duplexing capability can broadcast the NOMA mixture to the users, while receiving the uplink signals from the users. If the half duplexing mode is used, two times of bandwidth resources will be needed, compared to the full duplexing case. Similar to cooperative NOMA, the residual loopback interference can potentially degrade the system performance, i.e., the reception reliability for the base station to decode the uplink signals is deteriorated by the transmitted downlink signals. But unlike cooperative NOMA, the joint design of uplink and downlink results in another issue that uplink users can cause strong interference to downlink users. Because uplink and downlink transmission happen at the time, the performance of those downlink users can be significantly degraded by those nearby uplink users. However, in [80], it is shown that significant performance gains over half duplexing NOMA and full duplexing OMA can be still realized, if uplink and downlink users are placed in different sectors of a cell. More sophisticated algorithms for uplink and downlink user clustering as well as scheduling are needed to avoid this co-channel interference, by dynamically using the users' channel conditions.

C. NOMA Assisted Physical Layer Security

Similar to conventional multiple access techniques, the security issue was not originally considered when the concept of NOMA was developed. This is due to the fact that in the current mobile systems, security is realized by using upper layer cryptographic methods. For example, in OFDMA systems, a user which is allocated to the first subcarrier is capable of decoding the bits sent at the other subcarriers, but the use of cryptographic methods can ensure that this user cannot know the meaning of the decoded bits. Nevertheless, initial studies have indicated that the user of NOMA transmission is helpful to the implementation of physical layer security, as illustrated in the following.

In the presence of external eavesdroppers, it is shown that the use of NOMA can improve the secrecy rates of the legitimate users in various communication scenarios compared to OMA [81]–[83]. Take the simple downlink power-domain NOMA scheme as an example, where the base station superimposes the users' signals and broadcasts the generated mixture. Compared to OMA cases, signals are not separately sent in NOMA, and the power allocation coefficients are tailored to the channel conditions of the NOMA users, which makes the eavesdroppers more difficult to intercept the signals sent to the legitimate receivers. However, it is important to point out that there are two key challenges for secure NOMA transmission with external eavesdroppers. One is how to rigorously define the decoding rates at the eavesdroppers. A straightforward way is to assume that the eavesdroppers use the same SIC procedure as the legitimate users. However, there are alternatives for the eavesdropping strategies, e.g., a signal decoded during the last stage of SIC at legitimate users might be decoded first at the eavesdroppers, which can have an significant impact on the secrecy rates. The other is that different users' signals are protected in an unequal way, since the users' signals are amounted on different power levels and hence the users experience different security performance. Advanced resource allocation algorithms are needed to ensure that users' different security targets can be met.

Compared to the scenario with external eavesdroppers, the scenario in which some NOMA users are potential eavesdroppers is more challenging. For example, how to avoid a NOMA strong user decoding a weak user's information is a million-dollar question, to which no answer has been found. However, for some special cases, NOMA transmission is shown to be more secure than OMA, even if some legitimate users are potential eavesdroppers. For example, in [84], multicasting and unicasting services are integrated together by using the NOMA principle, where multicasting receivers are legitimate users in the system but may want to intercept the signal sent to the unicasting receiver. Without using NOMA, unicasting and multicasting transmission are separated, which means that the multicasting receivers can easily intercept the unicasting signals. By using the NOMA principle, the base station can use the multicasting signals as a type of jamming information to degrade the capabilities of the multicasting receivers for intercepting the unicasting signals. But it is still not clear whether this idea can be extended to general NOMA scenarios beyond joint multicast-unicast transmission.

D. NOMA Assisted Radio Frequency and VLC Based Energy Harvesting

Radio frequency based energy harvesting (RFEH), also termed as simultaneous wireless information and power transfer (SWIPT), is to use radio frequency signals for two purposes, namely energy harvesting and information delivery. RFEH is particularly important to energy constrained wireless networks, where communication nodes have limited energy supplies and have no access to conventional energy sources, such as solar power, winder power, etc [85]. The interaction between the two technologies, NOMA and RFEH, is bidirectional. On one hand, NOMA can be very useful to efficient

RFEH, e.g., a base station can superimpose two types of signals by using the NOMA principle, one for information transfer and the other for energy harvesting. On the other hand, RFEH is also important to many NOMA communication scenarios [86]. Take cooperative NOMA as an example, where a strong user is to act as a relay and help a weak user. Because relay transmission consumes the strong user's battery life, the strong user might not want to help the weak user, in order to save the energy. By applying RFEH, the strong user can harvest some energy from the radio frequency signals sent by the base station, which will be used to power the relay transmission. In this way, the strong user helps the weak user without reducing its own batter life, which provides more incentives for user cooperation. In addition to cooperative NOMA, RFEH has also been shown to be useful to joint NOMA uplink and downlink transmission [87]. It is worth pointing out that the use of RFEH makes the transmission power and users' channels coupled, which means that the performance analysis and system optimization in NOMA-RFEH are more challenging than conventional networks.

Using radio frequency signals is not the only way to charge nodes wirelessly, and recently the use of visible light as an alternative for RFEH has attracted a lot of attention [88], [89]. The main motivation to use visible light instead of radio frequency signals for energy harvesting is the safety concern, as explained in the following:

- At the transmitter side, there is a strict cap for the radio frequency transmission power, because of the public concern about electromagnetic pollution. Because of this power constraint and also path loss of radio frequency propagation, the amount of energy harvested at the receiver based RFEH can be quite small.
- At the receiver side, the receive power density is also strictly regulated, which means that the use of advanced smart antenna techniques to make the energy beamed to the desired location and hence improve the energy efficiency may not be allowed.

Compared to the aforementioned difficulty of RFEH, the use of visible light will cause less safety concerns. In addition, compared to energy harvesting circuits using radio frequency signals, the ones using light for energy harvesting are much more mature and cheaper. With these advantages, it is also worth pointing out that VLC channels are not as well understood as radio frequency channels, which imposes many challenges for the design and analysis for the system using light for energy harvesting and information transfer.

VI. Conclusions

In this paper, we have provided a detailed survey to illustrate the impact of the emerging communication technique, NOMA, on future wireless networks. Particularly, how the NOMA principle affects the design of the next generation multiple access techniques has been introduced first, where different practical forms of NOMA have also been described. Then, the applications of NOMA to other advanced communication techniques, such as wireless caching, MIMO techniques, and mmWave communications, and cooperative relaying, have been discussed. The impact of NOMA on communication systems beyond cellular networks has also been illustrated, by using digital TV, satellite communications, vehicular networks and VLC, as examples. Finally, important directions for future research about NOMA, such as the implementation of NOMA with imperfect CSI, the applications of NOMA to physical layer security, energy harvesting and full duplex transmission, have also been outlined in the paper.

References

[1] J. Proakis, *Digital Commun.*, 4th ed. McGraw-Hill, New York, 2000.

[2] T. Cover and J. Thomas, *Elements of Information Theory*, 6th ed. Wiley and Sons, New York, 1991.

[3] S. Verdu, *Multiuser Detection*. Cambridge University Press, Cambridge, UK, 1998.

[4] "5G, a techology vision," Huawei, Inc., Shengzheng, China, 5G Whitepaper, Mar. 2015.

[5] "5G radio access: Requirements, concepts and technologies," NTT DOCOMO, Inc., Tokyo, Japan, 5G Whitepaper, Jul. 2014.

[6] "5G innovation opportunities- a discussion paper," techUK, London, 5G Whitepaper, Aug. 2015.

[7] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. Bhargava, "A survey on non-orthogonal multiple access for 5G networks: Research challenges and future trends," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2181 –2195, 2017.

[8] Z. Wei, J. Yuan, D. W. K. Ng, M. Elkashlan, and Z. Ding, "A survey of downlink non-orthogonal multiple access for 5G wireless communication networks," *ZTE Commun.*, vol. 14, no. 4, pp. 17–26, Oct. 2016.

[9] 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.

[10] "World's first successful 5G trial using smartphone-sized NOMA chipset-embedded device to increase spectral efficiency," *DOCOMO Press Release*, Nov. 2017.

[11] Z. Ding, L. Dai, and H. V. Poor, "MIMO-NOMA design for small packet transmission in the Internet of Things," *IEEE Access*, vol. 4, pp. 1393–1405, Aug. 2016.

[12] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System level performance evaluation of downlink non-orthogonal multiple access (NOMA)," in *Proc. IEEE Int. Symposium on Personal, Indoor and Mobile Radio Commun.*, London, UK, Sept. 2013.

[13] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.

[14] J. Choi, "Power allocation for max-sum rate and max-min rate proportional fairness in NOMA," *IEEE Commun. Lett.*, vol. 20, no. 10, pp. 2055–2058, Oct. 2016.

[15] N. Nonaka, A. Benjebbour, and K. Higuchi, "System-level throughput of NOMA using intra-beam superposition coding and SIC in MIMO downlink when channel estimation error exists," in *Proc. IEEE Int. Conf. on Commun. Systems*, Macau, China, Nov 2014.

[16] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G non-orthogonal multiple access," *IEEE Trans. Veh. Tech.*, vol. 65, no. 8, pp. 6010–6023, Aug. 2016.

[17] R. Mitra and V. Bhatia, "Precoded Chebyshev-NLMS-Based predistorter for nonlinear LED compensation in NOMA-VLC," *IEEE Trans. Commu.*, vol. 65, no. 11, pp. 4845–4856, Nov. 2017.

[18] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "A general power allocation scheme to guarantee quality of service in downlink and uplink NOMA systems," *IEEE Trans. on Wireless Commun.*, vol. 15, no. 11, pp. 7244–7257, Nov 2016.

[19] Y. Sun, D. W. K. Ng, Z. Ding, and R. Schober, "Optimal joint power and subcarrier allocation for full-duplex multicarrier non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 65, no. 3, pp. 1077–1091, Mar. 2017.

[20] K. Yakou and K. Higuchi, "Downlink NOMA with SIC using unified user grouping for non-orthogonal user multiplexing and decoding order," in *Proc. Int. Symposium on Intelligent Signal Processing and Commu. Systems (ISPACS)*, Nov. 2015.

[21] X. Zhang, Q. Gao, C. Gong, and Z. Xu, "User grouping and power allocation for NOMA visible light communication multi-cell networks," *IEEE Commu. Lett.*, vol. 21, no. 4, pp. 777–780, Apr. 2017.

[22] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, Cambridge, UK, 2003.

[23] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. IEEE Int. Symposium on Personal Indoor and Mobile Radio Commun.*, London, UK, Sept. 2013.

[24] M. Taherzadeh, H. Nikopour, A. Bayesteh, and H. Baligh, "SCMA codebook design," in *Proc. IEEE Veh. Tech. Conf.*, Las Vegas, NV, US, Sept. 2014.

[25] Z. Yang, J. Cui, X. Lei, Z. Ding, P. Fan, and D. Chen, "Impact of factor graph on average sum rate for uplink sparse code multiple access systems," *IEEE Access*, vol. 4, pp. 6585–6590, Oct. 2016.

[26] D. Cai, P. Fan, X. Lei, Y. Liu, and D. Chen, "Multi-dimensional SCMA codebook design based on constellation rotation and interleaving," in *Proc. IEEE Veh. Tech. Conf.*, Nanjing, China, May 2016.

[27] L. Yu, P. Fan, X. Lei, and P. T. Mathiopoulos, "BER analysis of SCMA systems with codebooks based on Star-QAM signaling constellations," *IEEE Commu. Lett.*, vol. 21, no. 9, pp. 1925–1928, Sept. 2017.

[28] L. Yu, P. Fan, Z. Ma, X. Lei, and D. Chen, "An optimized design of irregular SCMA codebook based on rotated angles and EXIT chart," in *Proc. IEEE Veh. Tech. Conf.*, Montreal, QC, Canada, Sept. 2016.

[29] S. Chen, B. Ren, Q. Gao, S. Kang, S. Sun, and K. Niu, "Pattern division multiple access (PDMA) - a novel non-orthogonal multiple access for 5G radio networks," *IEEE Trans. Veh. Tech.*, vol. 66, no. 4, pp. 3185–3196, 2017.

[30] E. Bastug, M. Bennis, and M. Debbah, "Living on the edge: The role of proactive caching in 5G wireless networks," *IEEE Commun. Mag.*, vol. 52, no. 8, pp. 82–89, Aug. 2014.

[31] N. Golrezaei, A. F. Molisch, A. G. Dimakis, and G. Caire, "Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution," *IEEE Commun. Mag.*, vol. 51, no. 4, pp. 142–149, Apr. 2013.

[32] M. A. Maddah-Ali and U. Niesen, "Fundamental limits of caching," *IEEE Trans. Inform. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.

[33] X. Xu and M. Tao, "Modeling, analysis, and optimization of coded caching in small-cell networks," *IEEE Trans. Commun.*, vol. 65, no. 8, pp. 3415–3428, Aug. 2017.

[34] Z. Chen and M. Kountouris, "D2D caching vs. small cell caching: Where to cache content in a wireless network?" in *Proc. Int. Workshop on Signal Processing Advances in Wireless Commun.*, Edinburgh, UK, Jul. 2016.

[35] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "NOMA assisted wireless caching: Strategies and performance analysis," *IEEE Trans. Commun.*, (submitted) Available on-line at arXiv:1709.06951.

[36] G. Foschini and M. Gans, "On limits of wireless communication in a fading environment when using multiple antennas," *Wireless Personal Commun.*, vol. 6, no. 3, pp. 311–335, Mar. 1998.

[37] P. Xu, Z. Ding, X. Dai, and H. V. Poor, "A new evaluation criterion for non-orthogonal multiple access in 5G software defined networks," *IEEE Access*, vol. 3, pp. 1633–1639, 2015.

[38] Z. Chen, Z. Ding, P. Xu, and X. Dai, "Optimal precoding for a QoS optimization problem in two-user MISO-NOMA downlink," *IEEE Commun. Lett.*, vol. 20, no. 6, pp. 1263–1266, Jun. 2016.

[39] Z. Chen, Z. Ding, X. Dai, and G. K. Karagiannidis, "On the application of quasi-degradation to MISO-NOMA downlink," *IEEE Trans. Signal Process.*, vol. 64, no. 23, pp. 6174–6189, Dec. 2016.

[40] M. F. Hanif, Z. Ding, T. Ratnarajah, and G. K. Karagiannidis, "A minorization-maximization method for optimizing sum rate in non-orthogonal multiple access systems," *IEEE Trans. Signal Process.*, vol. 64, no. 1, pp. 76–88, Jan. 2016.

[41] M. Zeng, A. Yadav, O. A. Dobre, G. I. Tsiropoulos, and H. V. Poor, "Capacity comparison between MIMO-NOMA and MIMO-OMA with multiple users in a cluster," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2413–2424, Oct. 2017.

[42] J. Choi, "On the power allocation for MIMO-NOMA systems with layered transmissions," *IEEE Trans. Wireless Commun.*, vol. 15, no. 5, pp. 3226–3237, May 2016.

[43] Z. Ding, F. Adachi, and H. V. Poor, "The application of MIMO to non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 1, pp. 537–552, Jan. 2016.

[44] Z. Ding, R. Schober, and H. V. Poor, "A general MIMO framework for NOMA downlink and uplink transmissions based on signal alignment," *IEEE Trans. Wireless Commun.*, vol. 15, no. 6, pp. 4438–4454, Jun. 2016.

[45] Z. Ding and H. V. Poor, "Design of massive-MIMO-NOMA with limited feedback," *IEEE Signal Process. Lett.*, vol. 23, no. 5, pp. 629–633, May 2016.

[46] D. Zhang, Y. Liu, Z. Ding, Z. Zhou, A. Nallanathan, and T. Sato, "Performance analysis of non-regenerative massive-MIMO-NOMA relay systems for 5G," *IEEE Trans. Commu.*, vol. 65, no. 11, pp. 4777–4790, Nov. 2017.

[47] M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "A comparison of MIMO techniques in downlink millimeter wave cellular networks with hybrid beamforming," *IEEE Trans. Commun.*, vol. 64, no. 5, pp. 1952–1967, May 2016.

[48] R. W. Heath, N. Gonzalez-Prelcic, S. Rangan, W. Roh, and A. M. Sayeed, "An overview of signal processing techniques for millimeter wave MIMO systems," *IEEE J. Sel. Topics Signal Process.*, vol. 10, no. 3, pp. 436–453, Apr. 2016.

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

[50] Z. Zhang, Z. Ma, Y. Xiao, M. Xiao, G. K. Karagiannidis, and P. Fan, "Non-orthogonal multiple access for cooperative multicast millimeter wave wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 8, pp. 1794–1808, Aug. 2017.

[51] A. Alkhateeb, Y. H. Nam, J. Zhang, and R. W. Heath, "Massive MIMO combining with switches," *IEEE Wireless Commun. Lett.*, vol. 5, no. 3, pp. 232–235, Jun. 2016.

[52] X. Gao, L. Dai, Y. Sun, S. Han, and C.-L. I, "Machine learning inspired energy-efficient hybrid precoding for mmWave massive MIMO systems," in *Proc. IEEE Int. Conf. on Commun.*, Paris, France, May 2017.

[53] Z. Ding, L. Dai, R. Schober, and H. V. Poor, "NOMA meets finite resolution analog beamforming in massive MIMO and millimeter-wave networks," *IEEE Commun. Lett.*, vol. 21, no. 8, pp. 1879–1882, Aug. 2017.

[54] Z. Ding, M. Peng, and H. V. Poor, "Cooperative non-orthogonal multiple access in 5G systems," *IEEE Commun. Lett.*, vol. 19, no. 8, pp. 1462–1465, Aug. 2015.

[55] Z. Wei, L. Dai, D. W. K. Ng, and J. Yuan, "Performance analysis of a hybrid downlink-uplink cooperative NOMA scheme," in *Proc. IEEE Veh. Tech. Conf.*, Sydney, NSW, Australia, Jun. 2017.

[56] L. Lv, Q. Ni, Z. Ding, and J. Chen, "Application of non-orthogonal multiple access in cooperative spectrum-sharing networks over Nakagami-m fading channels," *IEEE Trans. Veh. Tech.*, vol. 66, no. 6, pp. 5506–5511, Jun. 2016.

[57] J.-B. Kim and I.-H. Lee, "Non-orthogonal multiple access in coordinated direct and relay transmission," *IEEE Commun. Lett.*, vol. 19, no. 11, pp. 2037–2040, Nov. 2015.

[58] S. Luo and K. C. Teh, "Adaptive transmission for cooperative NOMA system with buffer-aided relaying," *IEEE Commu. Lett.*, vol. 21, no. 4, pp. 937–940, Apr. 2017.

[59] Z. Ding, H. Dai, and H. V. Poor, "Relay selection for cooperative NOMA," *IEEE Wireless Commun. Lett.*, vol. 5, no. 4, pp. 416–419, Aug. 2016.

[60] Z. Yang, Z. Ding, Y. Wu, and P. Fan, "Novel relay selection strategies for cooperative NOMA," *IEEE Trans. Veh. Tech.*, vol. 66, no. 11, pp. 10 114–10 123, Nov. 2017.

[61] R. Molina-Masegosa and J. Gozalvez, "LTE-V for sidelink 5G V2X vehicular communications: A new 5G technology for short-range vehicle-to-everything communications," *IEEE Veh. Tech. Mag.*, vol. 12, no. 4, pp. 30–39, Dec. 2017.

[62] S. Chen, J. Hu, Y. Shi, Y. Peng, J. Fang, R. Zhao, and L. Zhao, "Vehicle-to-everything (V2X) services supported by LTE-based systems and 5G," *IEEE Commu. Stand. Mag.*, vol. 1, no. 2, pp. 70–76, Jan. 2017.

[63] I. W. H. Ho, K. K. Leung, and J. W. Polak, "Stochastic model and connectivity dynamics for VANETs in signalized road systems," *IEEE/ACM Trans. Networking*, vol. 19, no. 1, pp. 195–208, Feb. 2011.

[64] B. Di, L. Song, Y. Li, and G. Y. Li, "Non-orthogonal multiple access for high-reliable and low-latency V2X communications in 5G systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2383–2397, Oct. 2017.

[65] T. Komine and M. Nakagawa, "Fundamental analysis for visible-light communication system using LED lights," *IEEE Trans. Consumer Electronics*, vol. 50, no. 1, pp. 100–107, Feb. 2004.

[66] L. Yin, W. O. Popoola, X. Wu, and H. Haas, "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5162–5175, Dec. 2016.

[67] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat, "Non-orthogonal multiple access for visible light communications," *IEEE Photonics Tech. Lett.*, vol. 28, no. 1, pp. 51–54, Jan. 2016.

[68] L. Fay, L. Michael, D. Gomez-Barquero, N. Ammar, and M. W. Caldwell, "An overview of the ATSC 3.0 physical layer specification," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 159–171, Mar. 2016.

[69] L. Zhang, W. Li, Y. Wu, X. Wang, S. I. Park, H. M. Kim, J. Y. Lee, P. Angueira, and J. Montalban, "Layered-division-multiplexing: Theory and practice," *IEEE Trans. Broadcast.*, vol. 62, no. 1, pp. 216–232, Mar. 2016.

[70] X. Zhu, C. Jiang, L. Kuang, N. Ge, and J. Lu, "Non-orthogonal multiple access based integrated terrestrial-satellite networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2253–2267, Oct. 2017.

[71] M. Caus, M. A. Vazquez, and A. Perez-Neira, "NOMA and interference limited satellite scenarios," in *Proc. Asilomar Conference on Signals, Systems and Computers*, Pacific Grove, CA, USA, Nov. 2016.

[72] G. T. Zhou, M. Viberg, and T. McKelvey, "A first-order statistical method for channel estimation," *IEEE Signal Processing Lett.*, vol. 10, no. 3, pp. 57–60, Mar. 2003.

[73] Z. Yang, Z. Ding, P. Fan, and G. K. Karagiannidis, "On the performance of non-orthogonal multiple access systems with partial channel information," *IEEE Trans. Commun.*, vol. 64, no. 2, pp. 654–667, Feb. 2016.

[74] P. Xu, Y. Yuan, Z. Ding, X. Dai, and R. Schober, "On the outage performance of non-orthogonal multiple access with 1-bit feedback," *IEEE Trans. Wireless Commun.*, vol. 15, no. 10, pp. 6716–6730, Oct. 2016.

[75] J. Lee and T. Q. S. Quek, "Hybrid full-/half-duplex system analysis in heterogeneous wireless network," *IEEE Trans. Wireless Commun.*, vol. 14, no. 5, pp. 2883–2895, May, 2017.

[76] Z. Zhang, Z. Ma, M. Xiao, Z. Ding, and P. Fan, "Full-duplex device-to-device aided cooperative non-orthogonal multiple access," *IEEE Trans. Veh. Tech.*, vol. 66, no. 5, pp. 4467–4471, May 2017.

[77] L. Zhang, J. Liu, M. Xiao, G. Wu, Y.-C. Liang, and S. Li, "Performance analysis and optimization in downlink NOMA systems with cooperative full-duplex relaying," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 10, pp. 2398–2412, Oct. 2017.

[78] C. Zhong and Z. Zhang, "Non-orthogonal multiple access with cooperative full-duplex relaying," *IEEE Commun. Lett.*, vol. 20, no. 12, pp. 2478–2481, Dec. 2016.

[79] M. S. Elbamby, M. Bennis, W. Saad, M. Debbah, and M. Latva-aho, "Resource optimization and power allocation in full duplex non-orthogonal multiple access (FD-NOMA) networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2860–2873, Dec. 2017.

[80] Z. Ding, P. Fan, and H. V. Poor, "On the coexistence between full-duplex and NOMA," *IEEE Wireless Commun. Lett.*, (submitted).

[81] Y. Zhang, H. M. Wang, Q. Yang, and Z. Ding, "Secrecy sum rate maximization in non-orthogonal multiple access," *IEEE Commun. Lett.*, vol. 20, no. 5, pp. 930–933, May 2016.

[82] Y. Zhang, H. M. Wang, T. X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Tech.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[83] D. Xu, P. Ren, Q. Du, L. Sun, and Y. Wang, "Combat eavesdropping by full-duplex technology and signal transformation in non-orthogonal multiple access transmission," in *Proc. IEEE Int. Conf. Commu.*, Paris, France, May 2017.

[84] Z. Ding, Z. Zhao, M. Peng, and H. V. Poor, "On the spectral efficiency and security enhancements of NOMA assisted multicast-unicast streaming," *IEEE Trans. Commun.*, vol. 65, no. 7, pp. 3151–3163, Jul. 2017.

[85] L. Liu, R. Zhang, and K.-C. Chua, "Wireless information transfer with opportunistic energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 288 –300, Jan. 2013.

[86] Y. Liu, Z. Ding, M. Elkashlan, and H. V. Poor, "Cooperative non-orthogonal multiple access with simultaneous wireless information and power transfer," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 938–953, Apr. 2016.

[87] P. D. Diamantoulakis, K. N. Pappi, Z. Ding, and G. K. Karagiannidis, "Wireless-powered communications with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 15, no. 12, pp. 8422–8436, Dec. 2016.

[88] G. Pan, J. Ye, and Z. Ding, "On secure VLC systems with spatially random terminals," *IEEE Commu. Lett.*, vol. 21, no. 3, pp. 492–495, Mar. 2017.

[89] ——, "Secure hybrid VLC-RF systems with light energy harvesting," *IEEE Trans. on Commu.*, vol. 65, no. 10, pp. 4348–4359, Oct. 2017.