A Survey on Applications of Cache-Aided NOMA

Dipen Bepari, Soumen Mondal, Aniruddha Chandra, *Senior Member, IEEE*,
Rajeev Shukla, *Student Member, IEEE*, Yuanwei Liu, *Senior Member, IEEE*,
Mohsen Guizani, *Fellow, IEEE*, and Arumugam Nallanathan, *Fellow, IEEE*

***Abstract*—Contrary to orthogonal multiple-access (OMA), non-orthogonal multiple-access (NOMA) schemes can serve a pool of users without exploiting the scarce frequency or time domain resources. This is useful in meeting the future network requirements (5G and beyond systems), such as, low latency, massive connectivity, users' fairness, and high spectral efficiency. On the other hand, content caching restricts duplicate data transmission by storing popular contents in advance at the network edge which reduces data traffic. In this survey, we focus on cache-aided NOMA-based wireless networks which can reap the benefits of both cache and NOMA; switching to NOMA from OMA enables cache-aided networks to push additional files to content servers in parallel and improve the cache hit probability. Beginning with fundamentals of the cache-aided NOMA technology, we summarize the performance goals of cache-aided NOMA systems, present the associated design challenges, and categorize the recent related literature based on their application verticals. Concomitant standardization activities and open research challenges are highlighted as well.**

***Index Terms*—Caching, Non-orthogonal multiple access, Standardization.**

I. Introduction

DESPITE many challenges, there had been several successful trial-runs and limited-scale commercial deployments of the fifth generation (5G) cellular networks across the globe over the last couple of years [1], [2]. 5G implementations are, however, vastly heterogeneous as its three major use cases have conflicting requirements: enhanced mobile broadband (eMBB) promises Gbps connectivity on the go, massive machine type communication (mMTC) requires support for extremely high node density and low transmission power to enhance network lifetime, while ultra-reliable low-latency communication (URLLC) demands immediate response from a resilient network. Sixth generation (6G) networks aspire to touch all these three cornerstones, eMBB, mMTC and URLLC, all at once [3], but it is difficult to satisfy the data-rate, spectral-efficiency, and low-latency constraints, simultaneously. For example, a fully autonomous (level 5 autonomy) connected vehicle alone would generate 19 terabytes (TB) of sensory data per hour [4]. If we consider the future internet-of-everything (IoE), including bio, nano and space domains, the data we need to transfer over the backhaul network in an hour can easily reach the order of zettabytes (ZB; 1 ZB $\sim 10^9$ TB). The challenge is to deliver such a huge amount of data over limited-bandwidth links maintaining the strict latency constraints. 6G aims to break the *millisecond latency barrier* [5], while some applications, such as haptic interactions through a tactile internet, demand an end-to-end delay even lesser than 1 ms [6]. The pressure for a faster network is mounting as many game-changing device technologies are now mature enough to be prototyped: three-dimensional (3D) holographic displays are ready for thin gadgets [7], virtual reality/ augmented reality (VR/AR) microscopes are detecting cancer cells [8] and human body integrated wireless surfaces are being built for providing users with a truly-immersive extended reality (XR) experience [9]. Development of these devices further intensifies the struggle for building an agile 6G network. Undoubtedly, *backhaul is our next frontier*.

Cellular networks are more centralized than wired ones and *caching* has been under consideration for improving the backhaul latency since the introduction of 3G [10]. Considering the fact that *propagation delay is just one of the many components of the overall end-to-end delay*, and in addition there would be transmission delay (links are not of infinite capacity), buffering time (every in-between node has a finite storage capacity), multiple access delay (you are not the only one who is active), etc. the effective radius shrinks further down. Thus, either you have to bring the regional data centers (RDCs) to your locality or make sure the content is available (at least partially), in a local manner, i.e., perform caching.

The need for caching is more relevant than ever before with the paradigm shift in the internet protocol (IP) traffic pattern. In 2021, IP video consisted of 82% of total internet traffic. The surge is due to the increasing popularity of all three types of video services, namely free (e.g., YouTube), subscription-based (e.g., Netflix) and social media (e.g., WhatsApp). In 2016, the video traffic consumed by mobile users was almost equal to the PC users. Five years later, the ratio is now heavily skewed, the mobile users are consuming videos 3 times as much as the PC users. The request for specific high-quality multimedia content with low latency, irrespective of users' locations, converted the communication-centric networks to content-centric networks. A major amount of backhaul traffic

Manuscript received 29 April 2022; revised 31 October 2022 and XXX xx, 2023; accepted XXX xx, 2023. Date of publication XXX xx, 2023; date of current version XXX xx, 2023.

D. Bepari is with the Department of Electronics and Communication Engineering, National Institute of Technology Raipur, Chhattisgarh-492010, India (e-mail: dipen.jgec04@gmail.com).

S. Mondal, A. Chandra and R. Shukla are with the Department of Electronics and Communication Engineering, National Institute of Technology Durgapur, West Bengal-713209, India (e-mail: soumen.durgapur@gmail.com, aniruddha.chandra@ieee.org, rs.20ec1103@phd.nitdgp.ac.in).

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. (e-mail: yuanwei.liu@qmul.ac.uk, a.nallanathan@qmul.ac.uk).

M. Guizani is with the Machine Learning Department, Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. (e-mail: mguizani@ieee.org).

is due to frequently transmitting replica of the same content (say, Despacito by Luis Fonsi or Baby Shark Dance). Roughly 5% of the webpages, audio and video files are popular, and a large number of users request these popular files at different time instants, impelling the network to provide the same content, again and again, using the backhaul link.

Caching can reduce both backhaul use and latency; popular contents, asked by the users frequently, are stored near the network edge (e.g. at base stations, users' device) in advance during the off-peak period. When users request a common file, the network delivers the file from the cache without engaging the backhaul infrastructure all the way back to core. To store the popular contents in the cache, the network needs to access the backhaul links only once, thus avoiding accessing backhaul multiple times during peak hour [28]. Unlike, bandwidth and power, which are limited communication resources, content caching resources are adequately available, cost-effective, and suitably maintainable. Moreover, caching resources are growing following the Moore's law. *Installing memory for caching is cheaper than that of increasing backhaul capacity*; the retail price of a 2-3 TB memory is approximately 100 USD [29]. The non-causality characteristic of caching operation is particularly useful for mobile networks; in highly mobile 5G wireless environments caching at user equipment (UE) not only enhances the video streaming quality but also reduces the number of handovers, mitigates handover failure, and decreases energy consumption [30].

The cache technique is greatly compatible with many advanced communication systems, like millimeter-wave (mmWave) communications [31], [32], multiple-input multiple-output (MIMO) systems [33], [34], Mobile edge computing (MEC) [35], [36], terahertz communication [37] and others. However, our survey focuses on the cache-aided non-orthogonal multiple access (NOMA) technique. NOMA is one of the promising technologies for next-generation wireless communication [38], [39]. It is capable of efficiently realizing higher system throughput and spectral efficiency compared to the traditional orthogonal multiple access (OMA) [40]. Maintaining the fairness of users, NOMA serves multiple users simultaneously at the same frequency band/ time/code. The key idea of NOMA is to apply superposition coding at the transmitter for combing the signals of multiple users and successive interference cancellation (SIC) method at the receiver for decoding individual signal [12], [13], [41]. The SIC operation is a complex procedure, specifically, when AP serves a large number of users within a time-frequency block [42]. Cache-aided NOMA receivers exploiting the cache, when requested files of the other users are fully/partially available, can apply the cache-enabled interference cancellation (CIC) process and reduce the complexity of the SIC process [43]. A fundamental concept of the NOMA technique and its application in long-term evolution (LTE) and 5G have been reported in [44], [45]. It is also reported that the amalgamation of NOMA with cache technology can achieve a significant system performance enhancement. A cache-aided NOMA network reaps benefits from both the cache and NOMA techniques.

A. Motivation

Some of the existing literature explore various insights of cache strategies and NOMA principles separately. Table I illustrates the primary focus of the survey papers on NOMA and caching techniques. These articles analyze different features of NOMA and cache individually. Along with the state-of-the-art NOMA techniques for future networks (e.g., 5G and beyond system), the fundamental operating principles of NOMA and their comparative performance analysis over the OMA are the primary focus of [13]. Ding *at el.* provided a broader overview of research innovation of NOMA and their applications in advanced communications along with associated implementation challenges and constraints [12], [18]. The resource allocation and the performance analysis of MIMO-NOMA systems and the comparison between Welch-bound equality spread multiple access (WSMA)-based NOMA and multi-user-MIMO are reported in [14] and [17], respectively. Vaezi *at el.* [15] analyzed an interplay between NOMA and other technologies, like MIMO, massive MIMO, mmWave communications, cognitive communications, visible light communications, etc., and summarized how these combined technologies elevate network performance in terms of scalability, spectral efficiency, energy efficiency, etc.. The survey paper [16] analyzed various optimization scenarios to investigate only the maximum achievable sum-rate when PD-NOMA amalgamates with other promising technologies for 5G and beyond 5G (B5G) and ignored the analysis of other crucial performance metrics. The authors in [15], [16], completely overlooked a systematic analysis of wireless networks combining NOMA and cache technologies. The authors explored a comparative study of various adopted approaches (such as optimization techniques, analytical methods, game theory, matching theory, graph theory, machine learning techniques, etc.) to address the problem of resource allocation, signaling, and practical implementation of NOMA technologies in [18]. In this paper, the authors analyze the security aspects of NOMA technologies. Recently, Yahya *at el.* first enlightens the error rate performance of various NOMA configurations and designs in a holistic manner [19].

In [25], the authors studied recent development on the green content caching technique to explore various cache-equipped wireless networks, research-gap, solution methods, and application areas. In [24], Liying *et al.* presented a fundamental concept of caching techniques and their recent development in various types of cellular networks such as macro-cellular, heterogeneous, device-to-device, cloud-radio access, and fog-radio access networks. A few articles also studied the caching technique in cellular systems [23], [46]. In [23], the authors present the research challenges of cache-aided integrated networking in wireless communication systems. The survey [47] focused on caching techniques for vehicular communications. Although individual surveys on NOMA and caching do exist [13], [24], [25], explicit analysis of cache-aided NOMA systems and their applications have not been reported yet.

TABLE I: Timeline of existing surveys on NOMA and caching.

Tech.	Year	Ref.	Focus area	Metric	Application
NOMA	2016	[11]	Single carrier vs. multi-carrier NOMA	Sum rate	MIMO, CoCom
	2017	[12]	Single carrier vs. multi-carrier NOMA	Channel gain, Sum rate	MIMO, CoCom, mmWave
	2018	[13]	PD-NOMA vs. CD-NOMA	Spectral efficiency, Receiver complexity	LTE, 5G
	2018	[14]	NOMA vs. OMA	Sum rate, Outage probability	MIMO, CoCom
	2019	[15]	Coexistence of NOMA with other technologies	UL and DL architecture	MIMO, CoCom, mmWave, CR, VLC
	2020	[16]	PD-NOMA integration with other technologies	Optimal rate	MIMO, mmWave, CoMP, CR, VLC, UAV
	2020	[17]	Adoption of NOMA in future standards	Energy efficiency, end-to-end delay	5GNR
	2021	[18]	Security and resource allocation	PHY security, User pairing, Power allocation	MIMO, CoCom, CR, UAV, SWIPT
	2022	[19]	Error rate calculation	BER, SER, PEP, BLER	CoCom, VLC, FSO, IRS
Cache-aided NOMA (This paper)			Interplay of cache and NOMA	Sum rate, Delay, Decoding probability	MIMO, mmWave, SWIPT, IRS, V2X
Tech.	**Year**	**Ref.**	**Focus area**	**Architecture**	**Application**
Cache	2012	[20]	Caching in P2P	Web caching	CDN, P2P
	2013	[21]	Caching in international ICN projects	On-path, Off-path	ICN
	2013	[22]	ICN caching vs. web/P2P caching	ICN caching	ICN
	2018	[23]	Integration of networking, caching, computing	D2D, Cooperative	Wireless, Cloud, MEC
	2018	[24]	Cache enabled cellular networks	SBS, MBS, D2D	5G
	2020	[25]	Energy efficient caching	SBS, MBS, D2D	ICN
	2020	[26]	Cache utility to network operators	SBS, MBS, D2D	5G
	2022	[27]	Caching in IoT	ICN caching	IoT
Cache-aided NOMA (This paper)			Interplay of cache and NOMA	D2D, Cooperative, Edge	MEC, CFmMIMO

B. Contributions

The primary goal of this survey is to present a systematic study of the recent research development and innovations in the cache-aided NOMA systems. Numerous research articles analyze cache and NOMA-based wireless networks individually and exploit their benefits. However, to the best of our knowledge, this is the first article that introduces a survey on caching-aided NOMA systems and their practical applications in 5G and beyond systems. After a brief tutorial on the concept of wireless caching and NOMA, we explained the integration of NOMA with wireless caching by elaborating the underlining design principles, features and key performance indicators. We also presented a formal classification of cache-aided NOMA systems based on their diverse practical applications through highlighting the state-of-the-art, associated challenges, and promises. The forthcoming networks require massive data traffic to be carried over backhaul links. In this regard, we explored the fundamental impacts of cache-aided NOMA systems in terms of network efficiency, QoS and latency. Furthermore, this article presents a detailed account of the standardization activity and real-time development news of NOMA and cache technologies. Finally, this article identifies a wide range of potential future research opportunities and related technical challenges that need to be addressed for implementing cache-aided NOMA systems.

C. Organization

The organization of this paper is shown in Fig. 1. Section II provides an overview of cache technology. Section III and Section IV present an overview of cache-aided NOMA systems and explores various key performance indicator of cache-aided NOMA systems respectively. Next, Section V focuses on the application of cache-aided NOMA in the wireless communication domain, i.e., non-terrestrial networks, MEC, vehicle-to-everything (V2X) communication, cell-free massive MIMO, mmWave communications, etc. Section VI highlights standardization activities of cache-aided NOMA and identifies possible directions for future research on cache-aided NOMA. Finally, the conclusion of the paper is presented in Section VII. A list of acronyms is tabulated in Table II.

II. Overview of Cache Technology

The types of demand for content are changing day by day. In the early 1990s, Web pages and images were in high demand, and excess delivery of these contents was responsible for heavy network congestion. To cope with this problem Web caching technique becomes a promising approach for significantly reducing traffic load [48], [49]. In the early 2000s, the primary reason for network congestion was due to the demand for video content, and that was challenged by caching for content distribution networking [20] and information-centric networking (ICN) [22]. In ICN, the distance between cache and user was further shortened, and a new feature, *popularity of contents*, was incorporated in the

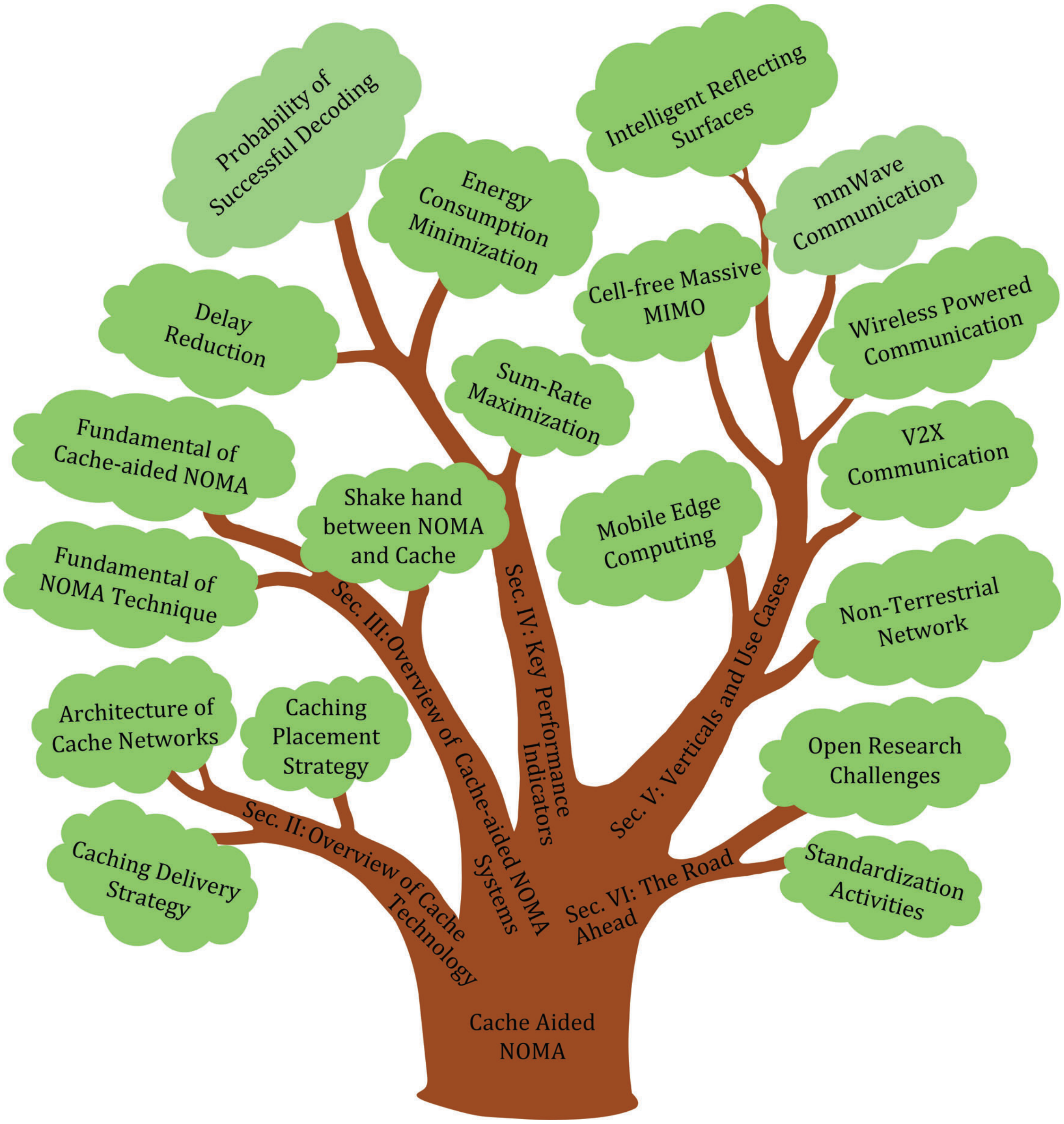

Probability of Successful Decoding
Intelligent Reflecting Surfaces
Energy Consumption Minimization
mmWave Communication
Delay Reduction
Cell-free Massive MIMO
Wireless Powered Communication
Sum-Rate Maximization
Fundamental of Cache-aided NOMA
V2X Communication
Shake hand between NOMA and Cache
Mobile Edge Computing
Fundamental of NOMA Technique
Sec. III: Overview of Cache-aided NOMA Systems
Sec. IV: Key Performance Indicators
Non-Terrestrial Network
Sec. V: Verticals and Use Cases
Architecture of Cache Networks
Caching Placement Strategy
Open Research Challenges
Sec. II: Overview of Cache Technology
Caching Delivery Strategy
Sec. VI: The Road Ahead
Standardization Activities
Cache Aided NOMA

Fig. 1: Organization of this survey.

content placement techniques [21]. Nowadays, users created video files and their delivery introduces additional network traffic load over wireless channels. Establishing a wireless caching network is more challenging than wired caching networks because of the unpredictable movement of the users and uncertain channel gain quality. Since 2009, researchers showing their interest in wireless caching for reducing traffic load and increasing the quality of communications [50], [51]. In 2010, the authors confirmed that the caching technique enhances the system throughput by as much as 400–500% [52]. The focus of the research was primarily limited to the development of the cache infrastructure, gateway design, routing, cooperative, and physical layers. However, after 2010, the researchers combined cache technology with other technologies and analyzed their performances in various wireless networks for next-generation communications.

The latency minimization is one of the stringent requirements and significant challenges for next-generation intelligent applications. Let us review the importance of caching from the 5G latency requirement perspective, where the projected latency is not more than 0.5 ms [53]. Electromagnetic (EM) waves travel at 3×10^5 km/s ($= c$) in an unguided medium and can cover a round-trip distance of 150 km over free space in 1 ms. However, most backhauls are not wireless, they are almost always created with the optical fiber. For a single-mode fiber having a core refractive index of 1.49 ($= n$), the velocity of light propagation becomes $v = c/n = 0.67c$, reducing the coverage to 100 km. Unless we invent a carrier that travels faster than light, regional data centers (RDCs) cannot be further away. Caching can be a great approach to minimizing the latency, where a few most popular contents are stored in a cache memory before being asked by users. The major challenges of wireless caching are:

TABLE II: List of acronyms frequently used in this survey

Acronyms	Description
AP	Access point
B5G	Beyond Fifth-Generation
CFmMIMO	Cell-free Massive Multi-Input Multiple-Output
CIC	Cache-enabled interference cancellation
CoCom	Cooperative Communication
CSI	Channel State Information
D2D	Device-to-Device
ETC	Edge Caching Technique
eMBB	Enhanced Mobile Broadband
ICN	Information Centric Networking
IoT	Internet of Thing
LFU	Least Frequently Used
LRU	Least Recently Used
LTE	long Term Evolution
MBS	Macro Base Station
MEC	Mobile Edge Computing
MIMO	Multiple-Input Multiple-Output
mMTC	Massive Machine Type Communication
mmWave	Millimeter-Wave
MUSA	Multiuser Shared Access
NOMA	Non-Orthogonal Multiple Access
OMA	Orthogonal Multiple Access
OTFS	Orthogonal Time-Frequency Space
PD-NOMA	Power-Domain NOMA
PER	Poll-Each-Read
QoS	Quality of Service
SBS	Small-cell Base Station
SCMA	Sparse Code Multiple Access
SDP	Successful Decoding Probability
SIC	Successive Interference Cancellation
SWPT	Simultaneous wireless information and power transfer
UAV	Unmanned Aerial Vehicle
UE	User Equipment
URLLC	Ultra-Reliable Low-Latency Communication
V2N	Vehicle-to-Network
V2V	Vehicle-to-Vehicle
V2X	Vehicle-to-everything
VLC	Visible Light Communications

- A wireless channel bandwidth is very limited compared to wired channels.
- A wireless channel gain quality is very poor due to noise, interference, shadowing, and so on.
- Wireless devices may be disconnected because of high mobility or/and poor channel gains.
- Limited battery life restricts the increase of transmit power.
- Limited cache memory demands efficient cache placement and access strategies.

The caching strategy involves two operating phases, the content placement phase, and the content delivery phase. In the content placement phase, the network stores popular contents in the cache memory during off-peak time. In the content delivery phase, the network serves the cached contents during peak traffic hours. Efficient caching placement with an updated strategy and delivery strategy is required to get maximum benefits from the caching strategies.

A. Caching Placement Strategy

Appropriate content placement is the baseline for achieving a significant performance gain from any cache-aided system. The caching placement strategy determines the size and location of files and decides how and where the selected files are to be downloaded to the cache memory. A content replacement mechanism that determines how to update the cached content regularly is an integral part of the cache placement strategy. Though a network requires a minimal up-gradation in the existing infrastructure for caching, significant challenges are involved with caching strategies. Because of variable content sizes, random demand for contents, limited cache resources, and the movement of the users, cache management becomes a challenging issue. Accommodating large numbers of files in the cache with limited memory space is one of the most severe issues.

The popularity of the files has to consider in the cache placement strategy for effectively reducing the use of the backhaul link in cache-aided systems. Increasing the availability of the requested files as much as possible is a key factor. Unpopular content selection for caching may lead to a considerable overhead cost [54]. The popularity of the randomly requested contents is widely modeled by the Zipf distribution [55]–[57]. The widely used Zipf distribution is proficient for measuring the polarity of video files [56]. In [58], based on the varying degree of popularity, the authors proposed a multiple-level non-uniform content popularity in their research work.

Two types of content placement strategy broadly found in literature- *coded placement strategy* [59]–[63] and *uncoded placement strategy* [64]–[66]. The basic principle of coded placement strategy is to divide the files into multiple small segments, encode the segments by a coding methods and place them in the cache memory. During the content delivery phase, a certain coding technique needs to employ to combine the requested files. Raptor codes [67] and fountain codes [68] are popularly used for combining the file segments. The uncoded placement strategy is comparatively simple where complete requested file or a portion of the file is kept in the cache.

The coded cache-enabled system with K users, cache capacity of F files, and N available files in the cache achieve a caching gain of $\frac{1}{1+\frac{KF}{N}}$ over the un-cached system [60]. The gain indicates a large amount of rate reduction in the shared link. The coded placement strategy is suitable for reducing cache memory consumption with increasing computational complexity, specifically for a large number of segment files [59]. The authors in [69], explore the advantages of coded caching strategies when cache-enabled access points (APs) like BS, SBS, MBS, etc.) are randomly distributed. In [70], authors have proposed a linear network coding-based cache content placement strategy that increases the amount of available data compared to triangular network coding. The coded caching strategies exploit coded multicast opportunities that further decrease the backhaul traffic, particularly for densely deployed cache-enabled APs [71]. Based on the random caching placement and multiple group cast index coding, the authors have proposed an order-optimal coded caching placement [72]. Niesen *et al.* [73] verified that optimal cache placement minimizes the traffic load of the shared link knowing the popularity distribution of files. Binbin *et al.* implemented an optimal cache content placement in cache-aided BS to reduce the backhaul traffic load of a wireless access network [74]. Jinbei *et al.* [71] presented an analysis that finds the lower bound on the data transmission rate of any coded caching strategy. Furthermore, for any popularity distributions and the system size, the authors also derived a constant factor that indicated the gap of achievable average transmission rate from the optimal.

A cache replacement strategy is accompanied by the cache placement strategy. It is an essential mechanism needed to employ for cache-aided systems. The least recently used (LRU) and least frequently used (LFU) are the traditional replacement policies found in the literature [48], [75]–[77]. The LRU replaces the least recently used files, and the LFU replaces the least frequently used files. The combination of cache access and replacement strategy makes a caching method. Various cache access and placement strategies and their performances are analyzed in [78]. A gain-based cache replacement policy named, Min-SAUD [79] and a hotspot-based caching scheme [80] are adopted to satisfy the caching replacement requirements. In [81], authors have proposed a cache replacement and content delivery strategy for regularly updating cached contents.

B. Caching Delivery Strategy

The availability of the requested files in the associated cache is not the only event that enhances the system performance, users need to receive and decode requested files in an error-free manner. In addition, a caching delivery strategy determines where the requested file will be transmitted from, the transmitting frequency band, the transmission power, and the encoding process so that files arrive at the requested user successfully and quickly. Poll-Each-Read (PER), Call-Back (CB), and Invalidated Report (IR) are some of the classical cache access schemes. BSs need to employ coding methods for coded transmission to combine user-requested files, whereas BSs deliver cached files individually for uncoded transmission techniques. Whenever any user requires a file, the network first searches it in the cache memory before downloading it from the internet server. Searching files in the cache may take substantial time, especially when the *miss rate* (cache memory fails to provide a shout file) is high due to a shortage of cache memory or/and storing less popular content in the cache [82]. Therefore, overall system performance primarily depends on how efficiently popular files are cached.

The backhaul traffic load and content delivery latency of cache-aided networks are subject to the cache memory size. The fundamental trade-off between the advantage of caching and the cache storage capacity has been studied in [62], [83] for coded and uncoded cache systems. They have analyzed the trade-off from an information-theoretic perspective, and carried out investigation based on the *normalized delivery time* metric, which measures the worst-case content delivery time subjected to the transmission rate of the requested files. Aiming to maximize the successful download probability of requested files, authors in [84] have optimized the cache memory size subjected to channel statistics, backhaul capacity, and distribution of file popularity in a cellular network. In [85], the authors have formulated an NP-hard optimization problem to minimize the time required to complete a file delivery for downlink transmission of a cache-aided network. Average latency in both the backhaul and cache link for delivering the requested file of a cache-enabled system under the constraint of quality of the recommended files has been minimized in [86]. Various types of caching strategies have been proposed, and the performance of these methods is analyzed mostly in terms of a cache hit rate.

C. Architecture of Cache Networks

Depending on the infrastructure for the downlink data transmission mechanism, the wireless caching network architectures are grouped into two categories. The first category is device-to-device (D2D) caching and the second category is edge caching.

1) D2D Caching Technique: In D2D caching, shown in Fig. 2(a), dedicated infrastructure for caching is not available. Users depend on the content stored by neighbors [87]–[89]. During the content placement phase, users store a few contents in their device, and during the content delivery phase, users communicate with the internet server through BS only when none of its neighbors has cached the requested file. A high density of users increases the availability of requested content at the nearby cache devices. D2D primarily relies on the cooperation of the neighboring users. D2D caching mainly improves spectral efficiency. Two types of D2D caching networks are found in the literature, *D2D Multihop relay* and *Cooperative D2D*.

D2D Multihop relay:- In D2D communication, intermediate users help to deliver the file from cache to the destination, as shown in Fig. 2(b). Multihop delivery empowers a user to access the desired file from a cache-user located far away [90], [91].

Cooperative D2D:- When a user requested file is available to multiple nearby cache the file can be transmitted cooperatively to the destination, as shown in Fig. 2(c). In this case, the implementation of MIMO technology accelerates the file transmission process. In [92], authors have considered a cooperative D2D caching for a wireless sensor network.

2) Edge Caching Technique: Unlike D2D Caching, in edge caching, dedicated caching infrastructure is available at the AP [93]–[95], as shown in Fig 2(d). In the content placement phase, popular contents are timely stored in the cache memory with high reliability before users ask for it. The primary aim of the content delivery phase is to provide the requested files without communicating with the internet server, thus enabling improvement in delivery latency. The latency of this technique is lower than that of in multihop transmission method. The edge caching technique primarily reduces the backhaul cost.

Cooperative Edge Caching Technique:- In this case, BSs communicate with the neighboring BS for the requested con-

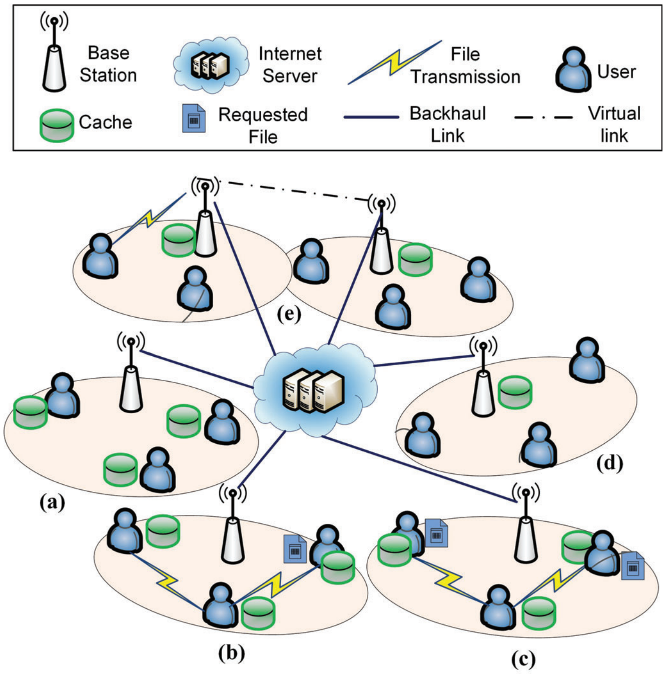


Fig. 2: Various architectures of cache networks (a) D2D caching technique, (b) D2D multihop relay, (c) Cooperative D2D, (d) edge caching technique, and (e) Cooperative edge caching technique

tents. When a user requested file is available neither in nearby cache device nor in its BS, the request is forwarded to the neighboring BS, and on the availability of the file in cache of the neighboring BS, the user can get its file via its own BS. Figure 2(e) shows the cooperative BSs delivery. Shan *et al.*, [96] analyses the performance of a cooperative edge caching in wireless cellular networks.

In the D2D networks, the users' devices are equipped with a cache, whereas in edge caching, APs are equipped with a dedicated cache facility. A few of the literature have considered hydride wireless caching networks, where cache infrastructure is available at both the transmitter (BS, SBS, and access points) and receiver (user) end [62], [97]. During the content placement phase, files are stored individually in the transmitter and users' devices, and during the content delivery, the BS searches its own and users' caches for the asked file before downloading from the internet servers. Fundamentally, edge caching is a centralized caching technique where based on the content popularity and network parameters, a BS decides which files at what time and where to be cached, and the BS also decides the requested file delivery strategies [55]. On the other hand, D2D caching is a distributed caching technique where each device determines cache placement and delivery strategy. Distributed caching algorithms are comparatively lesser complex than centralized ones but may fail to provide global optimality of the algorithms.

Summary:- The first and foremost task of establishing cache-aided networks is to make available the most popular contents in the cache before being requested. The Zipf distribution is the most commonly adopted approach for modeling the popularity of randomly requested contents. The popularity of the content, distribution of content popularity, correlation among contents, location of users, and finite storage availability constraints are required to be taken into consideration in the caching strategy, which introduces challenges to the caching strategies. Furthermore, a wide range of variety in the content and sudden change in popularity intensify difficulties and show the drawback of content placement during *off-hours*. Two online popularity prediction strategies, named the popularity prediction model (PPM) and the Grassmannian prediction model (GPM), were proposed to predict popularity in advance [98]. The rise of the content day by day may change statistical values of popularity distributions and content popularity. Therefore, caching demands an efficient cache placement strategy to encounter these changes, which increases challenges in designing such an algorithm [99]. Although codded caching techniques demand additional coding overhead, but extraordinarily enhance system performance in terms of reducing bandwidth requirements and transmission latency compared to that of the best-uncoded technique [59]. The researchers assume error-free channels during content pushing, which is questionable in real-time communication scenarios. Therefore, a more practical channel model needs to be taken into account during content placement.

III. Overview of cache-aided NOMA Systems

The primary aim of this section is to provide a brief overview of the NOMA operation and challenges associated with the cache-aided NOMA.

A. Fundamental of the NOMA Technique

With the increase of wireless devices, the multiple access (MA) technique becomes a popular approach to meet the demand of large numbers of wireless users. The MA techniques accommodate multiple users within the same resource blocks like frequency band, time slot, or spatial direction, concurrently. The MA can be categorized as OMA and NOMA. In OMA, resources like time, frequency, and code are orthogonal which reduces the interference introduced by other users. With the further advancement of wireless communication to fulfill the demand of massive connectivity, NOMA allows multiple users to use non-orthogonal resources simultaneously [100].

The power-domain NOMA (PD-NOMA) and code-domain NOMA (CD-NOMA) are two main types of NOMA. In PD-NOMA, a transmitter transmits signals to multiple users exploiting the power domain. The transmitter combines signals of multiple users with different power levels by applying superposition coding. Unlike the water-filling algorithm for power allocation in OMA [101], [102], in PD-NOMA, the total power is distributed among the users in such a way that signals of users with weaker channel conditions comparatively get more power than the signals of users with stronger channels. As a consequence, (i) a stronger user achieves a higher data transmission rate with low transmit power, and (ii) a weaker user experiences limited interference caused by stronger users, and simultaneously, higher power allocation to weaker users improves users' fairness, spectral efficiency, and sum-rate [103], [104]. At the receiver end, the receiver applies the SIC

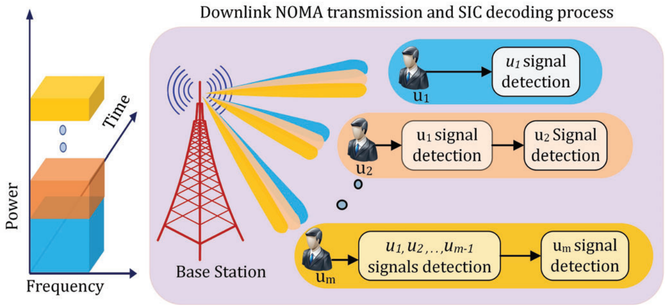


Fig. 3: Power distribution strategy for PD-NOMA. The transmitter allocates maximum power to user u_1 with the weakest channel and minimum power to user u_m with the strongest channel. u_1 decodes its own signal directly, and u_m first decodes signal of $u_1, u_2, ..., u_{m-1}$ then its own signal.

process to decode its signal. In the SIC process, the receiver first decodes the signal which has the highest assigned power and then subtracts the signal from the received composite signal. This process continues until the signal of the intended user is decoded [105]. The power allocation and decoding process of PD-NOMA is shown in Fig. 3.

The CD-NOMA assigns different codes to users and superimposes them over the same time and frequency. The operations of CD-NOMA are quite different from the PD-NOMA. The CD-NOMA is fundamentally built up on the concept of allocating different spreading codes that enable signals separation at the receiver. The multiuser shared access (MUSA) [106], sparse code multiple access (SCMA) [107], low density spreading (LDS) [108] are the main three types of CD-NOMA. In CD-NOMA, each user is supplied a codebook consisting of codewords. The transmitter first encodes the data of the user-requested content, then maps them into a complex codeword selected from the codebook and transmits over the channel allocating equal power to every user. The codebook of each user acts as a signature of the corresponding user. The BS delivers the user-requested content using CD-NOMA from the cache when it is available. Otherwise, the BS fetches it from the content server and then transmits it using the CD-NOMA principle. Though the CD-NOMA can remarkably enhance spectral efficiency, it requires a wide transmission bandwidth and considerable modification to the existing communication systems. On the contrary, PD-NOMA neither requires a major upgrade to the present communication networks nor a high transmission bandwidth [109]. In addition, PD-NOMA has a low complexity system compared to the CD-NOMA from a design perspective.

It is necessary to know the order of the average channel gains or instantaneous channel state information (CSI) for fixed power allocation [110] or dynamic channel allocation [111], respectively. The users' ordering is accomplished mostly based on the instantaneous value of the CSI [105], [112]. It is to note that a wrong user ordering leads to an incorrect choice of the power distribution, which may lead to a situation where a few users are always in outage [113]. Therefore, it is necessary to acquire the knowledge of the perfect channel gain. The availability of the perfect instantaneous CSI at the transmitter is not a valid assumption in practice, particularly when users of systems like 5G and B5G demand high mobile services. Unpredictable rapid users' movement frequently change channel characteristics that make the perfect CSI estimation process challenging. The error in the perfect channel estimation acts as a source of interference that degrades the overall system performance. The average sum-rate and the outage performance of the NOMA with a perfect CSI are always superior to that of the NOMA with imperfect CSI [114]. To give a *close to real-time* scenario, researchers consider imperfect CSI models for various network conditions such as slowly varying CSI, delayed CSI feedback, high mobility of users, and so on [115]–[118]. A second-order statistics (SOS)-based CSI model achieves superior system performance than the imperfect CSI based model [114].

Another challenging but key-enabling factor of NOMA is the implementation of an error-free SIC decoding. In imperfect SIC, the decoder completely cannot eliminate the signal power of other signals. Consequently, residue power affects the signal detection process as interference in subsequent signal detection. The imperfect SIC not only degrades the overall system performance but also increases processing time due to re-requesting for contents by the users who failed to decode their signals successfully. If a user fails to decode any signals with higher allocated power, it also fails to decode its signal (detailed discussed in section III-C2). The imperfect SIC is widely modeled as Gaussian distribution [119], [120]. However, there may be some typical scenarios where an error does not obey the Gaussian distribution [121]. One interesting fact of NOMA systems is that since the weakest user does not need to apply the SIC process for decoding its signal, there is no effect of imperfect SIC on the performance of the weakest user. A joint CSI- and QoS-based hybrid-SIC process is projected as a promising candidate for next generation multiple access in [122].

B. Fundamental of Cache-aided NOMA

Fundamentally, caching and NOMA are two completely different techniques; caching is a methodology for storing data temporarily in memory (cache), and NOMA is an advanced multiplexing technique for data transmission. Designing a NOMA-enabled system jointly with a cache facility can help both technologies. In a jointly developed cache-aided NOMA system, cache and NOMA individually help each other in their operations and enhance the performance of both techniques. Caching assists NOMA in the interference cancellation during the SIC process and increases the probability of successful decoding. On the other hand, NOMA helps a faster cache placement and delivery process than the OMA. Various studies validated the superiority of cache-aided NOMA over the conventional NOMA system in terms of energy efficiency [94], system latency [126], coverage performance [127], spectrum efficiency [128], successful decoding probability [129], and outage performance [61]. Caching with NOMA becomes an excellent communication technique that can provide support for next-generation communications.

The cache-aided NOMA evolved as an advanced communication concept for next-generation communications. In the

TABLE III: Comparison between cache-aided NOMA and cache-aided OMA

Parameters	Description
Content popularity	Contents popularity distribution for both the techniques is modeled as Zipf distributed.
Content selection	Content selection strategy does not depend on the multiplexing technique (OMA or NOMA).
Cache placement	• For cache-aided OMA, OMA is used for cache content placement and delivery. • For cache aided-NOMA, NOMA is used for cache content placement and delivery. • Content placement and delivery is faster in cache aided-NOMA than the cache aided-OMA.
Cache delivery	
System operation	• If the requested file is cached, the BS sends the file immediately for both techniques. • If requested file is un-cached, the BS first downloads the requested file from the data center using the backhaul link and then sends it to the users for both techniques.
System performance	• NOMA always performs better than any conventional OMA when both schemes are equipped with optimal resources [123]. • NOMA serves more users than the OMA under the same network condition [124]. • The average data rate and the spectral efficiency of NOMA-based system are higher than OMA [125].

OMA technique, users with better channel conditions get higher priority, and users with poorer channel conditions need to wait for access; that initiates fairness and high latency problems. On the other hand, NOMA serves multiple users with various channel gains simultaneously, which provides improved fairness with lower latency [44]. The performance of NOMA always surpasses any conventional OMA techniques when both are provided with the optimal resource allocation strategies [123]. The NOMA reaches an enhanced spectral efficiency and power efficiency over OMA [130] [131]. It also achieves an improved sum-rate and individual user rate over the time division multiple access (TDMA) [132]. Yang *et al.* validated that the NOMA outperforms traditional OMA in terms of outage probability even when partial CSI is available [114]. According to International Mobile Telecommunications (IMT) [133], 5G technology needs to support eMBB (requires 100Mbps user data rate), mMTC (needs to provide connectivity to 1 million devices per square kilometer), and URLLC (requires maximum 0.5ms end-to-end latency with reliability above 99.999% [53]). The OMA techniques cannot meet the above requirements. On the other hand, the NOMA technique efficiently improves the downlink and uplink spectral-efficiency by 30% and 100% respectively in eMBB compared to OMA [134]. NOMA-supported mMTC and URLLC applications can serve five and nine times more users, respectively [53]. The aforementioned advanced features encouraged an amalgamation of the concept of a caching technique with the NOMA over OMA. Table-III compares cache-aided NOMA and cache-aided OMA.

Due to the dynamic behaviour of the wireless channel and movement of the users, content placement and delivery become challenging in wireless caching networks. NOMA helps in fast content placement and delivery maintaining fairness. A caching strategy need to employ in cache-aided NOMA system to address the following challenges

- What and how to cache?
- What and when to update?
- How to design physical-layer transmission?

Superposition coding is widely used for combining signals in the NOMA technique. However, Yaru *et al.* proposed a method for combining signals, named index coding (IC), and claimed that IC is comparably more energy-efficient than superposition coding, particularly when requested files of a pair of users are available in the associated cache [135]. The SIC decoding process of a cache-aided NOMA network is slightly different from conventional NOMA [129]. Notably, caching helps in the decoding process even when the requested file is not cached (detail discussed in Sec III-C1).

A few cache placement techniques of NOMA systems has been summarised in Table IV. In [124], the authors proposed a cooperative NOMA- caching scheme to analyze the effect of physical storage of BS and available radio resource parameters like QoS, subcarrier assignment, and power allocation constraints in the network cost. The authors also validated that the proposed scheme can improve the network cost reduction compared to other caching strategies and OMA. An optimization problem for content placement in the NOMA system has been formulated to minimize the average transmit power taking into account cache capacity constraints [137]. In [138], the authors have optimized the cache placement strategy to reduce the average delay under the constraint of the storage capacity of a fog-computing AP. Xiang *et al.* proposed a coded caching delivery strategy and derived optimal transmit power and rate allocation based on cache status, file sizes, and channel conditions to minimize content delivery latency in cellular networks [85]. A joint cache content popularity prediction and access mode selection problem are formulated as the Stackelberg game in cache-aided NOMA-based F-RANs [139]. It is observed that a cache-aided network conventionally stores cache contents during an off-peak time, which may not be an efficient approach, particularly when a network frequently needs to update the content of the cache. Ding *et al.* proposed NOMA as the best candidate for efficiently storing cache contents during on-peak hours and developed two algorithms, *push-then-deliver strategy* and the *push-and-deliver strategy* [136], [143]. The push-then-deliver strategy stores popular content during on-peak times and delivers when requested. The push-and-deliver strategy deals with the scenario when a user's requested content is not cached.

C. Shake hand between NOMA and Cache

This subsection discusses how NOMA and cache jointly support each other in establishing next-generation communication systems. The caching assists NOMA in the interference cancellation and the successful decoding. On the other hand,

TABLE IV: Cache placement for cache-aided-NOMA systems

Ref.	Objective	System	Caching	Optimization	Technical Contribution	Major Outcome
[136]	On-peak hour content placement	Edge and D2D caching	Uncoded full file	Optimization is not employed	Proposed two on-peak hour cache content placement strategies	Improves cache hit and reduces delivery outage probability
[124]	Minimize the overall network cost	Cooperative edge caching	Uncoded full file	Hungarian algorithm and successive convex approximation method	Propose a novel joint resource allocation and cooperative caching scheme	Considerably reduce network cost compared to non-cooperative OMA
[137]	Minimizes BS average transmit power	D2D caching	Uncoded full file	Quadratic knapsack problem formulation	Proposed three methods for cache content placement	*Alternating upper plane* method is best for caching
[138]	Improves delivery delay and sum rate performance	Cooperative edge caching	Split file caching	Lagrange partial relaxation and McCormick envelopes methods	Optimized the cache placement strategy under the constraint of storage capacity	Proper resource utilization can achieve ultra-low latency and high throughput
[139]	Analyzes successful delivery probability and transmission rate	Edge caching	Uncoded full file	Hierarchical Stakelberg game theory method	Proposed joint cache content popularity prediction and access mode selection method	Proposed algorithm can achieve the evolutionary equilibrium very fast with 90% prediction effect
[140]	Improves spectrum efficiency and reduces outage probability	Edge caching	Uncoded full file	Gale-Shapley algorithm-based distributed method	Proposed NOMA-multicasting for pushing and multicasting content simultaneously	Proposed scheme is better than conventional OMA-based multicast scheme
[141]	Increases cache hit and outage performance	Edge caching	Uncoded full file	Optimization is not employed	Proposed a QoS-oriented dynamic power allocation strategy	Ensures correct detection of far user requested files and improves delay performance
[142]	Increases cache hit ratio and delivery delay performance	Edge caching	Uncoded full file	Q-learning based algorithm	Proposed long-term caching placement and resource allocation algorithm	Trade-off between performance and computational complexity exist

NOMA lends a helping hand to caching for placing multiple files within a short free window.

1) Interference Cancellation: The Cache-enabled interference cancellation (CIC) is an attractive feature of D2D cache-aided NOMA systems. During off-peak time, the BS stores popular content in the cache using the split file caching technique. Perfect knowledge of the cached content is available to the BS. Users get a few segments of the requested file from their associated cache memory and the rest of the portions from the BS. The CIC helps BS to remove a few segments of the requested file from the superimposed signal, which are available in the cache. By identifying the common portions between the received signal and cache contents, the receiver obtains the knowledge of the assigned power to the information associated with those segments. The CIC eliminates the common portions from the superposed signal and reduces the interference power [85], [144]. The CIC is a unique feature of the cache-aided NOMA scheme and does not exist in the conventional NOMA.

To understand the CIC, consider a D2D caching system with two cache-enabled users U_i and U_j request for files W_A and W_B respectively. This analysis can be extended for more than two users. The split caching technique is employed, where a file is divided into three segments and placed fully or partially into the cache sequentially to minimize the initial delivery delay. The segments are denoted by W_{fn}, $f \in \{A, B\}$, $n \in \{0, 1, 2\}$. The users cached $c_{kf} \in [0, 1]$ portions of W_f, $f \in \{A, B\}$. The minimum and maximum portions of W_f are defined as $\underline{c}_f = min_{k\in\{i,j\}} c_{kf}$ and $\overline{c}_f = max_{k\in\{i,j\}} c_{kf}$ respectively. The corresponding cache status $\underline{k}_f$ = arg $min_{k\in\{i,j\}} c_{kf}$ and $\overline{k}_f$ = arg $max_{k\in\{i,j\}} c_{kf}$.

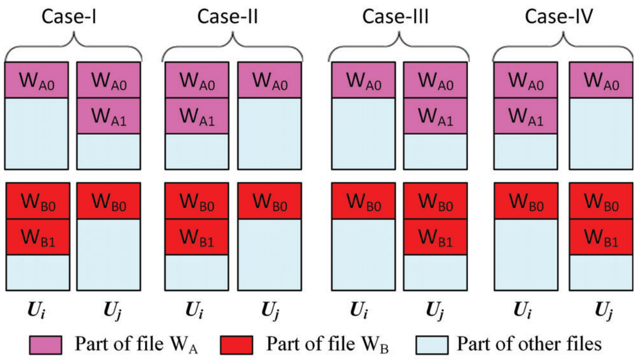


Fig. 4: Different cache status of U_i and U_j requested file W_{fn}, $f \in \{A, B\}$, $n \in \{0, 1, 2\}$. For all four cases, the file portion W_{f0} and W_{f2} are cached and uncached, respectively. Thus, the transmitter never transmits W_{f0} but always needs to transmit W_{f2} [85].

Four possible cache configurations at the time of the request are shown in Fig.4. Case-I is the unfavorable condition for both users, Case-II is favorable for U_i but unfavorable for U_j, Case-III is unfavorable for U_i but favorable for U_j, and Case-IV is favorable for both users. All these four scenarios can be expressed as follows.

Case-I: $i = \overline{k}_B$ and $j = \overline{k}_A$, i.e., $i = \underline{k}_A$ and $j = \underline{k}_B$;
Case-II: $i = \overline{k}_B$ and $j = \underline{k}_A$, i.e., $i = \overline{k}_A$ and $j = \underline{k}_B$;
Case-III: $i = \underline{k}_B$ and $j = \overline{k}_A$, i.e., $i = \underline{k}_A$ and $j = \overline{k}_B$;
Case-IV: $i = \underline{k}_B$ and $j = \underline{k}_A$, i.e., $i = \overline{k}_A$ and $j = \overline{k}_B$;

Due to the limited storage capacity of the cache, placing the whole file W_A or W_B in the cache memory might be impractical. Since the content placement is executed without

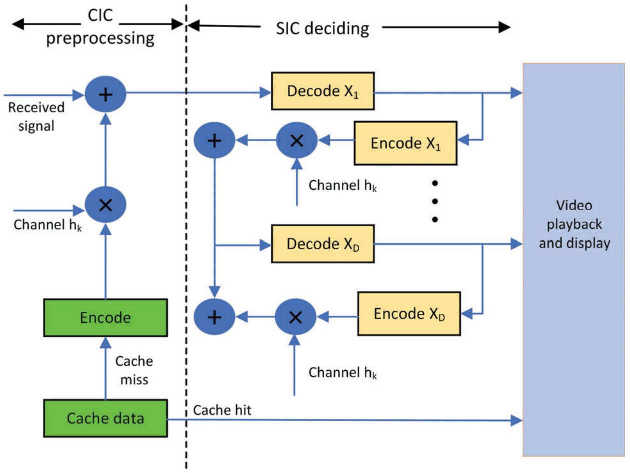


Fig. 5: Joint CIC and SIC decoding techniques. The cached portions are removed from the received signal by the CIC processing. The residual signals $X_1, X_2, ..., X_D$ which are not cached but sequentially decoded by applying traditional SIC process [85].

knowing the actual content demands, caching gain is reduced by full file caching. Therefore, it is assumed that the maximum W_{f0} and W_{f1}, $f \in \{A, B\}$ can be cached. The BS knows the perfect cache statuses, thus, transmits only the un-cached subfiles.The BS transmits x, a superimposed NOMA signal given by

$$x = \begin{cases} \sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} & \\ \qquad\qquad\qquad\quad +\sqrt{p_{j,2}}x_{B2}, & \text{Case-I}, \\ \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} + \sqrt{p_{j,2}}x_{B2}, & \text{Case-II}, \\ \sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}, & \text{Case-III}, \\ \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}, & \text{Case-IV}, \end{cases} \tag{1}$$

where x_{fs} is the codeword corresponding to un-cached subfile W_{fn}, $f \in \{A, B\}$, $s \in \{1, 2\}$, and p_{kn}, $k \in \{i, j\}$, $s \in \{1, 2\}$ is the transmit power to x_{fs}. The joint CIC and SIC receiver, shown in Fig. 5 performs the CIC pre-processing before applying the SIC-based decoding. The CIC process exploits the cache of U_i to remove X_{B1} and X_{A1} form Case-I and Case-II respectively of (1), and we get residual signal (2). Similarly, the CIC process exploits the corresponding cache of U_j to get (3) from (1).

$$y_i^{CIC} = \begin{cases} h_i\left(\sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}\right) & \\ \qquad\qquad\qquad +z_i, \ \text{Case-I \& III}, & \\ h_i\left(\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}\right) & \\ \qquad\qquad\qquad +z_i, \ \text{Case-II \& IV}. & \end{cases} \tag{2}$$

$$y_j^{CIC} = \begin{cases} h_j\left(\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} + \sqrt{p_{j,2}}x_{B2}\right) & \\ \qquad\qquad\qquad +z_j, \ \text{Case-I \& II}, & \\ h_j\left(\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}\right) & \\ \qquad\qquad\qquad +z_j, \ \text{Case-III \& IV}. & \end{cases} \tag{3}$$

The CIC process discards the file segments (all/a few) belonging to other users if those are available in the associated cache. After the CIC operation, the receiver applies the conventional SIC process to decode its signal. It is noted that before applying the SIC process, a receiver can remove a few amounts of interference from the received superimposed signal by using the CIC process. The caching technique helps SIC decoding even when the requested file is not cached. The combination of caching and NOMA can only provide this advantage. The joint CIC and SIC decoding process significantly increases the sum-rate and reduces the file delivery times [85], [144]. In [145], the authors have proposed a new D2D cache-aided NOMA system, where the cache infrastructure is available at both the users' end and the BS, and employing CIC enhances the system sum-rate. The caching technique also is implemented in MIMO-based wireless networks for canceling interference [146].

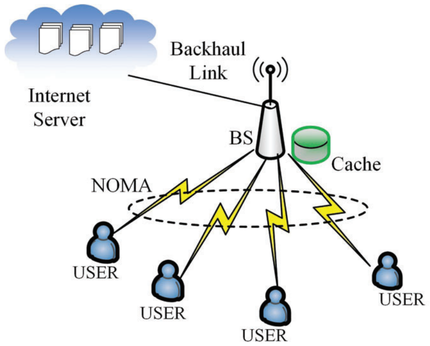


Fig. 6: Basic cache-aided NOMA network, where BS serves multiple users simultaneously other than time and frequency domain, and cache reduces use of backhaul link.

2) Probability of Successful Decoding: This section discusses how full file edge caching helps NOMA in the SIC-based decoding process. Unlike split file caching, BS stores complete files in the cache for full file caching. In NOMA, receivers applies SIC process to subtract a large number of signals intended for other users from the superposition signal before decoding its signal. A decoding failure occurs when a user fails not only to decode its signal but also anyone of the *other signals*. In a cache-aided NOMA system, a user needs to decode comparably a smaller number of *other signals* as contents of some users may be available in the cache. Consequently, a cache-aided NOMA attains an enormous improvement in SDP. To understand the decoding process, consider a simple edge cache-aided NOMA system with K users, $\mathcal{K} \in \{1, 2, ..., K\}$ and a cache-enabled BS, as shown in Fig.6. The channel coefficient of BS-to-kth user is h_k. Let u_1 and u_2 request for files f_1 and f_2 respectively to the BS. Now, four possible scenarios of cache status are:

Scenario-1: Both the f_1 and f_2 are cached
Scenario-2: f_1 is cached but f_2 is not cached
Scenario-3: f_1 is not cached but f_2 is cached
Scenario-4: Both the f_1 and f_2 are not cached

According to NOMA principles, the transmitter allocates the maximum power to the user having the poorest channel gain (weakest user) and minimum power to the user with the strongest channel gain (strongest user). The weakest user decodes its signal directly considering other signals as interference. The stronger users apply the SIC process during decoding their signals. Our aim is not to derive the SDP of a cache-aided NOMA network rather to illustrate the impact of caching on the decoding process. Hence, for simplicity, we assumed $|h_1|^2 < |h_2|^2$, i.e., u_1 is weaker than the u_2 therefore, BS allocates higher power to u_1 than the u_2.

Scenario-1: In this case, both f_1 and f_2 are available in the cache. Let $\mathcal{S}_1^{(1)}$ be the SINR of the signal for u_1 at u_1. Being a stronger user, u_2 first decodes the signal of u_1 and then decodes its signal using SIC. Consider, $\mathcal{S}_1^{(2)}$ and $\mathcal{S}_2^{(2)}$ are the SINRs of the u_1's signal and u_2's signal, respectively, at u_2. Now, $\mathcal{D}^{(1)}$, the overall SDP of the system for scenario-1 can be given as

$$\mathcal{D}^{(1)} = \mathcal{P}_r\bigg(\mathcal{S}_1^{(1)} \geq \gamma\bigg)\mathcal{P}_r\bigg(min\big(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}\big) \geq \gamma\bigg), \quad (4)$$

where γ is the predefined threshold SNR required to decode signal successfully. The first term $\mathcal{P}_r(\mathcal{S}_1^{(1)} \geq \gamma)$ and the second term $\mathcal{P}_r(min(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}) \geq \gamma)$ of (4) are the SDP of f_1 and f_2 respectively.

Scenario-2: In this scenario, f_2 is not cached. BS needs to access the internet server for f_2 through backhaul link. Let $\mathcal{S}_{BS}^{(2)}$ is the SNR of the u_2's signal at BS. Once the BS receives f_2 from internet server it delivers both the f_1 and f_2 using NOMA. Now, $\mathcal{D}^{(2)}$, the overall SDP for the scenario-2 can be expressed as

$$\mathcal{D}^{(2)} = \mathcal{P}_r\bigg(\mathcal{S}_1^{(1)} \geq \gamma\bigg)\mathcal{P}_r\bigg(min\big(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}, \mathcal{S}_{BS}^{(2)}\big) \geq \gamma\bigg). \quad (5)$$

Scenario-3: In this scenario, f_1 is not cached. Similar as scenario-2, after receiving f_1 from internet server, BS delivers both the files to the users. The SNR of the u_1's signal at BS is $\mathcal{S}_{BS}^{(1)}$. $\mathcal{D}^{(3)}$, the overall SDP for the scenario-3 is given by

$$\mathcal{D}^{(3)} = \mathcal{P}_r\bigg(min\big(\mathcal{S}_1^{(1)}, \mathcal{S}_{BS}^{(1)}\big) \geq \gamma\bigg)\mathcal{P}_r\bigg(min\big(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}\big) \geq \gamma\bigg). \quad (6)$$

Scenario-4: This is the case when none of the file is cached. BS downloads both the f_1 and f_2 from internet server using NOMA. It is assumed that internet server allocates more power to the f_1. After receiving both the files f_1 and f_2 from internet server, BS decodes the files and then delivers both the files to corresponding the users using NOMA. Now, $\mathcal{D}^{(4)}$, the overall SDP is expressed as

$$\begin{aligned}\mathcal{D}^{(4)} = \mathcal{P}_r\bigg(&min\big(\mathcal{S}_1^{(1)}, \mathcal{S}_{BS}^{(1)}\big) \geq \gamma\bigg)\\ &\mathcal{P}_r\bigg(min\big(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}, \mathcal{S}_{BS}^{(1)}, \mathcal{S}_{BS}^{(2)}\big) \geq \gamma\bigg).\end{aligned} \quad (7)$$

The first term and second term of (4)-(7), are the SDP of f_1 and f_2 respectively. It is worth noting that without caching (Scenario-4), multiple users need to decode a sequence of signals intended for other users before obtaining their signals. On the contrary, users can discard signals of others users using the cached content, which improves the SDP.

3) Cache Placement Using NOMA: The 5G and B5G need to share the spectrum among UEs to improve the spectral efficiency and simultaneously ensure massive connectivity with a high reliability. The cache placement during off-peak hours is not an efficient approach for 5G. In comparison with the OMA technique, the NOMA can place more content in the cache and serve a large number of users during the content delivery phase within a short duration of time. The OMA can store (or push) only a single file during a single time slot. Therefore, the BS pushes only the content with a maximum popularity during the first time slot and the second most popular file during the second time slot, and so on. When a comparatively longer period is available, the OMA-based content placement requires sophisticated methods which efficiently schedule the files based on their popularity [136]. Unlike the OMA, during a single time slot, applying the NOMA principle, the BS can push multiple files based on their popularity at the same time. The content delivery phase is divided into small time slots. During a single window, OMA can serve a single user whose requested file is available in the cache. An efficient user scheduling algorithm based on *first-in-first serve* is required for serving multiple users' requests. On the other hand, like the content pushing phase, the NOMA serves multiple users simultaneously. The cache-aided NOMA scheme efficiently improves the cache hit probability and reduces the delivery outage probability compared to conventional OMA-based caching.

Summary: Various studies validated that under the optimal scenario for both NOMA and caching, NOMA always outperforms OMA [123]. The performance of NOMA-based systems crucially depends on the degree of accuracy of the SIC process. NOMA with caching can remove interference fully or partially using cached contents and increase the SDP. Cache and NOMA individually help each other. The NOMA makes a faster cache placement process and delivers requested files simultaneously to multiple users maintaining fairness. On the other hand, caching helps NOMA in the SIC decoding process even when requested files are not cached. The conventional cache strategies push contents during off-peak hours, which is not an efficient approach, specifically when the popularity of contents changes suddenly and networks need to update cache contents. NOMA-aided caching can push multiple contents within a short duration and becomes the best candidate for content pushing during peak time [136]. Push-then-deliver and the push-and-deliver strategies are capable of content placement during on-peak hours [136]. A combination of NOMA with caching could be a breakthrough strategy for next-generation communication systems.

IV. Key Performance Indicators

This section presents a fundamental analysis of key performance indicators (KPIs) of cache-aided NOMA-based 5G and B5G systems. Various Performance metrics and the approaches adopted to enhance them have been summarised in Table-V.

A. Sum-Rate Maximization

Sum-Rate measures the successful data transmission rate over the communication channel of the unit bandwidth.

TABLE V: Different performance metrics and their solution techniques

Ref.	System	Metrics	Implemented Technique	Research Gap/ Merit
[145]	D2D	sum-rate	CIC process	Assumed perfect knowledge of CSI and error free SIC process
[147]	D2D	sum-rate	Subchannel allocation	Considered equal power allocation, contents are of equal size
[148]	ECT	sum-rate	Non-convex optimization problem	Imperfect CSI is taking into account
[85]	D2D	Sum-rate & delay minimizing	Optimal decoding order and optimal transmit power	Proposed delivery scheme is applicable for any caching scheme
[86]	ECT	Delay minimizing	Recommendation Mechanism	Perfect CSI is known at BS
[126]	ECT	Delay minimizing	Deep neural network-based dynamic power control	Assumed guaranteed successful decoding at receivers
[149]	ETC	Delay minimizing	Resource allocation approach	Considered a typical static single user scenario
[150]	ETC	Delay minimizing	Formulated optimization problem	Considered an equal file size
[129]	D2D	SDP	Optimal power allocation	Considered two-user scenario with perfect CSI
[151], [152]	D2D	SDP	Power allocation optimization	Considered a perfect SIC technique
[153]	D2D	SDP	Deep learning-based power allocation	Considered scheduling delay of content requested by users
[141]	ECT	Outage and cache hit probability	Dynamic power allocation strategy	BS cannot serve users during the periodic cache placement process
[154]	D2D	Outage performance	Inverse Laplace transform used for exact outage probabilities	Evaluated expression for exact outage probabilities
[155]	D2D	Outage performance	Power allocation and user pairing scheme	Hybrid delivery scheme can select NOMA or coded multicasting based on the channel conditions.

Achieving a higher sum-rate is a primary requirement for any communication system especially when a video file is streaming. The information-theoretic studies have demonstrated that NOMA cannot elevate the overall sum capacity of the system compared to conventional OMA [39], [156]. Hence, NOMA is exploited to maintain user fairness [12], [45], [130], [157]. Ding *et al.* demonstrated that fixed power allocation based NOMA system achieves remarkable throughput gain only for asymmetric channel gain quality of the users, and for symmetric channel gains performance of NOMA and OMA are identical [104]. Wireless caching is an efficient approach incorporated with the NOMA to increase the sum-tare of 5G and B5G networks [158]. The cache-aided NOMA significantly increases the achievable sum-rate compared to OMA by enabling joint CIC and SIC [85]. In [147], authors validated that cache-aided cloud radio access networks can achieve an improved sum-rate when NOMA is incorporated. The authors proposed a cache-aided NOMA-based D2D system, where a pair of users utilize the uplink channels for delivering cached contents [145]. The performance of the proposed network paradigm was evaluated in terms of the sum-rate. Xinyue *et al.* [148] formulated a sum-rate maximization problem under the constraints of the peak allocated power, backhaul capacity, minimum unicast rate, and maximum multicast outage probability for evaluating the performance of a cache-aided NOMA-based multiple-input single-output system.

B. Delay Reduction

The delay in delivering requested files depends on various network resources like transmit power, available bandwidth, cache status, etc. Overall transmission delay of a cache-aided NOMA system depends on the delay associated with both the backhaul and BS-to-user links. Under a cache hit condition, BS delivers requested files from the cache and reduces the delivery delay by saving the time required to fetch the requested file from the data center using the backhaul link. It is another advantage of cache when incorporated into a NOMA technique. To analyse the delay reduction technique, consider a cache-aided NOMA network, as shown in Fig.6, where *k*th user requests for *i*th $\forall i \in \mathcal{I} = \{1, 2, ..., I\}$ file of size L_i, $i \in \mathcal{I}$ to the BS. Transmission over the backhaul link is subjected to the availability of the requested file in the cache. The cache status of the *i*th content is symbolized as $C_i \in \{0, 1\}$. Particularly, $C_i = 1$ if the requested content is cached and 0 otherwise. Assuming R_B as the transmission rate of the backhaul link, $\mathcal{T}_B$, transmission time required for *i*th file is given by

$$\mathcal{T}_B = (1 - C_i)\frac{L_i}{R_B}. \quad (8)$$

The cache-aided NOMA system reduces the delivery delay by $\mathcal{T}_B = L_i/R_B$ when requested file is cached. The dynamic power allocation plays a vital role in reducing the delivery delay of cache-aided NOMA systems, where the volume of data is different [126]. In [126], for a cache-aided NOMA system, authors minimize the data transmission delay of each user under the constraints of the total available power and maximum tolerable transmission delay. Liu *et al.* performed joint tasks scheduling and resource management to minimize the transmission latency subjected to maximum transmission delay and network cost for a cache-enabled ultra-dense network [149]. The delivery delay of a cache-aided NOMA is minimized by jointly optimizing the decoding order of NOMA and power and rate allocations [144]. In [159], the authors have jointly optimized the transmission strategies for both backhaul and BS-to-users links under the constraint of transmission power to minimize the delivery time of an edge caching-enabled NOMA system. Recently, in [86], the average transmission delay of a cache-aided NOMA network with two-user is derived considering a scenario when both users request files of the same size. The average delay under the constraint of the minimum quality requirement of the file is minimized considering a *recommendation* mechanism [86].

C. Energy Consumption Minimization

Reducing the energy consumption is crucial for communication systems specifically battery operated systems. Various energy consumption minimization approaches adopted in cache-aided NOMA networks. The authors reduce the average signal transmitted power of a cache-aided NOMA-based cellular network under the constraint of cache memory capacity [137]. In [94], jointly optimizing the task offloading, computation and cache resource allocation under the constraint of caching and computing resources, authors have minimized the total energy consumption for the proposed cache-aided MEC network. In [160], authors minimized the total required transmitted power subjected to a minimum data rate of the users. Increasing energy efficiency is one of the challenging issues of Unmanned Aerial Vehicle (UAV)-assisted wireless networks. The energy efficiency of a UAV-assisted wireless NOMA system is maximized under the constraint of subchannel assignment and power allocation to cache-enabled UAVs [161]. In [150], authors propose a two-sided matching and swapping algorithm for maximizing the energy efficiency of a UAV-assisted NOMA-based fog wireless network. Authors are aiming to reduce the total consumption of energy by the UAV-assisted NOMA-based MEC networks taking into account the task computation allocation, computation capacity, and UAV trajectory in [162]. The authors in [163] studied resource allocation for enhancing the energy efficiency by optimizing subchannel allocation and power allocation in a NOMA hierarchical network.

D. Successful decoding probability

Successful decoding probability (SDP) defines the probability that a receiver decodes own signal correctly. Cache-aided NOMA systems achieves an improved SDP than NOMA and cache-aided OMA systems [129], [152]. The optimal power distribution is the most popular approach for enriching the SDP. The authors formulated optimal power distribution problems over the Rayleigh fading channel [129], Weibull, Nakagami-m, and Rician downlink fading channels in a cache-aided cellular network [152]. Depending on the QoS, Yin *et al.* proposed a dynamic power allocation strategy for a cache-aided cellular system that increases the probability of successfully decoding compared with the OMA and fixed power allocation-based NOMA schemes [141]. Doan *et al.* in [153] propose divide-and-conquer-based and deep-learning-based power allocation methods to maximize the SDP that also ensures QoS and fairness of the users.

Apart from the above-mentioned parameters, cache hit probability, backhaul cost, outage probability, network delay, spectral efficiency, and energy efficiency are KPIs for designing cache-aided NOMA systems.

Outage Probability- Outage probability and SDP are the two sides of the same coin. Outage probability tells the probability that a receiver fails to achieve QoS above the threshold level. On the other hand, SDP conveys the probability that a receiver successfully decodes its signal, i.e., QoS is above the threshold level. Outage probability also a widely used to evaluate the performance of cache-aided systems. Dani et al. evaluate the performance of the proposed hybrid NOMA or coded multicasting-based delivery scheme in terms of outage probability [61], [155]. These articles validate the superiority of NOMA over coded multicasting when channel gains of the paired users are highly distinctive, and coded multicasting is better than NOMA under similar channel gains. The outage performance of a MIMO NOMA cellular network is analyzed using inverse Laplace transform in [154].

Cache Hit Probability- The hit rate is an important parameter that tells how professionally popular files are selected to place in the cache. It is a ratio of the number of times the cache successfully delivers requested files and the total number of times users request files. The hit rate primarily evaluates the efficiency of cache placement techniques. A successful cache hit reduces communication costs by avoiding the use of the backhaul link and also reduces outage probability significantly [141]. The performance gain margin of cache-aided NOMA systems over the conventional NOMA systems crucially depends on the higher cache hit probability.

Backhaul Cost- BS communicates with other BSs or internet servers via a backhaul link. Generally, expensive optical fibers are used as a backhaul link to achieve high-speed data transfer [24]. Inefficient utilization of backhaul leads to an increased overall expenditure of wireless communication systems. The cache technique efficiently reduces the Backhaul cost. The computation time, successful cache hit, and content delivery latency play significant roles in increasing the energy efficiency of cache-aided wireless networks.

Summary: This section discussed some vital performance metrics of 5G and B5G systems. Table V shows interesting insights into caching techniques applied to evaluate the performance of cache-aided NOMA systems. The D2D caching techniques are widely implemented to maximize the SDP and minimize the outage probability. On the other hand, the edge caching has been implemented to reduce the delivery delay. The performance improvement of cache-aided NOMA systems in terms of any metric depends on the higher cache hit probability.

V. Verticals and Use Cases

This section focuses on various application scenarios of cache-aided NOMA networks and the challenges associated with these applications along with their solutions. The application scenarios of cache-aided NOMA are presented in Fig.7.

A. Non-Terrestrial Networks

Non-terrestrial networks (NTN) have evolved a lot over the last decade, beyond simple drones, and they now encompass a whole eco-system like hierarchy. The drones or unmanned aerial vehicles (UAVs) are at the lowest layer of the atmosphere superseded by cubesats in the second layer at the edge of atmosphere, the low-earth-orbit (LEO) satellites

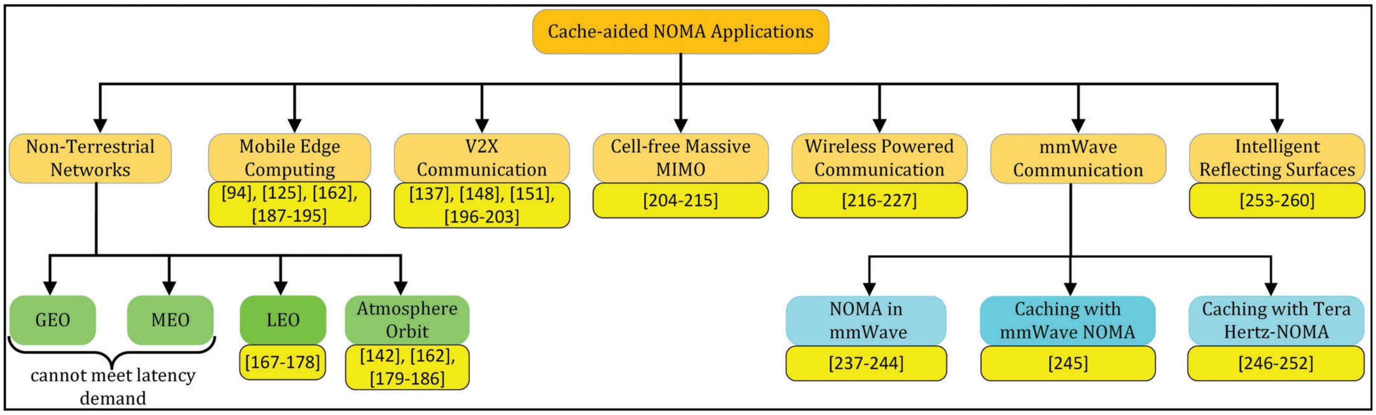


Fig. 7: Application areas of the cache-aided NOMA technique.

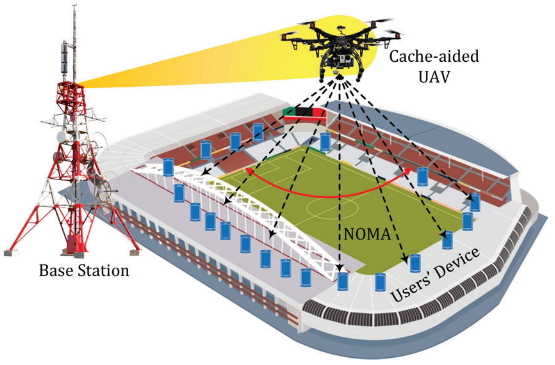


Fig. 8: One of the popular application scenarios of cache-aided NOMA-based UAV-assisted communication systems. The UAV was deployed temporarily to assist the BS to meet thousands of users' requests [142]

form the third layer beyond the atmosphere, and finally, the geostationary-earth-orbit (GEO) satellites are in the topmost layer in deep space. The backhaul link of terrestrial communication networks is connected with the data center via an optical fibers link which is prone to damage during disasters. It is challenging to deploy terrestrial infrastructures in remote locations like mountains, seas, and deserts. Moreover, the speed of light in optical fiber is 30-40% slower than that of free space, which has motivated researchers for non-terrestrial communications. A few companies like Google, Facebook have already launched satellites for providing better QoS with low latency to their customers. The non-terrestrial network can cover a large area on Earth for an instant, a LEO satellite approximately covers 1 million km^2 area. Various airborne platforms such as Balloons [164], Helikites [165], and UAVs [166] are recently emerging as potential approaches to meet wireless traffic demands, specifically for mobile users. Based on the orbital altitude, the non-terrestrial networks are classified into four categories, (i) GEO (35786 km), (ii) Medium Earth Orbit (MEO) (2000 – 35000 km), (iii) LEO (160 – 2000 km), and (iv) Atmosphere Orbit (a few hundred meters).

A signal requires a much higher round trip time between ground station and satellites deployed in MEO (or above), thus cannot meet the latency demand for 5G communications. In LEO satellite-based communications (LEOComm), as the satellite is deployed in relatively lower altitude orbits, the round trip time of a signal is only a few ms (10-15ms for the SpaceX Starlink system) [167] and can meet the latency requirement of Internet of Things (IoT), smart grid, and vehicular communication applications [168]. The caching facility in the satellite networks improves the latency performance and makes LEOComm a possible candidate for the 5G paradigms. Armon *et al.* have designed and addressed operating issues of cache-aided satellite distribution systems for web caching [169], [170]. To minimize the downlink and uplink traffic load, Wu *et al.* have proposed a two-layer caching model for satellite-terrestrial networks, where caching in the ground stations constitutes the first layer and caching in the satellite constitutes the second layer [171], [172]. The authors have validated that two content caching schemes, named as *most popular content-based* and *uniform content-based* schemes can efficiently improve the spectral efficiency in the Hybrid satellite-terrestrial relay networks [173]. Google Loon project is one of the industry projects where Google have installed Internet-delivery drone for providing global massive connectivity [174]. CubeSats are a class of miniaturized satellite for research build up by multiple cubic modules of dimensions 10 cm × 10 cm × 10 cm deployed into the lower altitude of LEO. It is reported that 1634 CubeSats already launched by Aug. 2021, and the future of nanosatellites is still to come [175]. Researchers are actively investigating the applicability of cache-aided NOMA in hybrid satellite-terrestrial (HST) networks to improve spectral efficiency, outage probability, hit probability and transmission latency [176]–[178]. In [176], Zhang *et al.* analyzed the transmission delay and cache hit probability of cache-aided NOMA HST systems, where users receive their requested contents from a cache-enabled relay node. In addition, they also investigated the outage probability and hit probability of a cache-aided NOMA HST network [177]. Recently, Vibhum *et. al* have incorporated a cache-aided NOMA in overlay-based cognitive hybrid satellite-terrestrial networks, where a secondary transmitter with cache capability is employed for cooperative relaying following the NOMA protocol [178]. Unlike [176] [177], users receive signals from

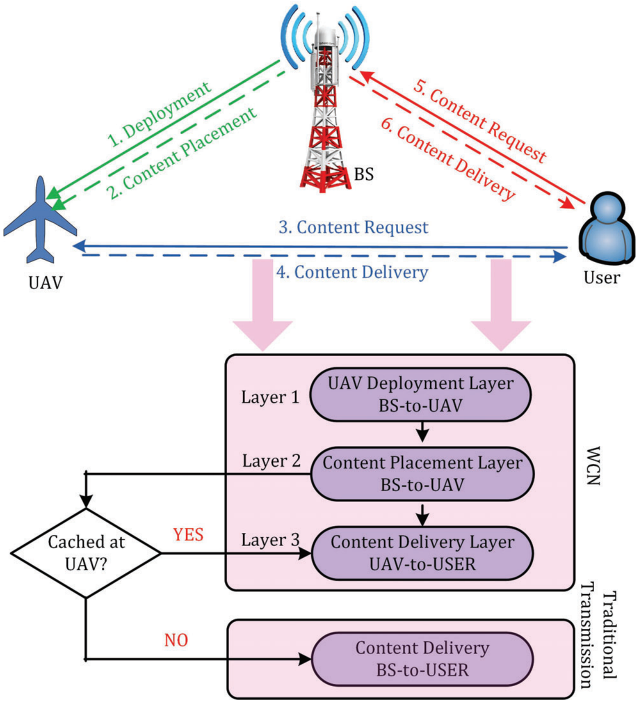


Fig. 9: Cache content delivery strategy and different layers of a UAV-assisted wireless caching system. User requests for the desired file to the BS only when the file is not cached [181].

the relay and the satellite directly in [178]. Herein, authors validated the superiority of the cache-aided NOMA over the cache-free NOMA model in terms of the outage probability. The cache-aided NOMA technique has not been explored much in satellite networks and could be a promising research domain.

UAVs deployed in the atmospheric orbit are the most popular and efficient commercial approaches to provide short-term connectivity in a hot-spot area. UAVs-aided wireless communication is gaining attention among researchers from both the industrial and academic communities for its low infrastructural cost, reduced size, line-of-sight communications, and flexible deployment process. Though UAV was developed for military applications but presently is being utilized for commercial applications also. To fulfill the rising demand for high data transmission rate with low latency, UAV exploited as an effective approach for highly dense wireless communication networks [179], [180]. The wireless systems deploy UAVs at low-altitude as a flying BS to meet traffic demands temporarily of a hot-spots area. One of the popular commercial application areas of UAV with cache-aided NOMA technology is depicted in Fig. 8, where a large number of mobile users in a hot-spot area are under the coverage of a ground macro base station (MBS). The MSB is overloaded and unable to satisfy the users' requirements during peak hours because of the limited available frequency band. Cache-enable UAVs are deployed to assist the MBS in delivering users' requested files. The battery-operated UAVs and the MBS are connected through wireless channels. When the battery is exhausted, it is recharged, or the UAV is replaced by a new one.

The UAV enabled cache-aided NOMA network is divided into three layers, (i) UAV deployment layer, (ii) content placement layer, and (iii) content delivery layer [181], as shown in Fig. 9. In the content placement layer, the MBS downloads popular content using the backhaul link and stores it in the cache of UAV using NOMA. The caching contents are replaced/updated regularly. In the content delivery layer, UAV groups users to deliver the requested contents using NOMA based on the CSI. A statistic QoS-based fixed (SQF) and instantaneous QoS-based adaptive (IQA) power distribution methods are applied in a UAV-enabled cache-aided NOMA system to improve the outage probability performance. Furthermore, an improved power allocation strategy named cross-layer based optimal method is employed to maximize the system hit probability [181]. A deep reinforcement learning (DRL) algorithm is proposed for content placement and delivering in a cache-enabling UAV-assisted cellular network [182]. The cache-enabled UAV serves users directly on the availability of the requested file in the cache. Otherwise, user requests for the files to the MBS directly [181] or via UAV [183].

The resource allocation in UAV with cache-aided NOMA system has been studied in [142], [183]–[185]. In [183], resource allocation for a UAV-assisted cellular system has been considered for maximizing the quality of experience (QoE) of the users by optimizing the content placement in the cache, location of UAV, and user association. In [184], to minimize the delivery delay, the authors have modelled an optimization problem for UAV deployment, caching placement, and power allocation of NOMA as a Stackelberg game. However, to minimize the content delivery delay, the authors in [142] have incorporated the Markov decision process for jointly optimising the content placement, user scheduling, and power allocation to NOMA users. Increasing the operation time of battery-operated UAVs is one of the challenging issues. The energy spent for cache content placement and replacement further reduces flying time. To prolong the operation time of UAVs, the authors in [185] have deployed a UAV that can harvest solar energy from the environment. Various algorithms researches have been applied to solve the optimization problem of UAV systems which have hardly considered the dynamic networks environment including the movement of UAV. The authors have applied a Markov decision process (MDP) to model caching placement and resource allocation with dynamic UAV locations and content requests [184] [142].

In [183] and [142], authors have proposed a UAV-assisted framework for delivering multimedia contents to the users located in a hotspot area. Here, cache-aided mobile UAV operated as a BS that reduces the backhaul link traffic providing the cached contents to the users' group by NOMA. In [186], Haibo *et al.* have developed a cache-aided UAV-assisted vehicle-to-network (V2N) communication system where the UAV operates as a flying base station to communicate with vehicles. Cache-aided UAV was deployed to maximize the sum fairness of the vehicles. In [182], cache-enabled UAV was deployed in a cellular network to assist the delivery of the user requested multimedia contents. Cache-enabled UAV was deployed in a NOMA-based MEC network to minimize

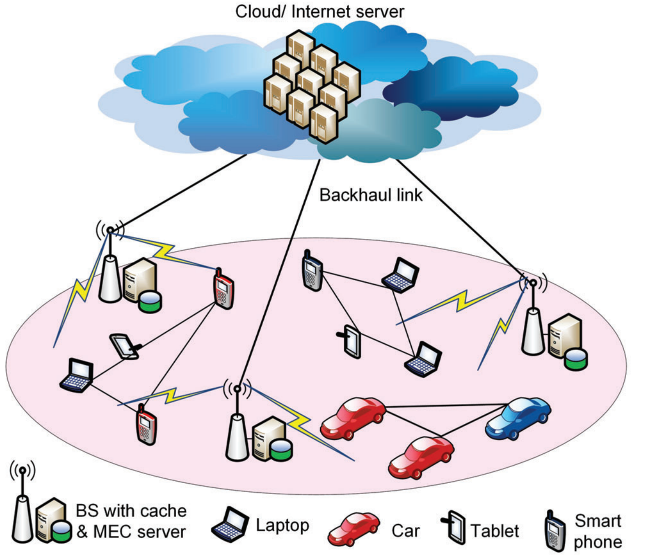


Fig. 10: Application scenario of cache-aided NOMA MEC networks. Users are connected and cooperatively sharing files. Devices offload their computational task to the MEC-cache-aided BS.

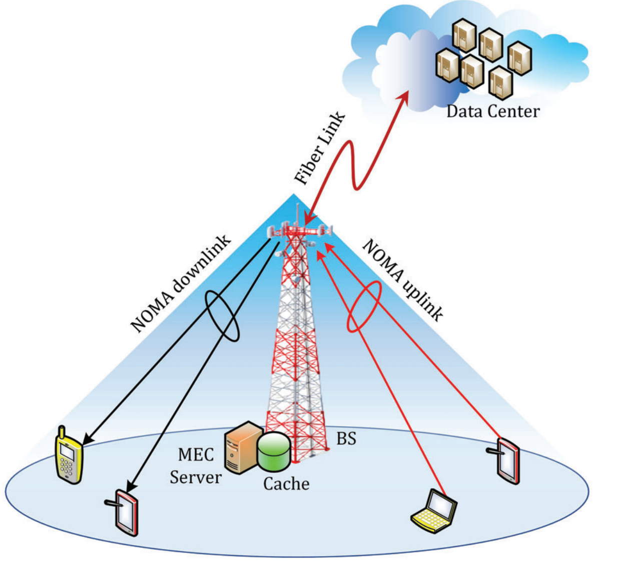


Fig. 11: Basic architecture of a cache-aided NOMA MEC network. User's devices offload computational tasks to cache and MEC-aided BS [188].

the consumption of total energy in [162].

B. Mobile Edge Computing

Mobile edge computing (MEC) improves the cloud computing capability by shifting computing facilities at the edge of highly latency-sensitive networks such as cloud gaming and multiplayer gaming, autonomous vehicle functions, real-time drone detection, etc. In addition, caching in the MEC server further enhances the quality of communications and reduces backhaul load. Offloading the computation workloads of the mobile users, MEC assists the existing applications to improve their performance in terms of congestion in networks, delivery latency, and QoE. Cache facilitated MEC significantly intensifies the performance further [187]. Generally, the nearest APs to the users are equipped with cache-enabled MEC servers. The users within the coverage area of an AP get access to the caching contents that significantly reduce the backhaul link traffic and data transmission rate. The NOMA strategy empowers MEC to cope with the massive connectivity and huge data traffic of mobile users. NOMA-based MEC (NOMA-MEC) networks are capable of offering flexible computing services to mobile users. Some of the applications of cache-aided NOMA-MEC networks are depicted in Fig.10. The task caching in the MEC refers to storing some of the popular completed tasks and their associated data in the cache. Unlike the other cache-aided applications, in MEC, task caching requires computation in addition to storage. The user requested task can be computed locally in the mobile device or offloaded to the MEC server.

Local Computation:- Consider a simple edge caching NOMA-MEC network with a BS, N mobile users and a remote cloud, as shown in Fig. 11. The BS is equipped with a MEC server with a finite storage capability. Let L_n is the size of the requested content, f_n^l is the local computing capability, W_n is the number of cycles required to finish the given task, R_n^{BH} and R_n^{DL} are the average data transmission rate through backhaul link and downlink NOMA, respectively. $\mathcal{T}_n^{Lo}$, the total computational latency for executing a task locally includes backhaul latency ($\mathcal{T}_n^{BL} = L_n/R_n^{BH}$), downlink latency ($\mathcal{T}^{DL} = L_n/R_n^{DL}$) and local processing time ($\mathcal{T}_n^{LP} = W_n/f_n^l$) which is given by

$$\mathcal{T}_n^{Lo} = (1 - C_n)\mathcal{T}_n^{BH} + \mathcal{T}_n^{DL} + \mathcal{T}_n^{LP}, \tag{9}$$

where C_n is the status of the cache content, $C_n = 1$ if the requested content is available in the BS cache, else it is 0. *Edge Offloading:-*The primary objective of the MEC is to offload the computational task to the BS as much as possible for the remote execution. The total computational latency for edge offloading includes the uplink transmission time, processing time of the MEC server, and the backhaul delay. If a U_n is the data of the offloaded task to the BS by the *n*th user and the R_n^{UL} is the average NOMA uplink data rate, the uplink delay is given as $\mathcal{T}_n^{UL} = U_n/R_n^{UL}$. The edge processing time for the offloaded task is $\mathcal{T}_n^{EP} = W_n/f_n$, where f_n is the resources allocated by the MEC server for executing the computational task. Now, the total edge offloading latency is expressed as

$$\mathcal{T}_n^{EO} = (1 - C_n)\mathcal{T}_n^{BH} + \mathcal{T}_n^{UL} + \mathcal{T}_n^{EP}. \tag{10}$$

Since the size of the result of the offloaded task is much smaller than that of the offloaded task, and the downlink data rate is much higher than that of the uplink (as the transmitting power of the BS is higher than that of the users' device), the downlink time delay and the energy consumption associated with the downloading the result of the offloaded task can be neglected [189]. Hao *et al.* studied the challenges related to the joint optimization of the task caching and the offloading in [187]. Based on the popularity of the historical tasks, the LSMT algorithm predicts the future task popularity as a function of time. When the popularity of the computational tasks is unknown, the Gated Recurrent Unit (GRU) algorithm can be applied to predict it for a time-varying system [190]. Depending on the predicted popularity of the task, a multi-agent Deep-Q-network (MADQN) algorithm was applied to deal with the problem of caching and offloading. A new collaborative task offloading scheme proposed in [191] is capable of reducing task execution delay up to 42.83% a for single-user caching-enhancement scheme.

Various types of resource allocation methods for cache-aided MEC with NOMA are found in the literature [94], [125], [162], [192], [193]. For efficiently completing the computation tasks of the users, a resource allocation optimization problem under the constraints of caching and computing resources is formulated and addressed by an SAQ-learning-based algorithm in [94]. In [192] also, authors have applied the SAQ-learning-based method to solve the problems associated with the optimization problem for minimizing total energy consumption subjected to offloading decision, computation resource, and caching decision. In [162], the authors designed a MEC network where UAV is deployed as a moving edge cloud server to offload the computation workloads of the mobile terminals. A resource allocation framework for video caching placement and delivery was developed in heterogeneous cache-aided MC-NOMA networks [125]. The delivery-aware cache placement strategy (DACPS) jointly allocates physical and radio resources during the cache placement phase, and the delivery-aware cache refreshment strategy (DACRS) deals with the dynamic behavior of the channel during the delivery phase [125]. Incorporating the overall tasks completion delay and total consumption of computational resources by the edge servers, the authors formulate a system cost function. Thereafter, jointly optimize the computational resource allocations at edge servers and radio resources for smart terminals to minimizes the system cost function [193]. The authors in [194] formulate a new utility function considering offloading time, available resources, and caching decision, and maximize it subjected to the transmission bandwidth, available computing resources, and storage resources. In [188], the authors aimed to reduce the total completion latency for all users of a cache-aided NOMA-MEC, formulated a joint optimization problem of offloading decision, caching strategy, computational resource, and power allocation under the constraint of energy consumption, offloading decision, and computation and storage capacity. In [195] also, the computation delay to finish mobile users' tasks was minimized by jointly optimizing the offloaded workloads and data transmission time.

C. V2X Communication

Vehicular communications have gained a huge attraction among researchers due to the possibility of improving travel experience in terms of road safety, internet access for on-board information, and entertainment facilities. The IEEE 802.11p technology-based communication for vehicular ad hoc networks (VANET) provides 6 - 27 Mbps data rate for a short distance communication [196]. LTE-based vehicle-to-vehicle (V2V) communication supported by the Third-Generation Partnership Project (3GPP) also emerges as an efficient approach [197]. The V2V communications not only provide an entertainment facility to the onboard user but also provide safety, traffic information, pollution control, and traffic applications that require a huge amount of data transmission. In addition, the short duration of connectivity between vehicle-and-infrastructure (V2I), frequent change of channel gain quality, and fast movement of the vehicle make V2V communication further challenging. Liang *et al.* have studied the fundamental challenges to empower efficient vehicular communications from the physical layer perspective in [198]. In [199], the authors verify the superiority of NOMA over conventional OMA in terms of enhancing the content delivery efficiency. Two dynamic cache content placement schemes are proposed for adaptive bitrate streaming of video in vehicular communications [200]. The NOMA technique is well recognized in vehicular communication for the capability to handle massive connectivity and outperforms the traditional OMA-based system [201]. The authors in [202] investigated the spectral efficiency and resource allocation of a NOMA-based vehicular system.

In V2V communication, cache-enabled vehicles store some of the popular contents. Vehicles communicate with the BS during the cache placement phase and on the unavailability of the requested file in the cache of neighboring vehicles. To understand the working principles of cache-aided NOMA in V2V communication, consider a simple vehicular communication model consisting of two cache-aided vehicles (V_1 and V_2) and a BS. Let users of V_1 and V_2 request for files f_1 and f_2 respectively. Consider an extreme case when neither f_1 nor f_2 is cached. The BS downloads the files from the internet using the backhaul link and then transmits files applying NOMA. The traffic load of the backhaul link is the same as of the conventional NOMA. When any one of the files (say, f_1) is cached, one user (V_1) receives its file from a neighbor (V_2), and another user (V_2) gets its file from BS (using backhaul link). Interestingly, for this case, interference can be removed completely, and users do not need to use SIC to decode their signal as the conventional OMA technique emplied to transmit both signals utilizing the whole available bandwidth. Hence, the average performance will be better than that of the conventional NOMA. Another extreme case is when both files f_1 and f_2 are cached. Here interference can also be avoided completely, and BS does not need to use the backhaul link. Hence, the average performance of the cache-aided NOMA in V2V communication will be significantly better than that of the conventional NOMA.

Gurugopinath *et al.* in [151] first proposed a cache-aided NOMA in vehicular communication. The authors consider full file caching and split file caching techniques in vehicular networks. In full file caching, each vehicle stores and requests entire files following NOMA principle; whereas the split file caching technique divides content into two parts. The challenges of cache-aided NOMA in vehicular communications addressed in [203]. A hybrid multicast/unicast scheme has been investigated in cache-aided NOMA-based vehicular networks [148]. In order to maximize the unicast sum-rate, the authors have formulated an optimization problem subjected to peak the transmit power, backhaul capacity, the minimum unicast rate, and the maximum multicast outage probability.

Chao *et al.* in [137] have studied the cache-assisted physical layer security of NOMA-based vehicular communications.

D. Cell-free Massive MIMO

Unlike conventional cellular topologies that mainly serve human users, next-generation communication systems provide mMTC also. The cellular networks cannot handle connectivity to billions of user terminals. Therefore, a cell-free communication topology with decentralized technology is required for next-generation communication, and cell-free massive MIMO (CFmMIMO) technology could be a potential approach [204]. The CFmMIMO comprises large numbers of low costs and low operating powered AP antennas distributed over a wide area and coherently serves user terminals by all nearby AP antennas. A front-haul network connects all the antennas of the APs to central processing units connected with internet servers through a backhaul network. Through this network design, user terminals get AP antenna very close to them, and consequently, achieve improved QoS. Though both the CFmMIMO and distributed massive MIMO can serve thousands of user terminals, they are different. In distributed massive MIMO, BS antennas are installed over a cell and serve only the users within that cell. On the other hand, CFmMIMO is free from a geographical boundary, and antennas serve all users.

Primarily, NOMA was employed in CFmMIMO to reuse pilot sequences within the same cluster which significantly serves more users than the conventional OMA [205]. However, under the low number of active users scenario in a CFmMIMO system, OMA is superior to the NOMA in achieving sum-rate because the NOMA suffers from intra-cluster pilot contamination and imperfect SIC [206]. To deal with this an adaptive NOMA/OMA mode-switching method is proposed in [207], [208]. The phase-related mismatch at the AP degrades the spectral efficiency of a NOMA-based CFmMIMO system. However, NOMA is still capable of outperforming OMA under mismatches and imperfect SIC [209]. The authors have applied Poisson point processes in the NOMA-based cell-free massive MIMO to model the random user and AP locations and found NOMA to be an efficient technique in enhancing the overall rate, especially under low path loss exponents and high AP densities [210], [211].

CFmMIMO is viewed as one of the new technologies for 5G and B5G communications due to its uniform service quality, robust diversity, and interference management ability [204], [212]. Distributed APs of CFmMIMO make it suitable for caching. Integrating with CFmMIMO, caching reduces backhaul traffic load and energy consumption significantly. Recently Chen *et al.* [213] have introduced a cache-aided CFmMIMO framework in 2021. However, in the year 2018, the authors have applied a coded caching in a cell-free environment and derived analytical expressions ergodic spectral efficiency and outage probability expressions for analyzing the performance of SIMO network [214]. Wang *et al.* proposed a smart caching scheme in MEC-enhanced small-cell Massive MIMO networks, where MBS and SBSs are equipped with caching memory [215]. Based on the user's request history, the MEC server can predict the next content that might be requested and starts caching that content in the SBS. Chen *et al.* [213] compared the performance of a CFmMIMO with small cells from caching strategies perspective and established CFmMIMO as a superior technology over small cells in terms of the successful content delivery probability and total energy consumption.

E. Wireless Powered Communication

Simultaneous wireless information and power transfer (SWIPT) through dedicated radio frequency emerged as a superior wireless energy harvesting (EH) technique that prolongs the battery life and provides uninterrupted network operation. Maximizing the energy efficiency is one of the fundamental objectives of 5G networks. Using the time switching (TS) or power splitting (PS) protocol, the SWIPT-enabled system extracts both information and energy from the ambient radio signals simultaneously. Consequently, SWIPT improves the energy efficiency of wireless systems [216]. The combined NOMA-and-SWIPT-based paradigms enhance spectral efficiency and energy efficiency of 5G systems, and support the services of the IoT and the mMTC [217]. Wu *et al.* designed a transceiver for NOMA-based SWIPT-enabled cooperative full-duplex relaying systems [218]. Yuan *et al.* considered a cooperative NOMA transmission scheme in a PS-based SWIPT system and formulated an optimization problem that maximizes energy efficiency and reduces the energy consumption of the system, especially in the low power region [219]. A few articles also found in the literature where the NOMA has been adopted in SWIP system and proposed various methods to analyze different metrics such as outage probability, throughput and energy efficiency [220]–[222].

Caching is popularly applied in sensor networks to reduce energy consumption and improve the energy efficiency of sensors. An AP employed as a gateway to sensors is facilitated with cache memory stores the sensing data temporarily and updates it periodically. The gateway retrieves cached sensing information and delivers it to multiple users without activating the sensors frequently (which consumes substantial energy). In [223], the authors have introduced caching in IoT sensing services and proposed a caching mechanism in EH-enabled sensor networks that improves the sensing performance significantly. The impact of caching and EH on the energy consumption at small cell base stations has been investigated in [224] and shown that instead of existing trade-off between the size of cache and harvesting equipment at the SBS, caching achieves desired system performance. In [225], the authors have proposed a new network paradigm named as GreenDelivery, where based on the popularity and harvested energy EH-enabled small cells cache and push the multimedia contents before it is requested. This network framework considerably reduces macro-BS activities and consequently decreases energy consumption.

In addition to the above articles, researchers have integrated the NOMA technique in cache-enabled EH networks. The caching in the NOMA-enabled SWIPT model functions efficiently in the enhancement of the quality of user experience (QoE) [226]. The authors have proposed a joint content push and transmission scheme in a cache-enabled SWIPT-based relying network [227]. With the help of the NOMA technique, a two-stage content push and the delivery scheme has been proposed to achieve superior spectral efficiency. Another joint content caching and EH method is proposed to improve the performance of EH and information transmission for a NOMA-based IoT network in [226]. Cache-aided NOMA is not explored much yet in the EH-enabled networks. However, researchers are implementing cache and NOMA separately and cache-aided NOMA primarily to enhance energy efficiency, QoE and reduce the energy consumption of SWIPT networks.

F. mmWave Communication

Millimeter Wave (mmWave) band ranges roughly from 30GHz to 300GHz, providing an alluring spectrum bandwidth of 270GHz. Compared to existing wireless technologies, mmWave proves advantageous in terms of available bandwidth, size of elements, and narrowed beams [228]. Despite all the privileges that mmWave technology offers, it is very challenging for the practical implementation mmWaves for 5G networks. The prime reason behind the shortcomings of mmWaves is it is channel characteristics tend to have high path loss, significant atmospheric absorption, and have difficulty in non-line-of-sight communication [228], [229].

While dealing with mmWave communications, one of the drawbacks is multiple access because of high power consumption and costly hardware [230]. MmWave, when integrated with NOMA, can overcome this limitation as NOMA can provide access to multiple users simultaneously in Power Domain. Not only can it increase the number of users, but also it can contribute to better data rates and reduced interference [231]. In [232], when compared with existing LTE systems, a significant capacity improvement was achieved on combining mmWave-NOMA with massive MIMO systems. Comparative analysis of the Outage probability for downlink NOMA-mmWave network over small-scale fading channels concluded that NOMA outperforms OMA in multi-cell-based mmWaves systems [233]. [234] suggested application of agile beam NOMA for mmWave networks and observes that NOMA-mmWave has better coverage probability and sum-rate than OMA mmWave networks. Compared with TDMA based MIMO-mmWave networks for D2D communications, for NOMA-based MIMO-mmWave networks, Outage Probability tends to decrease exponentially, whereas ergodic capacity increases linearly [235]. All the before-mentioned literature used the Nakagami-m Fading model to represent small-scale fading, whereas [236] used Fluctuating Two-Ray model (FTR) obtained the same results with a better precision. Thus, we can summarize that NOMA-mmWave based 5G networks can accomplish a better overall throughput than conventional OMA-based Networks.

1) NOMA in mmWave: Next, we will discuss some of the areas where integration of NOMA and mmWave can enhance the system throughput.

- **Beamforming mmWave-NOMA**: For mmWave-NOMA, beamforming is achievable in two ways: single beamforming and multi-beamforming. In the case of multiple users, single beamforming is rather disadvantageous because it will require wider beams, reducing the beam gain. Thus, authors in [230] suggested multi-beamforming for multi-users where the base station can simultaneously have multiple narrow beams directed towards multiple users. It will provide higher beam gain, robustness, and a better sum-rate. Despite its supremacy, mmWave-Noma with multi-beamforming has challenging Antenna Wave Vector (AWV) design due to Constant Modulus (CM) constraint. It occurs due to the presence of phase-shifters which are generally non-convex and high dimensional. Currently, research is going on in this direction, and further studies in this direction are required.
- **MmWave Massive MIMO**: With mmWave communications, using a large-scale antenna array is feasible, compensating for poor propagation conditions for mmWaves. The use of mmWave technology in massive MIMO enables low-cost-low-power components. However, the combination of mmWave and massive MIMO has its challenges. Pilot contamination, prior knowledge of accurate CSI, accurate channel estimation, channel feedback, and real-time realization are significant concerns [237]. Since the mmWave massive MIMO channels are not independent and identically distributed (i.i.d), the channel model can no longer be realized as orthogonal practically. Thus, exploration of non-orthogonal techniques such as NOMA was encouraged [237]. For a large number of users scenario, NOMA can further reduce the problem of pilot contamination using power domain NOMA and SIC [238]. NOMA also reduces the outage probability and improves spectral efficiency and energy efficiency of the mmWave massive MIMO systems [239]. MmWave-NOMA system needs the channel information at the base station, which produces significant overhead data. We can expect this overhead to increase further with the inclusion of MIMO, and with massive MIMO, this overhead data would be very large [240]. Thus, a detailed study on this aspect of the mmWave-massive MIMO-NOMA system is needed.
- **Cognitive mmWave Networks**: Cognitive Radios (CR) are the Dynamic Spectrum Access Networks where a licensed primary user shares spectrum with the unlicensed secondary users for transmission. Underlay CR networks enable the secondary user to transmit the data in the presence of the primary user on the condition that secondary user transmission power is low compared to that of the primary user. The Secondary user thus cannot transmit over a long range. Since mmWaves also work on short-range, mmWaves can be the enabling technology for CR systems to provide higher capacity and data rate at increased spectrum efficiency [241], [242]. Since power-domain NOMA can easily do justice on power handling for both primary and secondary users, integration of NOMA with Cognitive mmWave Networks has good potential. [243], [244] has discussed the security aspects of NOMA-based Cognitive mmWave Networks, but other aspects are still unexplored.

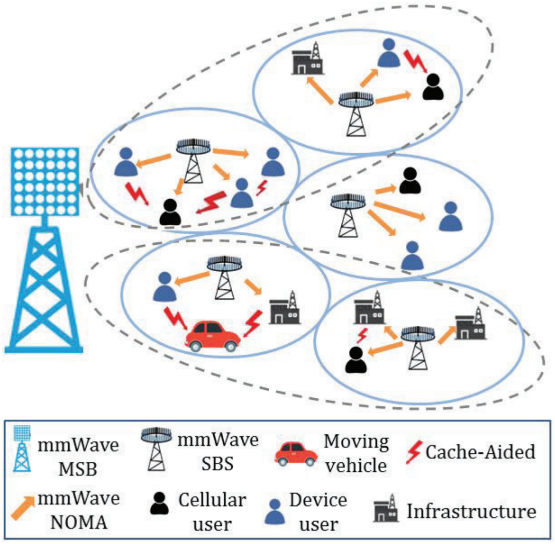


Fig. 12: Cache-Aided NOMA for mmWave.

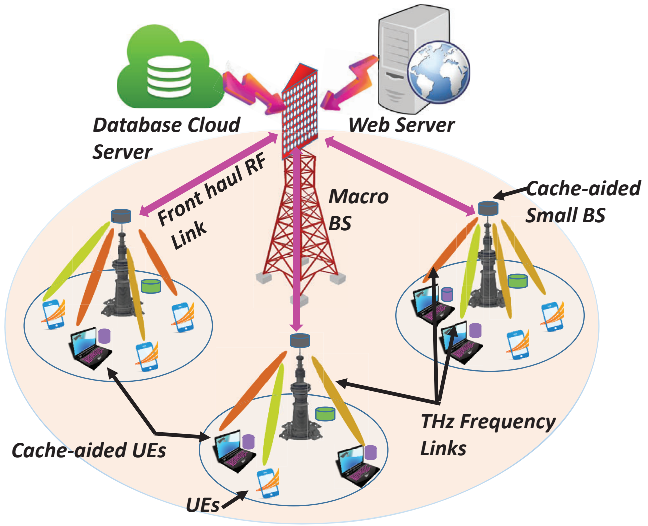


Fig. 13: Scheme for THz based Cache-Aided NOMA.

2) Caching with mmWave NOMA: MmWave-NOMA requires a considerable amount of backhaul overhead. Cache-aided systems can overcome this problem by saving the contents on the cache-enabled user device and small base stations. Caching for millimeter waves was suggested in [245], where the authors acknowledged that the use of cache in mmWaves could reduce frequent handovers and handover failures. Authors exploited the high storage capacity of the modern smartphone to store the data in mobile user equipment (MUE) and retrieve using high capacity mmWaves when required. It eases the overhead backhaul, especially in fast-moving MUEs, alleviates service delay problems.

Although the research community appreciated the use of cache in mmWave and NOMA; cache-aided mmWave NOMA is not adequately addressed. In this article, we propose a mmWave NOMA system that uses caching to reduce the backhaul data. MmWaves at 28GHz can have a larger coverage area than 60GHz as in prior frequency mmWaves get less attenuated. In Fig. 12 we have considered a mmWave macro base station (MBS) antenna transmitting at 28GHz using NOMA aided beamsteering. Each cell contains a small base station (SBS) antenna to receive the signals from the macro base station. Users of each cell get connected to SBS for communication. These users can be pedestrians, any infrastructure, vehicle, device, grid, etc. Also, these users may or may not be cache enabled. Due to the large storage capacity in modern-day devices, these devices can act as a cache-enabling platform. When a non-cached user requires specific data like a music file, the user sends its request to its SBS. SBS, in turn, asks the cache-enabled devices where the data gets stored in cache memory based on the principle of popularity. If the required file is available in any cache-enabled devices, the device sends the file to SBS, which forwards the file to the requesting user. The process will reduce the time and backhaul networking required by the SBS to reach out to MBS, which downloads the music file from the server. If the data is not available in the cell, the SBS may connect to nearby SBSs for the content, successively asking cache-enabled devices in their coverage area. If available, the content gets transferred through backhaul networking. If the content is not available in any cache-enabled devices, the SBS will request the content to MBS.

3) Caching with Tera Hertz-NOMA: Compared to mmWave bands, THz bands provide higher bandwidth, lower eavesdropping and less free-space diffraction [246]. Even at a distance of 5 meters, the NLOS THz propagation channel capacity reaches 100Gbps [247]. However, high frequency selective path loss, non-existent multipath gains and a complex mMIMO system threaten THz deployment [248]. Studies suggest that introducing NOMA systems for THz bands can fill these gaps efficiently [249], [250]. NOMA can reduce the processing complexity at the receiver, diversify highly correlated channels, and increase user fairness [250]. Ulgen *et al.* [248] showed that for a 350 GHz downlink, NOMA outperforms OMA in terms of the data rate. Further, the NOMA-based THz system has a data rate three times higher than OMA for the same distance and environmental conditions. Although NOMA-THz systems outsmart OMA-THz systems, studies suggest that NOMA-THz systems lags behind NOMA based mmWave/microwave systems.

In [251], authors showed that for a BS-mMTC communication, the spectral and energy efficiency for THz falls below microwave and mmWave. For the THz-NOMA, a high number of users (more than two) in a beam cluster can cause significant user interference causing degradation in received signal quality and delayed SIC. Also, optimizing the power allocation can become non-convex and computationally complex [250]. To mitigate the problem of spectral efficiency, Zhang *et al.* [252] recommended a hybrid MIMO-NOMA two-tier architecture where the MBS-SBS link is RF-based, and SBS-UEs links are THz based. Since RF-based networks can cover a larger area, multiple cache-aided SBS are connected to a single MBS. THz signals cannot travel long distances; thus, the SBS will have smaller coverage but a very high data rate compared to RF counterparts. However, such an arrangement will result in

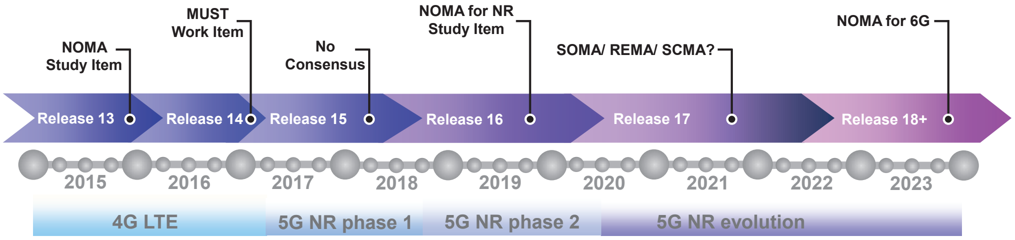


Fig. 14: Timeline of 3GPP releases and different NOMA related standardization activities. NOMA was included in one of the work items in release 16. In release 13 and release 16, discussions were limited to study items only.

a data rate mismatch between the two tiers. This mismatch can be reduced by utilizing the available cache at the SBS and in the UEs. Fig. 13 shows that one can combine the cache memory and NOMA to generate THz-based cache-aided NOMA networks. As discussed in previous articles, cache-aided NOMA can increase the spectral and energy efficiency. There are still wide open research questions and a detailed study on THz-based cache-aided NOMA is lacking.

G. Intelligent Reflecting Surfaces

Intelligent Reflecting Surfaces (IRS) is an intelligent metasurface manufactured of a large number of programmable metamaterials. An IRS can mitigate the wave propagation blockage problem, enhance signal power, and suppress interference by dynamically adjusting the phase and polarization of the incident wave [253], [254]. IRS is envisioned as a potential technology for 6G because of its ability to improve the QoS of wireless networks by customizing the wireless propagation environment. The IRS technique constructively tunes the channel vector of users and boosts the advantages of implementing the NOMA transmission technique. In [255], the authors have validated that an IRS-enabled NOMA outperforms the IRS-enabled OMA in terms of outage probability. The IRS technique mostly implemented in NOMA systems to further increase the coverage [256], energy efficiency [257], sum-rate [258], [259]. Ding *et al.* proposed an IRS-assisted NOMA transmission such that the network can serve more users compared with spatial division multiple access [256]. Article [260] introduces an IRS-aided edge caching system to realize the maximum benefit of caching. Here, the authors formulated a network cost minimization problem regarding backhaul capacity and the transmission power to optimize the content placement. Although no article has analyzed the performance of cache-aided IRS-enabled NOMA systems, we can further improve the performance of cache-aided NOMA networks by implementing IRS in the AP-to-user link.

VI. The Road Ahead

Cache-aided NOMA networks are constantly evolving and are getting diversified with advancement in other domains. In this section, we highlight the recent related standardization activities first, and then point out a few open research challenges.

A. Standardization Activities

A communication standard ensures interoperability among vendors, establishes conformity with local/ international regulations, boosts confidence of startups, and attracts venture capitals/ tech giants. Large investments demand a fixed turnaround time, which in turn, reduces the technology rollout phase. Although cache-aided NOMA is mostly a concept so far, the standardization activities is a proof that the concept will soon become a reality. For example, the discussions in the extreme high throughput (EHT) study group of IEEE 802.11 regarding semi-orthogonal multiple access (SOMA) [261] inspired a NOMA prototype based on software defined radio (SDR) for Wi-Fi [262].

Non-orthogonal multiple access has heavily influenced multiplexing schemes used for digital television (DTV). The advanced television systems committee (ATSC) 3.0 standard uses layered division multiplexing (LDM) [263]. LDM is a two-layer technique [264]. Unlike OMA, both of these layers use full frequency spectrum and full-time duration. Multiplexing between the layers is achieved through PD-NOMA, i.e., the layers are different at power levels. The upper layer has a higher power allocation and is meant for broadcast services to mobile terminals whereas, the lower layer has a lower power allocation and is used for multicasting to fixed reception terminals. While NOMA is fully incorporated in the American DTV standard ATSC, there had been many proposals to include NOMA in the other major DTV standard, Digital Video Broadcasting (DVB). These include the use of NOMA for satellite DVB-S2X [265] and for terrestrial DVB-T [266]. A comprehensive account of NOMA based broadcast services may be found in the recent article by Shariatzadeh *et al.* [267].

Within the 3rd Generation Partnership Project (3GPP), NOMA was first included in LTE Release 13 as a study item (SI) [268] and later, in Release 14, NOMA was included as a work item (WI) [269] under the name of multi-user superposition transmission (MUST) [270]. MUST is a grant-based downlink technique. Simulation studies showed that with MUST a 10% to 30% improvement in throughput is possible depending on other network deployment parameters. Release 15 in mid-2018 marks the beginning of 5G new radio (NR), which continued to advance in Release 16 also known as 5G NR Evolution. The grant-free uplink NOMA was included in Release 16 SI [271]. A complete summary of NOMA-based standardization activities within 3GPP is available in the article by Chen *et al.* [272] and later by Yuan *et al.* [273]. On the other hand, the article by Cirik *et al.* [274] discusses NOMA standardization in the larger grant-free landscape. Release 17 is due on 3rd quarter (Q3) of 2022, and there had been some indications that NOMA, in its contention-based grant-free form, will continue to play an important role. The proposed variants include rate-adaptive constellation expansion multiple access (REMA) [275].

Despite the interest, NOMA was not adopted in any of the work items for 5G NR, as seen in Fig. 14. Rather, OFDMA continues to be the downlink MA choice while single-carrier FDMA (SC-FDMA) has been finalized for uplink. This is because no consensus regarding NOMA could be formed by the large number of stakeholders [276]. However, as pointed out earlier, some variations of contention-based grant-free NOMA is being proposed for 6G. In [277], a variation of CD-NOMA, named as sparse-code multiple access (SCMA), is discussed. Exploiting the sparsity of the codebook matrix, the multi-user detection can be performed in SCMA with much lower complexity than maximum-likelihood detection.

NOMA has been the central topic for various companies' whitepapers [278]. The first in the league is NTT DOCOMO, which published a series of technical documents [279], [280]. Huawei, in addition to NOMA and SOMA, proposed a third variant, rate-adaptive constellation expansion multiple access (REMA) [281]. The interest in NOMA has been manifested by other large telecos (ZTE Corporation, SK Telecom) [282], chip suppliers (Intel, Qualcomm) [283], OEMs (LG Electronics, Samsung, Nokia) [284] and equipment vendors (Anritsu) [285].

Cache can aid NOMA based networks in multiple ways. One possible avenue is to reduce pilot overheads or completely get rid of pilots through prior statistical knowledge of data. Also, with prior knowledge it is possible to build connections keeping the radio resource control (RRC) in idle or in inactive state [286]. There is a related two-step random access channel (RACH) standardization within 3GPP as well [287].

B. Open Research Challenges

1) Joint Sensing and Communication Frameworks: The 6G wireless communications systems are envisioned as joint radar sensing and communication paradigms, which simultaneously sense targets and communicate with the users. An integrated sensing and communication (ISAC) network empowered by cache-aided NOMA shares the spectral resources and infrastructure and could be an evolutionary framework for the next-generation communication systems. The communication signal can be exploited for target sensing by suitably designing the co-variance matrix of the transmitted signal [288]. The cache empowered BS transmits a superimposed signal satisfying the necessary standard for target sensing hence, the superimposed signal can be exploited for communication and sensing also. The users employ the SIC process to recover their signals like the conventional process. The primary aim of the sensing system is to maximize the power of the probing signal towards the direction of targets [288]. Therefore, understanding the requirement, we can extend cache-aided NOMA for joint radar sensing and communications. The NOMA-aided joint radar and communication paradigm has been investigated to empower double spectrum sharing, where superimposed multicast and unicast communication signals have been exploited as radar probing waveforms [289]. A beamforming design problem of a NOMA-ISAC system has been addressed to maximize the sum throughput for the communication system and enhance the effective sensing power [290].

2) STAR-RIS Network for 360° Coverage: The only function of reflectors in the conventional IRS systems is to reflect incident signals constructively towards destinations. In this topology, transmitters and receivers need to be on the same side of the reflector, which restricts the flexible employment of the IRS systems. To deal with this shortcoming and facilitate more flexible communication systems, simultaneous transmitting and reflecting RISs (STAR-RISs) can be employed. Unlike conventional IRS reflectors, STAR-RIS divides incident signals into two parts, one part reflects from the surface, and another part propagates into the other side of the RIS. Recently, Mu *et al.* proposed three STAR-RIS operating protocols, namely energy splitting, mode switching, and time switching, and formulated a power consumption minimization problem for all the protocols under the constraint of data rate [291]. However, the NOMA and cache-aide NOMA in the STAR-RIS research domain is yet not explored and could be an evolutionary application for next-generation communications.

3) NOMA-Empowered Robotic Users: Future human societies in different fields will be surrounded by the application of robotic techniques from smart homes to smart factories. Instead of operating robots on their self-centered individual computational units, robots can be connected with wireless networks as robotic users and operated exchanging information with the APs [292]. The challenges associated with the application areas of robotic users make difficulties in resource management. Operating large numbers of robots is much more challenging, especially when they are deployed for different tasks. To deal with this scenario, researchers have recently started initial research work to investigate the capability of NOMA in robotic communications [293].

4) Orthogonal Time Frequency Space (OTFS)-NOMA: Providing communication maintaining 5G standards to various types of users with different mobility profiles is one of the essential objectives of 5G and B5G systems. Doppler frequency shifts and frequent channel estimation with reliability are two central challenges for high-mobility users. Doppler frequency shift introduces inter-carrier interference, and channel parameters realization timely causes additional system overhead. Orthogonal time-frequency space (OTFS) modulation has been proposed recently to encounter high mobility-related issues [294]. In OTFS, the primary task is to place high-mobility users' signals in a delay-Doppler plane and converts the channels that are time-varying in the time-frequency plane to time-invariant channels in the delay-Doppler plane. As a result, delay-Doppler plane can directly estimate the channel parameters. Now consider a scenario when highly mobile users occupy the bandwidth resources and time slots, and the users do not require a high data rate or channel gain quality is poor. In this case, the spectral efficiency of OTFS may be low and NOMA-based OTFS can be a solution to this. In [295], the authors proposed OTFS-NOMA to improve the spectral efficiency and delivery latency with heterogeneous mobility profiles. The OTFS-NOMA technique groups users with different mobility for implementing the NOMA principle. In this domain, NOMA and cache have not been explored much.

5) NOMA-aided Internet of Health: Internet of health (IoH) is steadily emerging as a necessary service for human health monitoring under the forthcoming 6G communications. IoH services are required to communicate with a massive number patients' electronic medical devices to improve their quality of health. IoH services include real-time remote diagnosis, remote treatment (telemedicine), and remote surgery for emergencies. NOMA is a promising candidate capable of simultaneously transmitting information to multiple patients maintaining coordination among numerous smart devices. To establish in-home medical networks, Xuewan *et al.* proposed a multi-carrier NOMA framework that connects comparatively more monitoring units and transfers more information bits compared to OMA-based designs [296]. Based on patients' medical history, some medical advices as first aid services can be cached in the health monitoring unit. Cache enabled NOMA is particularly useful because, often, the medical information is data heavy (high resolution tomography or video) and local storage of patient data is useful considering their limited mobility.

6) NOMA-aided Visible Light Communications: For short-range communications, visible light communication (VLC) is a promising communication scheme operated through an unlicensed spectrum with high secrecy and low energy. Efficient MA techniques are required to be implemented for VLC systems to improve spectral efficiency since the modulation bandwidth is narrow in the VLC. NOMA could be an attractive MA for the VLC. Each transmitter of practical VLC serves has a limited number of users which leads to a reduced SIC-based NOMA decoding process keeping control of the outage probabilities. Furthermore, the slow-varying channel characteristic of the VLC reduces the overhead and complexity required for accurate CSI estimation at the NOMA transmitter. These two unique features of the VLC make the NOMA-based VLC advantageous over the OMA-VLC in terms of the ergodic sum-rate [297]. MIMO-NOMA-based VLC is also becoming a popular approach to enhance the sum-rate [298]. Depending on the requirement and network model, the NOMA-based VLC can be extended to cache-aided NOMA-based VLC systems.

7) Hybrid Free Space Optical Communication: Like the VLC, free-space optical (FSO) communication is another short-range, line-of-sight communication system exploiting the optical domain for the next generation. The FSO communication system fundamentally works on intensity modulation of laser and direct detection by a photodetector in the near-infrared range ($750 - 1600$ nm) and transmits data in Gbps range through high bandwidth optical channels. The FSO communication uses a laser as a transmitter with low implementation cost, and the directional communication nature provides higher security compared to traditional communication systems. Despite so many advantages, weather conditions, atmospheric turbulence, and pointing errors (misalignment between the optical transmitter and receiver) are the inescapable challenges of FSO systems. In [299] the authors have analyzed the performance of a mixed RF-FSO system and validated the superiority of the FSO backhauling and high-reliability NOMA systems over the conventional RF backhauling. Cooperative relay transmission capability can make NOMA the most suitable technology for hybrid FSO/radio frequency communication systems. Cache-aided NOMA systems are particularly attractive for hybrid RF/FSO systems as cache can help in alleviating issues like difference of data rates over parallel RF and FSO links.

VII. Conclusion

This article presented a comprehensive survey and reviewed the state-of-the-art research contributions of the cache-aided NOMA technique in wireless communications. First, we explained the fundamental concepts, operating principles, and major challenges associated with the cache and NOMA techniques. We then flexibly amalgamated cache with NOMA and discussed the motivations and goals of cache-aided NOMA-based wireless networks. This article explicitly presented the primary considerations related to cache-aided NOMA network design, including cache memory allocation, most popular content selection, and optimal content placement. This survey paper thoroughly reviewed the frameworks for achieving primary goals such as the increased probability of successful decoding, sum-rate maximization, delay reduction, interference cancellation, and energy consumption minimization. We found efficient placement of popular contents and suitable power allocation are the two essential requirements for designing cache-aided NOMA systems. Next, we categorized the research articles depending on the application scenarios of cache-aided NOMA systems. We then identified that the cache-aided NOMA technology is mostly applied in vehicular communications, UAV-based networks, MEC, and cellular communications. Advantages and challenges related to the utilization of cache-aided NOMA technology in these scenarios are presented. We discussed the benefits of using cache-aided NOMA in these scenarios, and concluded that balancing the performance and energy consumption is the most challenging tasks. Finally, we highlighted existing open challenges and future research directions of the cache-aided NOMA technology.

References

[1] K. Buchholz. (2021, Aug.) Where 5G technology has been deployed. Infographic. Statista. [Online]. Available: https://www.statista.com/chart/23194/5g-networks-deployment-world-map/ (accessed Sep. 10, 2021).

[2] C. Casetti, "Promises come to fruition as 5G reaches critical mass," *IEEE Veh. Technol. Mag.*, vol. 16, no. 3, pp. 6–13, Sep. 2021.

[3] S. Dang, O. Amin, B. Shihada, and M. S. Alouini, "What should 6G be?" *Nature Electron.*, vol. 3, no. 1, pp. 20–29, Jan. 2020.

[4] F. Götz. (2021, Jan.) The data deluge: What do we do with the data generated by AVs? Blog. Siemens. [Online]. Available: https://blogs.sw.siemens.com/polarion/the-data-deluge-what-do-we-do-with-the-data-generated-by-avs/ (accessed Sep. 10, 2021).

[5] S. Panwar, "Breaking the millisecond barrier: Robots and self-driving cars will need completely reengineered networks," *IEEE Spectrum*, vol. 57, no. 11, pp. 44–49, Nov. 2020.

[6] Z. Xiang, F. Gabriel, E. Urbano, G. T. Nguyen, M. Reisslein, and F. H. P. Fitzek, "Reducing latency in virtual machines: Enabling tactile internet for human-machine co-working," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 5, pp. 1098–1116, May 2019.

[7] J. Park, K. Lee, and Y. Park, "Ultrathin wide-angle large-area digital 3D holographic display using a non-periodic photon sieve," *Nature Commun.*, vol. 10, no. 1304, pp. 1–8, Mar. 2019.

[8] P.-H. C. Chen, K. Gadepalli, R. MacDonald, Y. Liu, S. Kadowaki, K. Nagpal, T. Kohlberger, J. Dean, G. S. Corrado, J. D. Hipp, C. H. Mermel, and M. C. Stumpe, "An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis," *Nature Medicine*, vol. 25, no. 9, pp. 1453–1457, Sep. 2019.

[9] e. a. Yu, Xinge, "Skin-integrated wireless haptic interfaces for virtual and augmented reality," *Nature*, vol. 575, no. 7783, pp. 473–479, Nov. 2019.

[10] J. Erman, A. Gerber, M. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: The 3G case," *IEEE Internet Comput.*, vol. 15, no. 2, pp. 27–34, Mar. 2011.

[11] 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 Communications*, vol. 14, no. 4, pp. 17–25, Oct. 2016.

[12] Z. Ding, X. Lei, G. K. Karagiannidis, R. Schober, J. Yuan, and V. K. 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, Oct. 2017.

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

[14] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, "A tutorial on nonorthogonal multiple access for 5G and beyond," *Wirel. commun. and mobile computing*, vol. 2018, 2018.

[15] M. Vaezi, G. A. Aruma Baduge, Y. Liu, A. Arafa, F. Fang, and Z. Ding, "Interplay between NOMA and other emerging technologies: A survey," *IEEE Trans. Cognitive Commun. Networking*, vol. 5, no. 4, pp. 900–919, Dec. 2019.

[16] O. Maraqa, A. S. Rajasekaran, S. Al-Ahmadi, H. Yanikomeroglu, and S. M. Sait, "A survey of rate-optimal power domain NOMA with enabling technologies of future wireless networks," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2192–2235, Fourthquarter 2020.

[17] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, "A survey of NOMA: Current status and open research challenges," *IEEE Open J. Commun. Soc.*, vol. 1, pp. 179–189, Jan. 2020.

[18] A. Akbar, S. Jangsher, and F. A. Bhatti, "NOMA and 5G emerging technologies: A survey on issues and solution techniques," *Comput. Networks*, vol. 190, p. 107950, May 2021.

[19] H. Yahya, E. Alsusa, A. Al-Dweik *et al.*, "Error rate analysis of NOMA: Principles, survey and future directions," Jan. 2022.

[20] A. Passarella, "A survey on content-centric technologies for the current internet: CDN and P2P solutions," *Comput. Commun.*, vol. 35, no. 1, pp. 1–32, Jan. 2012.

[21] G. Xylomenos, C. N. Ververidis, V. A. Siris, N. Fotiou, C. Tsilopoulos, X. Vasilakos, K. V. Katsaros, and G. C. Polyzos, "A survey of information-centric networking research," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 2, pp. 1024–1049, Second Quarter 2014.

[22] G. Zhang, Y. Li, and T. Lin, "Caching in information centric networking: A survey," *Comput. netw.*, vol. 57, no. 16, pp. 3128–3141, Nov. 2013.

[23] C. Wang, Y. He, F. R. Yu, Q. Chen, and L. Tang, "Integration of networking, caching, and computing in wireless systems: A survey, some research issues, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 1, pp. 7–38, Fourth quarter 2018.

[24] L. Li, G. Zhao, and R. S. Blum, "A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1710–1732, Third quarter 2018.

[25] M. I. A. Zahed, I. Ahmad, D. Habibi, Q. V. Phung, M. M. Mowla, and M. Waqas, "A review on green caching strategies for next generation communication networks," *IEEE Access*, vol. 8, pp. 212 709–212 737, Nov. 2020.

[26] A. Kabir, G. Rehman, S. M. Gilani, E. J. Kitindi, Z. Ul Abidin Jaffri, and K. M. Abbasi, "The role of caching in next generation cellular networks: A survey and research outlook," *Trans. Emerg. Telecommun. Technol.*, vol. 31, no. 2, p. e3702, 2020.

[27] H. Al-Ward, C. K. Tan, and W. H. Lim, "Caching transient data in information-centric internet-of-things (IC-Iot) networks: A survey," *Journal of Network and Computer Applications*, p. 103491, 2022.

[28] E. Zeydan, E. Bastug, M. Bennis, M. A. Kader, I. A. Karatepe, A. S. Er, and M. Debbah, "Big data caching for networking: moving from cloud to edge," *IEEE Commun. Mag.*, vol. 54, no. 9, pp. 36–42, Sep. 2016.

[29] 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.

[30] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Caching meets millimeter wave communications for enhanced mobility management in 5g networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 779–793, Feb. 2018.

[31] ——, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 779–793, Feb. 2018.

[32] W. Hao, M. Zeng, G. Sun, and P. Xiao, "Edge cache-assisted secure low-latency millimeter-wave transmission," *IEEE Internet Things J.*, vol. 7, no. 3, pp. 1815–1825, Mar. 2020.

[33] Y. Cao, M. Tao, F. Xu, and K. Liu, "Fundamental storage-latency tradeoff in cache-aided MIMO interference networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5061–5076, Aug. 2017.

[34] N. Garg, M. Sellathurai, V. Bhatia, and T. Ratnarajah, "Function approximation based reinforcement learning for edge caching in massive MIMO networks," *IEEE Trans. Commun.*, vol. 69, no. 4, pp. 2304–2316, Apr. 2021.

[35] X. Wang, Y. Han, C. Wang, Q. Zhao, X. Chen, and M. Chen, "In-edge AI: Intelligentizing mobile edge computing, caching and communication by federated learning," *IEEE Netw.*, vol. 33, no. 5, pp. 156–165, Oct. 2019.

[36] T. X. Tran and D. Pompili, "Adaptive bitrate video caching and processing in mobile-edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 18, no. 9, pp. 1965–1978, Sep. 2019.

[37] H. Zhang, H. Zhang, J. Dong, V. C. Leung *et al.*, "Energy efficient user clustering and hybrid precoding for terahertz MIMO-NOMA systems," in *Proc. ICC*. IEEE, Jul 2020, pp. 1–5.

[38] Y. Liu, Z. Qin, M. Elkashlan, Z. Ding, A. Nallanathan, and L. Hanzo, "Non-orthogonal multiple access for 5G and beyond," *Proc. IEEE*, vol. 105, no. 12, pp. 2347–2381, Dec. 2017.

[39] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.

[40] Y. Saito, A. Benjebbour, Y. Kishiyama, and T. Nakamura, "System-level performance of downlink non-orthogonal multiple access (NOMA) under various environments," in *Proc. VTC (Spring)*, May 2015, pp. 1–5.

[41] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2667–2680, Dec. 2017.

[42] X. Yue, Z. Qin, Y. Liu, S. Kang, and Y. Chen, "A unified framework for non-orthogonal multiple access," *IEEE Trans. Commun.*, vol. 66, no. 11, pp. 5346–5359, Nov. 2018.

[43] C. Yang, X. Wang, B. Xia, and H. Ding, "Joint interference cancellation in cache- and SIC-enabled networks," in *Proc. Globecom*, Dec. 2017, pp. 1–6.

[44] Y. Saito, Y. Kishiyama, A. Benjebbour, T. Nakamura, A. Li, and K. Higuchi, "Non-orthogonal multiple access (NOMA) for cellular future radio access," in *Proc. VTC (Spring)*, 2013, pp. 1–5.

[45] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, 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.

[46] M. Ji, G. Caire, and A. F. Molisch, "Wireless device-to-device caching networks: Basic principles and system performance," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 1, pp. 176–189, Jan. 2016.

[47] S. Glass, I. Mahgoub, and M. Rathod, "Leveraging MANET-based cooperative cache discovery techniques in VANETs: A survey and analysis," *IEEE Commun. Surveys Tuts.*, vol. 19, no. 4, pp. 2640–2661, Fourth quarter 2017.

[48] J. Wang, "A survey of web caching schemes for the internet," *ACM SIGCOMM Computer Communication Review*, vol. 29, 10 1999.

[49] W. Ali, S. M. Shamsuddin, A. S. Ismail *et al.*, "A survey of web caching and prefetching," *Int. J. Advance. Soft Comput. Appl*, vol. 3, no. 1, pp. 18–44, Mar. 2011.

[50] U. Niesen, D. Shah, and G. Wornell, "Caching in wireless networks," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, Jun./Jul. 2009.

[51] K.-T. M. K.-Y. C. K.-S. L. V. Tam, "Improving data centric storage with diffuse caching in wireless sensor networks," *Wireless Commun. and Mobile Comput.*, vol. 9, Apr. 2009.

[52] "Video aware wireless networks," https://software.intel.com/content/www/us/en/develop/articles/video-aware-wireless-networks.html, accessed: 2012-07-30.

[53] "Multiple access for 5G new radio interface," *Tech. Rep. 3GPP R1-162305*, CATT, Busan, Korea, Apr. 2016.

[54] P. Nuggehalli, V. Srinivasan, C.-F. Chiasserini, and R. Rao, "Efficient cache placement in multi-hop wireless networks," *IEEE/ACM Trans. Networking*, vol. 14, no. 5, pp. 1045–1055, Oct. 2006.

[55] K. Shanmugam, N. Golrezaei, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless content delivery through distributed caching helpers," *IEEE Trans. Inf. Theory*, vol. 59, no. 12, pp. 8402–8413, Dec. 2013.

[56] L. Breslau, P. Cao, L. Fan, G. Phillips, and S. Shenker, "Web caching and zipf-like distributions: evidence and implications," in *Proc. INFOCOM*, vol. 1, Apr 1999, pp. 126–134.

[57] M. Ji, A. M. Tulino, J. Llorca, and G. Caire, "Order-optimal rate of caching and coded multicasting with random demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 6, pp. 3923–3949, Jun. 2017.

[58] J. Hachem, N. Karamchandani, and S. N. Diggavi, "Coded caching for multi-level popularity and access," *IEEE Trans. Inf. Theory*, vol. 63, no. 5, pp. 3108–3141, May 2017.

[59] Y. Fadlallah, A. M. Tulino, D. Barone, G. Vettigli, J. Llorca, and J.-M. Gorce, "Coding for caching in 5G networks," *IEEE Commun. Mag.*, vol. 55, no. 2, pp. 106–113, Feb. 2017.

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

[61] M. N. Dani and D. K. So, "On the performance of NOMA and coded multicasting in cache-aided wireless networks," in *Proc. ICC*. IEEE, May 2019, pp. 1–6.

[62] F. Xu, M. Tao, and K. Liu, "Fundamental tradeoff between storage and latency in cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 63, no. 11, pp. 7464–7491, Nov. 2017.

[63] J. Pedersen, A. Graell i Amat, I. Andriyanova, and F. Brännström, "Optimizing MDS coded caching in wireless networks with device-to-device communication," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 286–295, Jan. 2019.

[64] J. Rao, H. Feng, C. Yang, Z. Chen, and B. Xia, "Optimal caching placement for D2D assisted wireless caching networks," in *Proc. ICC*, May 2016, pp. 1–6.

[65] N. Dimokas, D. Katsaros, and Y. Manolopoulos, "Cooperative caching in wireless multimedia sensor networks," *Mobile Networks and Applications*, vol. 13, no. 3, pp. 337–356, 2008.

[66] X. Li, X. Wang, and V. C. M. Leung, "Weighted network traffic offloading in cache-enabled heterogeneous networks," in *Proc. ICC*, 2016, pp. 1–6.

[67] A. Shokrollahi, "Raptor codes," *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.

[68] D. MacKay, "Fountain codes," *IEE Proc. Commun.*, vol. 152, Dec. 2005.

[69] E. Altman, K. Avrachenkov, and J. Goseling, "Coding for caches in the plane," *arXiv preprint arXiv:1309.0604*, 2013.

[70] P. Ostovari, A. Khreishah, and J. Wu, "Cache content placement using triangular network coding," in *Proc. WCNC*, Apr. 2013, pp. 1375–1380.

[71] J. Zhang, X. Lin, and X. Wang, "Coded caching under arbitrary popularity distributions," *IEEE Trans. Inf. Theory*, vol. 64, no. 1, pp. 349–366, Jan. 2018.

[72] M. Ji, A. Tulino, J. Llorca, and G. Caire, "Caching-aided coded multicasting with multiple random requests," in *Proc. IEEE Inf. Theory Workshop*, May 2015, pp. 1–5.

[73] U. Niesen and M. A. Maddah-Ali, "Coded caching with nonuniform demands," *IEEE Trans. Inf. Theory*, vol. 63, no. 2, pp. 1146–1158, Feb. 2017.

[74] B. Dai and W. Yu, "Joint user association and content placement for cache-enabled wireless access networks," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2016, pp. 3521–3525.

[75] L. Podlipnig, Stefan; Böszörmenyi, "A survey of web cache replacement strategies," *ACM Computing Surveys*, vol. 35, 12 2003.

[76] M. Balamash, Abdullah; Krunz, "An overview of web caching replacement algorithms," *IEEE Commun. Surveys Tuts.*, vol. 6, Second Quarter 2004.

[77] Y.-B. L. . W.-R. L. . J.-J. Chen, "Effects of cache mechanism on wireless data access," *IEEE Trans. Wireless Commun.*, vol. 2, no. 6, Nov. 2003.

[78] H. Chen and Y. Xiao, "Cache access and replacement for future wireless internet," *IEEE Commun. Mag.*, vol. 44, no. 5, pp. 113–123, May 2006.

[79] J. Xu, Q. Hu, W.-C. Lee, and D. L. Lee, "Performance evaluation of an optimal cache replacement policy for wireless data dissemination," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 1, pp. 125–139, Jan. 2004.

[80] C. Zhang, C. Xia, Y. Li, H. Wang, and X. Li, "A hotspot-based probabilistic cache placement policy for ICN in MANETs," *EURASIP J. Wireless Commun. Networks*, vol. 2019, 12 2019.

[81] L. Lei, T. X. Vu, L. Xiang, X. Zhang, S. Chatzinotas, and B. Ottersten, "Optimal resource allocation for NOMA-enabled cache replacement and content delivery," in *Proc. PIMRC*. IEEE, Sep. 2019, pp. 1–6.

[82] J. L. Hennessy and D. A. Patterson, *Computer architecture: A quantitative approach*, 5th ed. Elsevier Science, 2012.

[83] A. Sengupta, R. Tandon, and O. Simeone, "Cache aided wireless networks: Tradeoffs between storage and latency," in *2016 Annual Conference on Information Science and Systems (CISS)*, Apr. 2016, pp. 320–325.

[84] X. Peng, J. Zhang, S. H. Song, and K. B. Letaief, "Cache size allocation in backhaul limited wireless networks," in *Proc. ICC*, Jul. 2016, pp. 1–6.

[85] L. Xiang, D. W. K. Ng, X. Ge, Z. Ding, V. W. Wong, and R. Schober, "Cache-aided non-orthogonal multiple access: The two-user case," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 436–451, Jun. 2019.

[86] Y. Fu, Z. Shi, J. Ke, H. Wang, A. K. Wong, and T. Q. Quek, "Efficient delay minimization algorithm for cache-enabled NOMA systems," *IEEE Wireless Commun. Lett.*, Early access 2021.

[87] F. Cheng, Y. Yu, Z. Zhao, N. Zhao, Y. Chen, and H. Lin, "Power allocation for cache-aided small-cell networks with limited backhaul," *IEEE Access*, vol. 5, pp. 1272–1283, Jan. 2017.

[88] A. Asadi, Q. Wang, and V. Mancuso, "A survey on device-to-device communication in cellular networks," *IEEE Commun. Surveys Tuts.*, vol. 16, no. 4, pp. 1801–1819, Apr. 2014.

[89] R. Wang, J. Zhang, S. H. Song, and K. B. Letaief, "Mobility-aware caching in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 16, no. 8, pp. 5001–5015, Aug. 2017.

[90] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage tradeoff in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, July 2013, pp. 1461–1465.

[91] D.-Y. Kim and J. Cho, "Active caching: a transmission method to guarantee desired communication reliability in wireless sensor networks," *IEEE Commun. Lett.*, vol. 13, no. 6, pp. 378–380, Jun. 2009.

[92] M. Taghizadeh, A. Plummer, and S. Biswas, "Cooperative caching for improving availability in social wireless networks," in *The 7th IEEE International Conference on Mobile Ad-hoc and Sensor Systems (IEEE MASS 2010)*. IEEE, 2010, pp. 342–351.

[93] R. Ma, L. Wang, Y. Chen, M. Pan, and L. Xu, "Enabling edge caching through full-duplex non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 12 338–12 342, Oct. 2020.

[94] Z. Yang, Y. Liu, Y. Chen, and N. Al-Dhahir, "Cache-aided NOMA mobile edge computing: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6899–6915, Oct. 2020.

[95] L. Pu, L. Jiao, X. Chen, L. Wang, Q. Xie, and J. Xu, "Online resource allocation, content placement and request routing for cost-efficient edge caching in cloud radio access networks," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 8, pp. 1751–1767, Aug. 2018.

[96] Y. Shan, Q. Zhu, and Y. Wang, "Performance analysis on a cooperative transmission scheme of multicast and NOMA in cache-enabled cellular networks," *IET Commun.*, vol. 15, no. 7, pp. 946–956, Feb. 2021.

[97] J. Hachem, U. Niesen, and S. N. Diggavi, "Degrees of freedom of cache-aided wireless interference networks," *IEEE Trans. Inf. Theory*, vol. 64, no. 7, pp. 5359–5380, Jul. 2018.

[98] N. Garg, M. Sellathurai, V. Bhatia, B. N. Bharath, and T. Ratnarajah, "Online content popularity prediction and learning in wireless edge caching," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 1087–1100, Feb. 2020.

[99] X. Wang, M. Chen, T. Taleb, A. Ksentini, and V. C. Leung, "Cache in the air: exploiting content caching and delivery techniques for 5G systems," *IEEE Commun. Mag.*, vol. 52, no. 2, pp. 131–139, Feb. 2014.

[100] Z. Ding, Y. Liu, J. Choi, Q. Sun, M. Elkashlan, I. Chih-Lin, 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.

[101] D. Bepari and D. Mitra, "Improved power loading scheme for orthogonal frequency division multiplexing based cognitive radio," *IET Commun.*, vol. 9, pp. 2033–2040, Nov. 2015.

[102] D. Bepari, A. K. Bojja, B. S. Kumar, and D. Mitra, "A spectral distance based power control scheme for capacity enhancement of OFDM cognitive radio," *Wireless Pers. Commun.*, vol. 90, no. 1, pp. 157–173, Apr. 2016.

[103] J. Men and J. Ge, "Performance analysis of non-orthogonal multiple access in downlink cooperative network," *IET Commun.*, vol. 9, no. 18, pp. 2267–2273, Dec. 2015.

[104] Z. Ding, P. Fan, and H. V. Poor, "Impact of user pairing on 5G nonorthogonal multiple-access downlink transmissions," *IEEE Trans. Veh. Technol.*, vol. 65, no. 8, pp. 6010–6023, 2015.

[105] Z. Yang, W. Xu, C. Pan, Y. Pan, and M. Chen, "On the optimality of power allocation for NOMA downlinks with individual QoS constraints," *IEEE Commun. Lett.*, vol. 21, no. 7, pp. 1649–1652, Jul. 2017.

[106] Y. Yuan, Z. Yuan, G. Yu, C.-h. Hwang, P.-k. Liao, A. Li, and K. Takeda, "Non-orthogonal transmission technology in LTE evolution," *IEEE Commun. Mag.*, vol. 54, no. 7, pp. 68–74, Jul. 2016.

[107] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. PIMRC*. IEEE, Sep. 2013, pp. 332–336.

[108] R. Hoshyar, F. P. Wathan, and R. Tafazolli, "Novel low-density signature for synchronous CDMA systems over AWGN channel," *IEEE Trans. Signal Process.*, vol. 56, no. 4, pp. 1616–1626, Apr. 2008.

[109] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-lin, and Z. Wang, "Non-orthogonal multiple access for 5G: solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sept. 2015.

[110] H. Liu, Z. Ding, K. J. Kim, K. S. Kwak, and H. V. Poor, "Decode-and-forward relaying for cooperative NOMA systems with direct links," *IEEE Trans. Wireless Commun.*, vol. 17, no. 12, pp. 8077–8093, Dec. 2018.

[111] Z. Yang, Z. Ding, P. Fan, and N. Al-Dhahir, "The impact of power allocation on cooperative non-orthogonal multiple access networks with SWIPT," *IEEE Trans. Wireless Commun.*, vol. 16, no. 7, pp. 4332–4343, Jul. 2017.

[112] S. Mondal, S. D. Roy, and S. Kundu, "Outage analysis for NOMA-based energy harvesting relay network with imperfect CSI and transmit antenna selection," *IET Commun.*, vol. 14, no. 14, pp. 2240–2249, Aug. 2020.

[113] 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.

[114] 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.

[115] S. Guo and X. Zhou, "Robust resource allocation with imperfect channel estimation in NOMA-based heterogeneous vehicular networks," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2321–2332, Mar. 2019.

[116] W. Sun, D. Yuan, E. G. Ström, and F. Brännström, "Cluster-based radio resource management for D2D-supported safety-critical V2X communications," *IEEE Trans. Wireless Commun.*, vol. 15, no. 4, pp. 2756–2769, Apr. 2016.

[117] L. Liang, J. Kim, S. C. Jha, K. Sivanesan, and G. Y. Li, "Spectrum and power allocation for vehicular communications with delayed CSI feedback," *IEEE Wireless Commun. Lett.*, vol. 6, no. 4, pp. 458–461, Aug. 2017.

[118] M. R. Zamani, M. Eslami, M. Khorramizadeh, and Z. Ding, "Energy-efficient power allocation for noma with imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 1009–1013, Jan. 2019.

[119] M. F. Kader, M. B. Shahab, and S. Y. Shin, "Exploiting non-orthogonal multiple access in cooperative relay sharing," *IEEE Commun. Lett.*, vol. 21, no. 5, pp. 1159–1162, May 2017.

[120] G. Im and J. H. Lee, "Outage probability for cooperative NOMA systems with imperfect SIC in cognitive radio networks," *IEEE Commun. Lett.*, vol. 23, no. 4, pp. 692–695, Apr. 2019.

[121] S. Li, M. Derakhshani, and S. Lambotharan, "Outage-constrained robust power allocation for downlink MC-NOMA with imperfect SIC," in *Proc. ICC*, May 2018, pp. 1–7.

[122] Y. Liu, W. Yi, Z. Ding, X. Liu, O. A. Dobre, and N. Al-Dhahir, "Developing NOMA to next generation multiple access: Future vision and research opportunities," *IEEE Wirel. Commun.*, Dec. 2022.

[123] Z. Chen, Z. Ding, X. Dai, and R. Zhang, "An optimization perspective of the superiority of NOMA compared to conventional OMA," *IEEE Trans. Signal Process.*, vol. 65, no. 19, pp. 5191–5202, Oct. 2017.

[124] M. Moghimi, A. Zakeri, M. R. Javan, N. Mokari, and D. W. K. Ng, "Joint radio resource allocation and cooperative caching in PD-NOMA-based HetNets," *IEEE Trans. Mobile Comput.*, vol. 21, no. 6, pp. 2029–2044, Jun. 2022.

[125] S. Rezvani, S. Parsaeefard, N. Mokari, M. R. Javan, and H. Yanikomeroglu, "Cooperative multi-bitrate video caching and transcoding in multicarrier NOMA-assisted heterogeneous virtualized mec networks," *IEEE Access*, vol. 7, pp. 93 511–93 536, 2019.

[126] Y. Fu, W. Wen, Z. Zhao, T. Q. Quek, S. Jin, and F.-C. Zheng, "Dynamic power control for NOMA transmissions in wireless caching networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1485–1488, Oct. 2019.

[127] Z. Zhao, M. Xu, W. Xie, Z. Ding, and G. K. Karagiannidis, "Coverage performance of NOMA in wireless caching networks," *IEEE Commun. Lett.*, vol. 22, no. 7, pp. 1458–1461, Jul. 2018.

[128] 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.

[129] K. N. Doan, W. Shin, M. Vaezi, H. V. Poor, and T. Q. Quek, "Optimal power allocation in cache-aided non-orthogonal multiple access systems," in *Proc. ICC*. IEEE, May 2018, pp. 1–6.

[130] 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.

[131] Z. Wei, D. W. K. Ng, J. Yuan, and H.-M. Wang, "Optimal resource allocation for power-efficient MC-NOMA with imperfect channel state information," *IEEE Trans. Commun.*, vol. 65, no. 9, pp. 3944–3961, Sep. 2017.

[132] 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.

[133] "Framework and overall objectives of the future development of IMT for 2020 and beyond," *Tech. Rep. ITU-R M.2083-0*, 2015.

[134] "Candidate solution for new multiple access," *Tech. Rep. 3GPP R1-162306*, CATT, Busan, Korea, Apr. 2016.

[135] Y. Fu, Y. Liu, H. Wang, Z. Shi, and Y. Liu, "Mode selection between index coding and superposition coding in cache-based NOMA networks," *IEEE Commun. Lett.*, vol. 23, no. 3, pp. 478–481, Mar. 2019.

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

[137] Y. Li, M. Jiang, Q. Zhang, and J. Qin, "Cache content placement optimization in non-orthogonal multiple access networks," *IEEE Trans. Commun.*, vol. 68, no. 7, pp. 4580–4591, Jul. 2020.

[138] R. Rai, H. Zhu, and J. Wang, "Performance analysis of NOMA enabled fog radio access networks," *IEEE Trans. Commun.*, vol. 69, no. 1, pp. 382–397, Jan. 2021.

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

[140] Z. Zhao, M. Xu, Y. Li, and M. Peng, "A non-orthogonal multiple access-based multicast scheme in wireless content caching networks," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 12, pp. 2723–2735, Dec. 2017.

[141] Y. Yin, M. Liu, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "QoS-oriented dynamic power allocation in NOMA-based wireless caching networks," *IEEE Wireless Commun. Lett.*, vol. 10, no. 1, pp. 82–86, Jan. 2021.

[142] T. Zhang, Z. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Caching placement and resource allocation for cache-enabling UAV NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 12 897–12 911, Nov. 2020.

[143] Z. Ding, P. Fan, G. K. Karagiannidis, R. Schober, and H. V. Poor, "On the application of NOMA to wireless caching," in *Proc. ICC*. IEEE, May 2018, pp. 1–7.

[144] L. Xiang, D. W. K. Ng, X. Ge, Z. Ding, V. W. S. Wong, and R. Schober, "Cache-aided non-orthogonal multiple access," in *Proc. ICC*. IEEE, May 2018, pp. 1–7.

[145] K. Z. Shen, T. E. Alharbi, and D. K. So, "Cache-aided device-to-device non-orthogonal multiple access," in *Proc. VTC (Spring)*. IEEE, May 2020, pp. 1–6.

[146] X. Wei, L. Xiang, L. Cottatellucci, T. Jiang, and R. Schober, "Cache-aided massive MIMO: Linear precoding design and performance analysis," in *Proc. ICC*, Jul. 2019, pp. 1–7.

[147] J. Zhao, Y. Liu, T. Mahmoodi, K. K. Chai, Y. Chen, and Z. Han, "Resource allocation in cache-enabled CRAN with non-orthogonal multiple access," in *Proc. ICC*. IEEE, May 2018, pp. 1–6.

[148] X. Pei, H. Yu, Y. Chen, M. Wen, and G. Chen, "Hybrid multicast/unicast design in NOMA-based vehicular caching system," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16 304–16 308, Dec. 2020.

[149] Y. Liu, F. R. Yu, X. Li, H. Ji, H. Zhang, and V. C. M. Leung, "Joint access and resource management for delay-sensitive transcoding in ultra-dense networks with mobile edge computing," in *Proc. ICC*. IEEE, Jul. 2018, pp. 1–6.

[150] Y. Li, H. Zhang, K. Long, S. Choi, and A. Nallanathan, "Resource allocation for optimizing energy efficiency in NOMA-based fog UAV wireless networks," *IEEE Netw.*, vol. 34, no. 2, pp. 158–163, Mar. 2019.

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

[152] S. Mohan, S. Morgansgate, P. Basket, S. Gurugopinath, and S. Muhaidat, "Cache-aided non-orthogonal multiple access over fading channels in downlink cellular networks," in *2020 International Conference on Communication Systems & NetworkS (COMSNETS)*. IEEE, Jan. 2020, pp. 452–459.

[153] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 630–644, Jan. 2020.

[154] J. Zheng, Q. Zhang, and J. Qin, "Outage probabilities of cache and SIC enabled downlink MIMO NOMA cellular networks with randomly distributed users," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 13 942–13 946, Nov. 2020.

[155] M. N. Dani, D. K. So, J. Tang, and Z. Ding, "NOMA and coded multicasting in cache-aided wireless networks," *IEEE Trans. Wireless Commun.*, vol. 21, no. 4, pp. 2506–2520, 2021.

[156] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.

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

[158] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key technologies for 5G wireless systems*. Cambridge university press, Apr. 2017.

[159] J. Kim, D. Yu, S.-H. Moon, and S.-H. Park, "Grouped NOMA multicast transmission for F-RAN with wireless fronthaul and edge caching," in *2019 16th International Symposium on Wireless Communication Systems (ISWCS)*. IEEE, Aug. 2019, pp. 145–149.

[160] Y. Fu, H. Wang, and C. W. Sung, "Optimal power allocation for the downlink of cache-aided NOMA systems," in *2018 10th International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, Oct. 2018, pp. 1–6.

[161] Y. Li, H. Zhang, W. Huangfu, K. Long, and J. Liu, "Subchannel assignment and power optimization in caching based UAV networks with NOMA," in *Proc. ICC*. IEEE, Jun. 2020, pp. 1–6.

[162] I. Budhiraja, N. Kumar, S. Tyagi, and S. Tanwar, "Energy consumption minimization scheme for NOMA-based mobile edge computation networks underlaying UAV," *IEEE Syst. J.*, Dec. 2021.

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

[164] "Project loon. accessed: Jul. 15, 2017." [Online]. Available: https://x.company/projects/loon/.

[165] S. Chandrasekharan, K. Gomez, A. Al-Hourani, S. Kandeepan, T. Rasheed, L. Goratti, L. Reynaud, D. Grace, I. Bucaille, T. Wirth, and S. Allsopp, "Designing and implementing future aerial communication networks," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 26–34, May 2016.

[166] Y. Zeng, R. Zhang, and T. J. Lim, "Wireless communications with unmanned aerial vehicles: opportunities and challenges," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 36–42, May 2016.

[167] Y. Su, Y. Liu, Y. Zhou, J. Yuan, H. Cao, and J. Shi, "Broadband LEO satellite communications: Architectures and key technologies," *IEEE Wirel. Commun.*, vol. 26, no. 2, pp. 55–61, Apr. 2019.

[168] N. U. Hassan, C. Huang, C. Yuen, A. Ahmad, and Y. Zhang, "Dense small satellite networks for modern terrestrial communication systems: Benefits, infrastructure, and technologies," *IEEE Wirel. Commun.*, vol. 27, no. 5, pp. 96–103, Oct. 2020.

[169] A. Armon and H. Levy, "Cache satellite distribution systems: modeling, analysis, and efficient operation," *IEEE J. Sel. Areas Commun.*, vol. 22, no. 2, pp. 218–228, 2004.

[170] ——, "Cache satellite distribution systems: modeling and analysis," in *IEEE INFOCOM 2003. Twenty-second Annual Joint Conference of the IEEE Computer and Communications Societies (IEEE Cat. No.03CH37428)*, vol. 1, Apr. 2003, pp. 240–250.

[171] H. Wu, J. Li, H. Lu, and P. Hong, "A two-layer caching model for content delivery services in satellite-terrestrial networks," in *Proc. Globecom*, Dec. 2016, pp. 1–6.

[172] Q. T. Ngo, T. K. Phan, W. Xiang, A. Mahmood, and J. Slay, "Two-tier cache-aided full-duplex hybrid satellite–terrestrial communication networks," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 58, no. 3, pp. 1753–1765, Jun. 2022.

[173] K. An, Y. Li, X. Yan, and T. Liang, "On the performance of cache-enabled hybrid satellite-terrestrial relay networks," *IEEE Wireless Commun. Lett.*, vol. 8, no. 5, pp. 1506–1509, Oct. 2019.

[174] Facebook, "Connecting the world from the sky," Facebook Technical Report, Tech. Rep., 2014.

[175] E. Kulu, "Nanosatellite & cubesat database," [Online]. Available: https://www.nanosats.eu/, 28 Aug. 2021.

[176] X. Zhang, B. Zhang, K. An, B. Zhao, Y. Jia, Z. Chen, and D. Guo, "On the performance of hybrid satellite-terrestrial content delivery networks with non-orthogonal multiple access," *IEEE Wireless Commun. Lett.*, vol. 10, no. 3, pp. 454–458, Mar. 2020.

[177] X. Zhang, B. Zhang, K. An, G. Wu, Y. Jia, S. Qi, and D. Guo, "Noma-based proactive content caching in hybrid satellite-aerial-terrestrial networks," in *Proc. WCNC*. IEEE, Mar. 2021, pp. 1–6.

[178] V. S. et al., "On the performance of cache-free/cache-aided stbc-noma in cognitive hybrid satellite-terrestrial networks," *IEEE Wireless Commun. Lett.*, 2022.

[179] M. Mozaffari, W. Saad, M. Bennis, and M. Debbah, "Wireless communication using unmanned aerial vehicles (UAVs): Optimal transport theory for hover time optimization," *IEEE Trans. Wireless Commun.*, vol. 16, no. 12, pp. 8052–8066, Dec. 2017.

[180] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[181] Y. Yin, M. Liu, G. Gui, H. Gacanin, H. Sari, and F. Adachi, "Cross-layer resource allocation for UAV-assisted wireless caching networks with NOMA," *IEEE Trans. Veh. Technol.*, vol. 70, no. 4, pp. 3428–3438, Apr. 2021.

[182] Z. Wang, T. Zhang, Y. Liu, and W. Xu, "Deep reinforcement learning for caching placement and content delivery in UAV NOMA networks," in *2020 International Conference on Wireless Communications and Signal Processing (WCSP)*. IEEE, Oct. 2020, pp. 406–411.

[183] T. Zhang, Y. Wang, Y. Liu, W. Xu, and A. Nallanathan, "Cache-enabling UAV communications: Network deployment and resource allocation," *IEEE Trans. Wireless Commun.*, vol. 19, no. 11, pp. 7470–7483, Nov. 2020.

[184] Z. Wang, T. Zhang, Y. Liu, and W. Xu, "Caching placement and resource allocation for AR application in UAV NOMA networks," in *Proc. Globecom*. IEEE, Dec. 2020, pp. 1–6.

[185] P. D. Thanh, H. T. H. Giang, and I. Koo, "UAV-assisted NOMA downlink communications based on content caching," in *2020 International Conference on Information and Communication Technology Convergence (ICTC)*, Oct. 2020, pp. 786–791.

[186] H. Dai, L. Zhang, H. Bian, and B. Wang, "UAV relaying assisted transmission optimization with caching in vehicular networks," *Physical Commun.*, p. 101214, Dec. 2020.

[187] Y. Hao, M. Chen, L. Hu, M. S. Hossain, and A. Ghoneim, "Energy efficient task caching and offloading for mobile edge computing," *IEEE Access*, vol. 6, pp. 11 365–11 373, Mar. 2018.

[188] L. N. Huynh, Q.-V. Pham, T. D. Nguyen, M. D. Hossain, Y.-R. Shin, and E.-N. Huh, "Joint computational offloading and data-content caching in NOMA-MEC networks," *IEEE Access*, vol. 9, pp. 12 943–12 954, Jan. 2021.

[189] F. Wang, J. Xu, X. Wang, and S. Cui, "Joint offloading and computing optimization in wireless powered mobile-edge computing systems," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 1784–1797, Mar. 2018.

[190] S. Li, B. Li, and W. Zhao, "Joint optimization of caching and computation in multi-server NOMA-MEC system via reinforcement learning," *IEEE Access*, vol. 8, pp. 112 762–112 771, Jun. 2020.

[191] S. Yu, R. Langar, X. Fu, L. Wang, and Z. Han, "Computation offloading with data caching enhancement for mobile edge computing," *IEEE Trans. Veh. Technol.*, vol. 67, no. 11, pp. 11 098–11 112, Nov. 2018.

[192] Z. Yang, Y. Liu, and Y. Chen, "Distributed reinforcement learning for NOMA-enabled mobile edge computing," in *Proc. ICC*. IEEE, Jun. 2020, pp. 1–6.

[193] L. P. Qian, B. Shi, Y. Wu, B. Sun, and D. H. K. Tsang, "NOMA-enabled mobile edge computing for internet of things via joint communication and computation resource allocations," *IEEE Internet Things J.*, vol. 7, no. 1, pp. 718–733, Jan. 2020.

[194] Z. Zhang, Q. Li, W. Chen, and Z. Hong, "Distributed resource allocation for NOMA-based mobile edge computing with content caching," in *Proc. WCNC*. IEEE, Mar. 2021, pp. 1–6.

[195] Y. Wu, K. Ni, C. Zhang, L. P. Qian, and D. H. K. Tsang, "NOMA-assisted multi-access mobile edge computing: A joint optimization of computation offloading and time allocation," *IEEE Trans. Veh. Technol.*, vol. 67, no. 12, pp. 12 244–12 258, Dec. 2018.

[196] D. Jiang and L. Delgrossi, "IEEE 802.11p: Towards an international standard for wireless access in vehicular environments," in *Proc. VTC (Spring)*, May 2008, pp. 2036–2040.

[197] S. Chen, J. Hu, Y. Shi, and L. Zhao, "LTE-V: A TD-LTE-based V2X solution for future vehicular network," *IEEE Internet Things J.*, vol. 3, no. 6, pp. 997–1005, Dec. 2016.

[198] L. Liang, H. Peng, G. Y. Li, and X. Shen, "Vehicular communications: A physical layer perspective," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10 647–10 659, Dec. 2017.

[199] S. Fang, H. Chen, Z. Khan, and P. Fan, "On the content delivery efficiency of NOMA assisted vehicular communication networks with delay constraints," *IEEE Wireless Commun. Lett.*, vol. 9, no. 6, pp. 847–850, Jun. 2020.

[200] Y. Guo, Q. Yang, F. R. Yu, and V. C. M. Leung, "Cache-enabled adaptive video streaming over vehicular networks: A dynamic approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 6, pp. 5445–5459, Jun. 2018.

[201] B. Di, L. Song, Y. Li, and G. Y. Li, "NOMA-based low-latency and high-reliable broadcast communications for 5G V2X services," in *Proc. Globecom*. IEEE, Dec. 2017, pp. 1–6.

[202] B. Di, L. Song, Y. Li, and Z. Han, "V2X meets NOMA: Non-orthogonal multiple access for 5G-enabled vehicular networks," *IEEE Wirel. Commun.*, vol. 24, no. 6, pp. 14–21, Dec. 2017.

[203] S. Gurugopinath, Y. Al-Hammadi, P. C. Sofotasios, S. Muhaidat, and O. A. Dobre, "Non-orthogonal multiple access with wireless caching for 5G-enabled vehicular networks," *IEEE Netw.*, vol. 34, no. 5, pp. 127–133, Sep. 2020.

[204] J. Zhang, S. Chen, Y. Lin, J. Zheng, B. Ai, and L. Hanzo, "Cell-free massive MIMO: A new next-generation paradigm," *IEEE Access*, vol. 7, pp. 99 878–99 888, Jul. 2019.

[205] Y. Zhang, H. Cao, M. Zhou, and L. Yang, "Spectral efficiency maximization for uplink cell-free massive MIMO-NOMA networks," in *Proc. ICC*, May 2019, pp. 1–6.

[206] Y. Li and G. A. Aruma Baduge, "Noma-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 950–953, Dec. 2018.

[207] M. Bashar, K. Cumanan, A. G. Burr, H. Q. Ngo, L. Hanzo, and P. Xiao, "On the performance of cell-free massive MIMO relying on adaptive NOMA/OMA mode-switching," *IEEE Trans. Commun.*, vol. 68, no. 2, pp. 792–810, Feb. 2020.

[208] ——, "NOMA/OMA mode selection-based cell-free massive MIMO," in *Proc. ICC*, May 2019, pp. 1–6.

[209] A. A. Ohashi, D. B. d. Costa, A. L. P. Fernandes, W. Monteiro, R. Failache, A. M. Cavalcante, and J. C. W. A. Costa, "Cell-free massive MIMO-NOMA systems with imperfect SIC and non-reciprocal channels," *IEEE Wireless Commun. Lett.*, vol. 10, no. 6, pp. 1329–1333, Jun. 2021.

[210] S. Kusaladharma, W. P. Zhu, W. Ajib, and G. Amarasuriya, "Achievable rate analysis of NOMA in cell-free massive MIMO: A stochastic geometry approach," in *Proc. ICC*, May 2019, pp. 1–6.

[211] S. Kusaladharma, W.-P. Zhu, W. Ajib, and G. A. A. Baduge, "Achievable rate characterization of NOMA-aided cell-free massive MIMO with imperfect successive interference cancellation," *IEEE Trans. Commun.*, vol. 69, no. 5, pp. 3054–3066, May 2021.

[212] J. Zhang, E. Björnson, M. Matthaiou, D. W. K. Ng, H. Yang, and D. J. Love, "Prospective multiple antenna technologies for beyond 5G," *IEEE J. Sel. Areas Commun.*, vol. 38, no. 8, pp. 1637–1660, Aug. 2020.

[213] S. Chen, J. Zhang, E. Björnson, S. Wang, C. Xing, and B. Ai, "Wireless caching: Cell-free versus small cells," in *Proc. ICC*, Jun. 2021, pp. 1–6.

[214] M. Bayat, R. K. Mungara, and G. Caire, "Coded caching in a cell-free SIMO network," in *WSA 2018; 22nd International ITG Workshop on Smart Antennas*, Mar. 2018, pp. 1–8.

[215] C. Wang, R. C. Elliott, D. Feng, W. A. Krzymien, S. Zhang, and J. Melzer, "A framework for MEC-enhanced small-cell hetnet with massive MIMO," *IEEE Wirel. Commun.*, vol. 27, no. 4, pp. 64–72, Aug. 2020.

[216] W. Feng, J. Tang, Y. Yu, J. Song, N. Zhao, G. Chen, K.-K. Wong, and J. Chambers, "UAV-enabled SWIPT in IoT networks for emergency communications," *IEEE Wirel. Commun.*, vol. 27, no. 5, pp. 140–147, Oct. 2020.

[217] J. Tang, J. Luo, M. Liu, D. K. C. So, E. Alsusa, G. Chen, K.-K. Wong, and J. A. Chambers, "Energy efficiency optimization for NOMA with SWIPT," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 452–466, Jun. 2019.

[218] W. Wu, X. Yin, P. Deng, T. Guo, and B. Wang, "Transceiver design for downlink SWIPT NOMA systems with cooperative full-duplex relaying," *IEEE Access*, vol. 7, pp. 33 464–33 472, Mar. 2019.

[219] Y. Yuan, Y. Xu, Z. Yang, P. Xu, and Z. Ding, "Energy efficiency optimization in full-duplex user-aided cooperative SWIPT NOMA systems," *IEEE Trans. Commun.*, vol. 67, no. 8, pp. 5753–5767, Aug. 2019.

[220] X. Li, J. Li, and L. Li, "Performance analysis of impaired SWIPT NOMA relaying networks over imperfect weibull channels," *IEEE Syst. J.*, vol. 14, no. 1, pp. 669–672, Mar. 2020.

[221] T. N. Do and B. An, "Optimal sum-throughput analysis for downlink cooperative SWIPT NOMA systems," in *2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications Computing (SigTelCom)*, Jan. 2018, pp. 85–90.

[222] 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.

[223] D. Niyato, D. I. Kim, P. Wang, and L. Song, "A novel caching mechanism for internet of things (IoT) sensing service with energy harvesting," in *Proc. ICC*, May 2016, pp. 1–6.

[224] A. Kumar and W. Saad, "On the tradeoff between energy harvesting and caching in wireless networks," in *Proc. ICC*. IEEE, Sep. 2015, pp. 1976–1981.

[225] S. Zhou, J. Gong, Z. Zhou, W. Chen, and Z. Niu, "Greendelivery: proactive content caching and push with energy-harvesting-based small cells," *IEEE Commun. Mag.*, vol. 53, no. 4, pp. 142–149, Apr. 2015.

[226] H. Li, J. Li, M. Liu, Z. Ding, and F. Gong, "Energy harvesting and resource allocation for cache-enabled UAV based IoT NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 70, no. 9, pp. 9625–9630, Sep. 2021.

[227] X. Zhang, T. Lv, Y. Ren, and Z. Lin, "Joint content push and transmission in NOMA with SWIPT caching helper," *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 922–925, Apr. 2020.

[228] X. Wang, L. Kong, F. Kong, F. Qiu, M. Xia, S. Arnon, and G. Chen, "Millimeter wave communication: A comprehensive survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 3, pp. 1616–1653, Thirdquarte 2018.

[229] T. S. Rappaport, S. Sun, R. Mayzus, H. Zhao, Y. Azar, K. Wang, G. N. Wong, J. K. Schulz, M. Samimi, and F. Gutierrez, "Millimeter wave mobile communications for 5G cellular: It will work!" *IEEE Access*, vol. 1, pp. 335–349, 2013.

[230] L. Zhu, Z. Xiao, X.-G. Xia, and D. Oliver Wu, "Millimeter-wave communications with non-orthogonal multiple access for B5G/6G," *IEEE Access*, vol. 7, pp. 116 123–116 132, 2019.

[231] S. A. R. Naqvi and S. A. Hassan, "Combining NOMA and mmWave technology for cellular communication," in *Proc. VTC (Fall)*, 2016, pp. 1–5.

[232] D. Zhang, Z. Zhou, C. Xu, Y. Zhang, J. Rodriguez, and T. Sato, "Capacity analysis of NOMA with mmWave massive MIMO systems," *IEEE J. Sel. Areas Commun.*, vol. 35, no. 7, pp. 1606–1618, Jul. 2017.

[233] 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.

[234] J. Li, X. Jing, Y. Zhang, and J. Mu, "Performance analysis of agile-beam NOMA in millimeter wave networks," *IEEE Access*, vol. 8, pp. 6638–6649, 2020.

[235] J. Li, X. Li, A. Wang, and N. Ye, "Performance analysis for downlink MIMO-NOMA in millimeter wave cellular network with D2D communications," *Wireless Communications and Mobile Computing*, vol. 2019, p. 1914762, Jun. 2019.

[236] Y. Tian, G. Pan, and M.-S. Alouini, "On NOMA-based mmwave communications," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 15 398–15 411, Dec. 2020.

[237] S. A. Busari, K. M. S. Huq, S. Mumtaz, L. Dai, and J. Rodriguez, "Millimeter-wave massive MIMO communication for future wireless systems: A survey," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 836–869, Secondquarter 2018.

[238] Y. Li and G. A. Aruma Baduge, "NOMA-aided cell-free massive MIMO systems," *IEEE Wireless Commun. Lett.*, vol. 7, no. 6, pp. 950–953, Dec. 2018.

[239] J. Ghosh, V. Sharma, H. Haci, S. Singh, and I.-H. Ra, "Performance investigation of NOMA versus OMA techniques for mmwave massive MIMO communications," *IEEE Access*, vol. 9, pp. 125 300–125 308, Aug. 2021.

[240] M. R. G. Aghdam, B. M. Tazehkand, and R. Abdolee, "On the performance analysis of mmWave MIMO-NOMA transmission scheme," *IEEE Trans. Veh. Technol.*, vol. 69, no. 10, pp. 11 491–11 500, Oct. 2020.

[241] Y. Song, W. Yang, X. Yang, Z. Xiang, and B. Wang, "Physical layer security in cognitive millimeter wave networks," *IEEE Access*, vol. 7, pp. 109 162–109 180, 2019.

[242] R. K. Saha, "Underlay cognitive radio millimeter-wave spectrum access for in-building dense small cells in multi-operator environments toward 6G," in *2020 23rd International Symposium on Wireless Personal Multimedia Communications (WPMC)*, 2020, pp. 1–6.

[243] Y. Song, W. Yang, Z. Xiang, N. Sha, H. Wang, and Y. Yang, "An analysis on secure millimeter wave NOMA communications in cognitive radio networks," *IEEE Access*, vol. 8, pp. 78 965–78 978, 2020.

[244] Y. Song, W. Yang, Z. Xiang, B. Wang, and Y. Cai, "Secure transmission in mmWave NOMA networks with cognitive power allocation," *IEEE Access*, vol. 7, pp. 76 104–76 119, 2019.

[245] O. Semiari, W. Saad, M. Bennis, and B. Maham, "Caching meets millimeter wave communications for enhanced mobility management in 5G networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 2, pp. 779–793, Feb. 2018.

[246] H. Elayan, O. Amin, R. M. Shubair, and M.-S. Alouini, "Terahertz communication: The opportunities of wireless technology beyond 5G," in *2018 International Conference on Advanced Communication Technologies and Networking (CommNet)*, 2018, pp. 1–5.

[247] A. Moldovan, M. A. Ruder, I. F. Akyildiz, and W. H. Gerstacker, "LOS and NLOS channel modeling for terahertz wireless communication with scattered rays," in *Proc. Globecom*, 2014, pp. 388–392.

[248] O. Ulgen, S. Erkucuk, and T. Baykas, "Non-orthogonal multiple access for terahertz communication networks," in *2020 11th IEEE Annual Ubiquitous Computing, Electronics and Mobile Communication Conference (UEMCON)*, 2020, pp. 0737–0742.

[249] S. B. Melhem and H. Tabassum, "User pairing and outage analysis in multi-carrier NOMA-THz networks," 2022. [Online]. Available: https://arxiv.org/abs/2201.09357

[250] A. Magbool, H. Sarieddeen, N. Kouzayha, M.-S. Alouini, and T. Y. Al-Naffouri, "Terahertz-band non-orthogonal multiple access: System- and link-level considerations," 2021. [Online]. Available: https://arxiv.org/abs/2111.01412

[251] S. R. Sabuj, A. M. S. Khan, and M. Hamamura, "Application of non-orthogonal multiple access for machine type communication in sub-terahertz band," *Computer Networks*, vol. 182, p. 107508, 2020. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1389128620311737

[252] H. Zhang, H. Zhang, W. liu, K. long, J. Dong, and V. C. M. Leung, "Energy efficient user clustering and hybrid precoding for terahertz MIMO-NOMA systems," in *Proc. ICC*, 2020, pp. 1–5.

[253] B. Zheng, Q. Wu, and R. Zhang, "Intelligent reflecting surface-assisted multiple access with user pairing: NOMA or OMA?" *IEEE Commun. Lett.*, vol. 24, no. 4, pp. 753–757, Apr. 2020.

[254] S. Gong, X. Lu, D. T. Hoang, D. Niyato, L. Shu, D. I. Kim, and Y.-C. Liang, "Toward smart wireless communications via intelligent reflecting surfaces: A contemporary survey," *IEEE Commun. Surveys Tuts.*, vol. 22, no. 4, pp. 2283–2314, Fourthquarter 2020.

[255] Z. Ding, R. Schober, and H. V. Poor, "On the impact of phase shifting designs on IRS-NOMA," *IEEE Wireless Commun. Lett.*, vol. 9, no. 10, pp. 1596–1600, Oct. 2020.

[256] Z. Ding and H. Vincent Poor, "A simple design of IRS-NOMA transmission," *IEEE Commun. Lett.*, vol. 24, no. 5, pp. 1119–1123, May 2020.

[257] F. Fang, Y. Xu, Q.-V. Pham, and Z. Ding, "Energy-efficient design of IRS-NOMA networks," *IEEE Trans. Veh. Technol.*, vol. 69, no. 11, pp. 14 088–14 092, Nov. 2020.

[258] M. Zeng, X. Li, G. Li, W. Hao, and O. A. Dobre, "Sum rate maximization for IRS-assisted uplink NOMA," *IEEE Commun. Lett.*, vol. 25, no. 1, pp. 234–238, Jan. 2021.

[259] X. Mu, Y. Liu, L. Guo, J. Lin, and N. Al-Dhahir, "Exploiting intelligent reflecting surfaces in NOMA networks: Joint beamforming optimization," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6884–6898, Oct 2020.

[260] Y. Chen, M. Wen, E. Basar, Y.-C. Wu, L. Wang, and W. Liu, "Exploiting reconfigurable intelligent surfaces in edge caching: Joint hybrid beamforming and content placement optimization," *IEEE Trans. Wireless Commun.*, vol. 20, no. 12, pp. 7799–7812, Dec. 2021.

[261] J. Suh, O. Aboul-Magd, J. Jia, and E. Au. (2018, Sep.) SOMA for EHT. Doc: IEEE 802.11-18/1462r0. Huawei. [Online]. Available: https://mentor.ieee.org/802.11/dcn/18/11-18-1462-00-0eht-soma-for-eht.pptx (accessed Nov. 03, 2021).

[262] E. Khorov, A. Kureev, I. Levitsky, and I. F. Akyildiz, "Prototyping and experimental study of non-orthogonal multiple access in Wi-Fi networks," *IEEE Netw.*, vol. 34, no. 4, pp. 210–217, Jul. 2020.

[263] ATSC A/322:2021, "ATSC standard: Physical layer protocol," ATSC 3.0, Tech. Rep., Jan. 2021.

[264] 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.

[265] T. Ramírez, C. Mosquera, N. Noels, M. Caus, J. Bas, L. Blanco, and N. Alagha, "Study on the application of NOMA techniques for heterogeneous satellite terminals," in *10th Advanced Satellite Multimedia Systems Conference and the 16th Signal Processing for Space Communications Workshop (ASMS/SPSC)*, Oct. 2020, pp. 1–8.

[266] P. Vanichchanunt, P. La-aiddee, P. Sasithong, and S. Paripurana, "Implementation of non-orthogonal multiple access on DVB-T using software-defined radio," in *36th International Technical Conference on Circuits/Systems, Computers and Communications (ITC-CSCC)*, Jun. 2021, pp. 1–4.

[267] H. Shariatzadeh, S. Ghazi-Maghrebi, and B. Karakaya, "An improving performance cellular DTV broadcasting with hybrid non-orthogonal LDM and orthogonal eMBMS configuration," *Array*, vol. 11, no. 100073, pp. 1–14, Sep. 2021.

[268] A. Benjebbour, A. Li, K. Saito, Y. Saito, Y. Kishiyama, and T. Nakamura, "NOMA: From concept to standardization," in *IEEE conference on standards for communications and networking (CSCN)*, Oct. 2015, pp. 18–23.

[269] A. Benjebbour, "An overview of non-orthogonal multiple access," *ZTE Commun.*, vol. 15, no. S1, pp. 21–30, Jun. 2017.

[270] 3GPP TR 36.859 V13.0.0, "Study on downlink multiuser superposition transmission (MUST) for LTE," 3GPP, Tech. Rep., Dec. 2015.

[271] 3GPP TR 38.812 V16.0.0, "Study on non-orthogonal multiple access (NOMA) for NR," 3GPP, Tech. Rep., Dec. 2018.

[272] Y. Chen, A. Bayesteh, Y. Wu, B. Ren, S. Kang, S. Sun, Q. Xiong, C. Qian, B. Yu, Z. Ding, S. Wang, S. Han, X. Hou, H. Lin, R. Visoz, and R. Razavi, "Toward the standardization of non-orthogonal multiple access for next generation wireless networks," *IEEE Commun. Mag.*, vol. 56, no. 3, pp. 19–27, Mar. 2018.

[273] Y. Yuan, Z. Yuan, and L. Tian, "5G non-orthogonal multiple access study in 3GPP," *IEEE Commun. Mag.*, vol. 58, no. 7, pp. 90–96, Jul. 2020.

[274] A. C. Cirik, N. M. Balasubramanya, L. Lampe, G. Vos, and S. Bennett, "Toward the standardization of grant-free operation and the associated NOMA strategies in 3GPP," *IEEE Commun. Stand. Mag.*, vol. 3, no. 4, pp. 60–66, Dec. 2019.

[275] A. G. Perotti and B. M. Popović, "Non-orthogonal multiple access for degraded broadcast channels: RA-CEMA," in *Proc. WCNC*, Mar. 2015, pp. 735–740.

[276] Y. Yuan, S. Wang, Y. Wu, H. V. Poor, Z. Ding, X. You, and L. Hanzo, "NOMA for next-generation massive IoT: Performance potential and technology directions," *IEEE Commun. Mag.*, pp. 1–7, 2021.

[277] L. Yu, Z. Liu, M. Wen, D. Cai, S. Dang, Y. Wang, and P. Xiao, "Sparse code multiple access for 6G wireless communication networks: Recent advances and future directions," *IEEE Commun. Stand. Mag.*, vol. 5, no. 2, pp. 92–99, Jun. 2021.

[278] S. M. Riazul Islam, M. Zeng, and O. A. Dobre. (2017, Jun.) NOMA in 5G systems: Exciting possibilities for enhancing spectral efficiency. IEEE 5G Tech Focus. IEEE Future Networks. [Online]. Available: https://futurenetworks.ieee.org/tech-focus/june-2017/noma-in-5g-systems (accessed Nov. 03, 2021).

[279] DOCOMO 5G Whitepaper, "5G radio access: Requirements, concepts and technologies," NTT DOCOMO, Tech. Rep., Jul. 2014.

[280] A. Benjebbour, Y. Kishiyama, and Y. Okumura, "Field trials of improving spectral efficiency by using a smartphone-sized NOMA chipset," *NTT DOCOMO Tech. J.*, vol. 20, no. 1, pp. 4–13, Jul. 2018.

[281] Z. Ding. (2016, Apr.) Non-orthogonal multiple access (NOMA): Evolution towards 5G cellular networks. Lecture Slides. School of Computing and Communications, Lancaster University. [Online]. Available: https://www.lancaster.ac.uk/staff/dingz/NOMA.pdf (accessed Nov. 10, 2021).

[282] SK Telecom 5G Whitepaper, "SK Telecom's view on 5G vision, architecture, technology, and spectrum," SK Telecom, Tech. Rep., Oct. 2014.

[283] Qualcomm. (2018, Sep.) Expanding the 5G NR ecosystem: 5G NR roadmap in 3GPP Release 16 and beyond. [Online]. Available: https://www.qualcomm.com/media/documents/files/expanding-the-5g-nr-ecosystem-and-roadmap-in-3gpp-rel-16-beyond.pdf (accessed Nov. 10, 2021).

[284] H. Lee, H. Ko, K. Choi, K. Noh, D. Kim, and S. Lee, "Method for transmitting and receiving terminal grouping information in non-orthogonal multiple access scheme," U.S. Patent 10 595 325, Mar. 17, 2020.

[285] M. Fuse and K. Shioiri, "Non-orthogonal multiple access and massive MIMO for improved spectrum efficiency," Anritsu, Tech. Rep. 91, Mar. 2016.

[286] Y. Ma, Z. Yuan, W. Li, and Z. Li, "Lightweight and instant access technologies and protocols to boost digital transformations," in *ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K)*, Dec. 2020, pp. 1–5.

[287] 3GPP RP-190711, "Revised work item proposal: 2-step RACH for NR," 3GPP, Tech. Rep., Mar. 2019.

[288] P. Stoica, J. Li, and Y. Xie, "On probing signal design for MIMO radar," *IEEE Trans. Signal Process.*, vol. 55, no. 8, pp. 4151–4161, Aug. 2007.

[289] X. Mu, Y. Liu, L. Guo, J. Lin, and L. Hanzo, "NOMA-aided joint radar and multicast-unicast communication systems," *IEEE J. Sel. Areas Commun.*, Mar. 2022.

[290] Z. Wang, Y. Liu, X. Mu, Z. Ding, and O. A. Dobre, "NOMA empowered integrated sensing and communication," *IEEE Commun. Lett.*, vol. 26, no. 3, pp. 677–681, Mar. 2022.

[291] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications," *IEEE Trans. Wireless Commun.*, vol. 21, no. 5, pp. 3083–3098, May 2022.

[292] Y. Liu, X. Liu, X. Gao, X. Mu, X. Zhou, O. A. Dobre, and H. V. Poor, "Robotic communications for 5g and beyond: Challenges and research opportunities," *IEEE Commun. Mag.*, vol. 59, no. 10, pp. 92–98, Oct. 2021.

[293] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, "Intelligent reflecting surface enhanced indoor robot path planning: A radio map-based approach," *IEEE Trans. Wireless Commun.*, vol. 20, no. 7, pp. 4732–4747, Jul. 2021.

[294] P. Raviteja, Y. Hong, E. Viterbo, and E. Biglieri, "Practical pulse-shaping waveforms for reduced-cyclic-prefix OTFS," *IEEE Trans. Veh. Technol.*, vol. 68, no. 1, pp. 957–961, Jan. 2019.

[295] Z. Ding, R. Schober, P. Fan, and H. V. Poor, "OTFS-NOMA: an efficient approach for exploiting heterogenous user mobility profiles," *IEEE Trans. Commun.*, vol. 67, no. 11, pp. 7950–7965, Nov. 2019.

[296] X. Zhang, L. Yang, Z. Ding, J. Song, Y. Zhai, and D. Zhang, "Sparse vector coding-based multi-carrier NOMA for in-home health networks," *IEEE J. Sel. Areas Commun.*, vol. 39, no. 2, pp. 325–337, Feb. 2021.

[297] 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.

[298] C. Chen, W.-D. Zhong, H. Yang, and P. Du, "On the performance of MIMO-NOMA-based visible light communication systems," *IEEE Photonics J.*, vol. 30, no. 4, pp. 307–310, Feb. 2018.

[299] M. V. Jamali and H. Mahdavifar, "Uplink non-orthogonal multiple access over mixed RF-FSO systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 5, pp. 3558–3574, May 2020.

Dipen Bepari received his B-Tech degree in Electronics and Communication Engineering from Jalpaiguri Government Engineering College, West Bengal, India, and completed M.Tech. from National Institute of Technology, Durgapur, India. He has received Ph.D. degree from Department of Electronics Engineering, IIT (ISM), Dhanbad, India. He received scholarship from the University Grant Commission (UGC), Government of India for the period 2008–2010 during M.Tech., and from the Ministry of Human Resource and Development (MHRD), Government of India for the period 2012–2017 during Ph.D. Presently he is working at National Institute of Technology, Raipur, India. His research interests include cognitive radio networks, wireless sensor networks, energy harvesting, and NOMA technique.

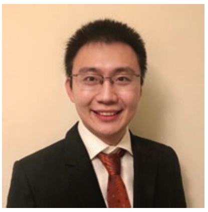
Soumen Mondal (S'16-S'20) received his B.Tech degree in Electronics and Communication Engineering in 2008 from Haldia Institute of Technology, Haldia, India, M.Tech. degree in Telecommunication Engineering in 2010, and Ph.D. degree in 2021 from National Institute of Technology, Durgapur, India. Dr. Mondal has published about 15 research papers in refereed journals. His research interests include Cognitive Radio Networks, Energy Harvesting, Intelligent Reflecting Surface, FSO and NOMA.

Aniruddha Chandra (M'08–SM'16) received BE, ME, and PhD degrees from Jadavpur University, Kolkata, India, in 2003, 2005 and 2011, respectively.

He joined the Electronics and Communication Engineering Department, National Institute of Technology, Durgapur, India, in 2005. He is currently an Associate Professor there. In 2011, he was a Visiting Lecturer at the Asian Institute of Technology, Bangkok. From 2014 to 2016, he worked as a Marie Curie fellow at Brno University of Technology, Czech Republic. In 2019, he worked as a Visiting Researcher at the Slovak University of Technology, Slovakia. In 2022, he was a Guest Researcher at Niigata University, Japan.

Dr. Chandra has published about 130 research papers in refereed journals and peer-reviewed conferences. He is a co-recipient of the best short paper award at IEEE VNC 2014, held in Paderborn, Germany, and delivered a keynote lecture at IEEE MNCApps 2012, held in Bangalore, India. He is currently the secretary of the IEEE P2982 Standard working group and IEEE ComSoc RCC SIG on Propagation Channels for 5G and Beyond. His primary area of research is physical layer issues in wireless communication.

Rajeev Shukla (S'20) received his BE degree in Electronics and Telecommunication Engineering from Rungta College of Engineering and Technology in 2011 and ME degree in Communication Engineering from Chhattrapati Shivaji Institute of Engineering and Technology in 2015. He is currently pursuing his PhD degree in Wireless Communication from National Institute of Technology, Durgapur, India. He has worked as an Assistant Professor from 2012 to 2015 in School of Engineering, MATS University, Raipur and from 2016 to 2018 in Shekhawati Institute of Engineering and Technology, Sikar. From 2018 to 2020, he worked as a Junior Research Fellow at Indian Institute of Technology, Jodhpur, where he worked on a project on SAR image processing. His research interests include 5G, millimeter waves, channel modelling, and cognitive radio.



Yuanwei Liu (S'13–M'16–SM'19, http://www.eecs.qmul.ac.uk/ yuanwei) received the B.S. and M.S. degrees from the Beijing University of Posts and Telecommunications in 2011 and 2014, respectively, and the PhD degree in electrical engineering from the Queen Mary University of London, U.K., in 2016. He was with the Department of Informatics, King's College London, from 2016 to 2017, where he was a Post-Doctoral Research Fellow. He has been a Senior Lecturer (Associate Professor) with the School of Electronic Engineering and Computer Science, Queen Mary University of London, since Aug. 2021, where he was a Lecturer (Assistant Professor) from 2017 to 2021. His research interests include non-orthogonal multiple access, 5G/6G networks, RIS, integrated sensing and communications, and machine learning.

Yuanwei Liu is a Web of Science Highly Cited Researcher 2021. He is currently a Senior Editor of IEEE Communications Letters, an Editor of the IEEE Transactions on Wireless Communications and the IEEE Transactions on Communications. He serves as the leading Guest Editor for IEEE JSAC special issue on Next Generation Multiple Access, a Guest Editor for IEEE JSTSP special issue on Signal Processing Advances for Non-Orthogonal Multiple Access in Next Generation Wireless Networks. He received IEEE ComSoc Outstanding Young Researcher Award for EMEA in 2020. He received the 2020 IEEE Signal Processing and Computing for Communications (SPCC) Technical Early Achievement Award, IEEE Communication Theory Technical Committee (CTTC) 2021 Early Achievement Award. He received IEEE ComSoc Young Professional Outstanding Nominee Award in 2021. He has served as the Publicity Co-Chair for VTC 2019-Fall. He is the leading contributor for "Best Readings for Non-Orthogonal Multiple Access (NOMA)" and the primary contributor for "Best Readings for Reconfigurable Intelligent Surfaces (RIS)". He serves as the chair of Special Interest Group (SIG) in SPCC Technical Committee on the topic of signal processing Techniques for next generation multiple access (NGMA), the vice-chair of SIG Wireless Communications Technical Committee (WTC) on the topic of Reconfigurable Intelligent Surfaces for Smart Radio Environments (RISE), and the Tutorials and Invited Presentations Officer for Reconfigurable Intelligent Surfaces Emerging Technology Initiative.

Mohsen Guizani (M'89–SM'99–F'09) received the BS (with distinction), MS and PhD degrees in Electrical and Computer engineering from Syracuse University, Syracuse, NY, USA in 1985, 1987 and 1990, respectively. He is currently a Professor of Machine Learning and the Associate Provost at Mohamed Bin Zayed University of Artificial Intelligence (MBZUAI), Abu Dhabi, UAE. Previously, he worked in different institutions in the USA. His research interests include applied machine learning and artificial intelligence, Internet of Things (IoT), intelligent autonomous systems, smart city, and cybersecurity. He was elevated to the IEEE Fellow in 2009 and was listed as a Clarivate Analytics Highly Cited Researcher in Computer Science in 2019, 2020 and 2021. Dr. Guizani has won several research awards including the "2015 IEEE Communications Society Best Survey Paper Award", the Best ComSoc Journal Paper Award in 2021 as well five Best Paper Awards from ICC and Globecom Conferences. He is the author of ten books and more than 800 publications. He is also the recipient of the 2017 IEEE Communications Society Wireless Technical Committee (WTC) Recognition Award, the 2018 AdHoc Technical Committee Recognition Award, and the 2019 IEEE Communications and Information Security Technical Recognition (CISTC) Award. He served as the Editor-in-Chief of IEEE Network and is currently serving on the Editorial Boards of many IEEE Transactions and Magazines. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer.

Arumugam Nallanathan (S'97–-M'00–-SM'05–-F'17) has been a Professor of Wireless Communications and the Head of the Communication Systems Research Group, School of Electronic Engineering and Computer Science, Queen Mary University of London, since September 2017. He was with the Department of Informatics, King's College London from December 2007 to August 2017, where he was a Professor of Wireless Communications from April 2013 to August 2017 and a Visiting Professor from September 2017. He was an Assistant Professor with the Department of Electrical and Computer Engineering, National University of Singapore from August 2000 to December 2007. He published nearly 500 technical papers in scientific journals and international conferences. His research interests include artificial intelligence for wireless systems, B5G wireless networks, Internet of Things, and molecular communications.

He is a co-recipient of the Best Paper Awards presented at the IEEE Communications Society SPCE Outstanding Service Award 2012, the IEEE Communications Society RCC Outstanding Service Award 2014, the IEEE International Conference on Communications in 2016, IEEE Global Communications Conference 2017, and the IEEE Vehicular Technology Conference in 2018. He served as the Chair for the Signal Processing and Communication Electronics Technical Committee of IEEE Communications Society and Technical Program Chair and member of Technical Program Committees in numerous IEEE conferences. He is an Editor-at-Large for IEEE TRANSACTIONS ON COMMUNICATIONS and a Senior Editor for IEEE WIRELESS COMMUNICATIONS LETTERS. He was an Editor for IEEE TRANSACTIONS ON WIRELESS COMMUNICATIONS from 2006 to 2011, IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY from 2006 to 2017, and IEEE SIGNAL PROCESSING LETTERS. He has been selected as a Web of Science Highly Cited Researcher in 2016 and an AI 2000 Internet of Things Most Influential Scholar in 2020. He is an IEEE Distinguished Lecturer.