

# A Survey on Cache-Aided NOMA for 6G Networks

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 sixth generation (6G) network requirements, 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 6G 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 cache-aided NOMA technology, we summarize the performance goals of cache-aided NOMA systems, present the associated design challenges, and categorize related recent 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

This work was partly supported by Core Research Grant (CRG), Science and Engineering Research Board, Department of Science and Technology, Government of India, Grant No. CRG/2018/000175.

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: 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).

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 back-haul latency since the introduction of 3G [10]. Let us review the importance of caching from the 6G latency requirement perspective. Electromagnetic (EM) waves travel at  $3 \times 10^5$  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 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. 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 RDC 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 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 the 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 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 [11]. 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 [12]. 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 [13].

The cache technique is greatly compatible with many advanced communication systems, like millimeter-wave (mmWave) communications [14], [15], multiple-input multiple-output (MIMO) systems [16], [17], Mobile edge computing (MEC) [18], [19], teraHertz communication [20] 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 [21], [22]. It is capable of efficiently realizing higher system throughput and spectral efficiency compared to the traditional orthogonal multiple access (OMA) [23]. 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 combining the signals of multiple users and successive interference cancellation (SIC) method at the receiver for decoding individual signal [24]–[26]. A fundamental concept of the NOMA technique and its application in long-term evolution (LTE) and 5G have been reported in [27], [28]. It is also reported that the amalgamation of NOMA with cache technology can achieve a significant

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

Tech.	Year	Ref.	Primary Focus to survey
NOMA	2016	[29]	• Recent innovations • Performance analysis • Resource allocation • Associated challenges and solutions • NOMA applications • Integration with other technologies • Error rate performance of NOMA • Future research directions
	2017	[25]	
	2018	[24], [30]	
	2019	[31]	
	2020	[32], [33]	
	2021	[34]	
	2022	[35]	
Cache	2009	[36]	• Contributions of existing caching
	2012	[37]	• Explores cache-aided network types
	2013	[38], [39]	• Research challenges and solutions
	2018	[40], [41]	• Content placement & delivery strategy • Future research scope
	2020	[42]	The existing studies on green caching
	2020	[43]	The edge caching in cellular network
Cache-aided NOMA	This paper		Design principles, challenges, key features, and diverse practical applications of cache-aided NOMA systems.

system performance enhancement. A cache-aided NOMA network reaps benefits from both the cache and NOMA techniques.

#### A. Motivation

A few survey papers on cache strategy and NOMA principles available in the literature are shown in Table I. Along with the state-of-the-art of NOMA techniques for future 5G systems, the fundamental operating principles of NOMA and their comparative performance analysis over the OMA were the primary focus of Linglong *et al.* [24]. The interplay between NOMA and other technologies such as MIMO, massive MIMO, mmWave communications, cognitive communications, visible light communications, wireless caching, etc. have been presented, and how the combination of NOMA and these technologies elevate network performance such as scalability, spectral efficiency, energy efficiency, etc. summarized in [31]. But a systematic interplay between NOMA and wireless caching was completely overlooked. The survey paper [32] analyses various optimization scenarios to investigate the maximum achievable sum-rate when power domain NOMA amalgamates with other promising technologies for 5G and beyond 5G (B5G). However, analysis of other important performance metrics was ignored. A detailed analysis of wireless networks combining NOMA and cache technologies was not presented in [31], [32]. Ding *et al.* provide a broader overview of recent research on NOMA and their applications along with research challenges in various enabling technology for advanced communications [25]. The recent survey paper analyses the error rate performance of the NOMA in a holistic manner [35]. However, applications of cache-aided NOMA were not discussed. The authors present a comparative study of various methods (such as optimization techniques, analytical methods, game theory, matching theory, graph theory, and machine learning techniques) involved to address the problem of resource allocation, signaling, practical implementation, and security aspects of NOMA technologies in [34].

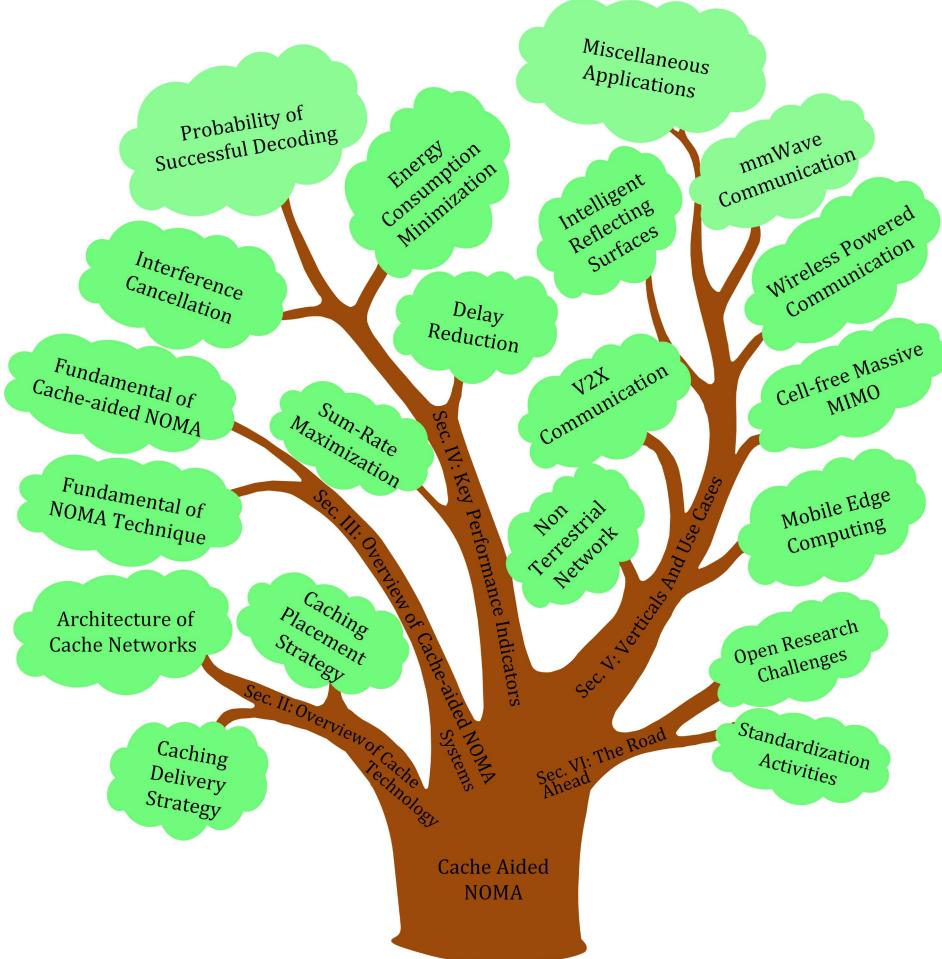


Fig. 1: Organization of this survey.

In [42], 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 [41], Liying *et al.* present 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 study the caching technique in cellular systems [40], [44]. In [40], the authors present the research challenges of cache-aided integrated networking in wireless communication systems. The survey paper [45] focuses on caching techniques for vehicular communication. Although individual surveys on NOMA and caching do exist [24], [41], [42], explicit analysis of cache-aided NOMA systems and their applications have not been reported yet.

### 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 6G 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 6G 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, 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 the list of acronyms used in this survey paper tabulated in Table II.

TABLE II: List of acronyms frequently used in this survey

Acronyms	Description
AAT	Average Access Time
AP	Access point
B5G	Beyond Fifth-Generation
CIC	Cache-enabled interference cancellation
CFmMIMO	Cell-free Massive Multi-Input Multiple-Output
CSI	Channel State Information
D2D	Device-to-Device
eMBB	Enhanced Mobile Broadband
ICN	Information Centric Networking
IoT	Internet of Thing
LFU	Least Frequently Used
LRU	Least Recently Used
LSTM	Long-Short-Term Memory
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
SCMA	Sparse Code Multiple Access
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
VLC	Visible Light Communications

## II. OVERVIEW OF CACHE TECHNOLOGY

The types of demand 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, implementation

of a Web caching technique becomes a promising approach for significantly reducing traffic load [46], [47]. In the early 2000s, primary reason for network congestion was due to the demand for video content, and that was challenged by caching for content distribution networking [37] and information centric networking (ICN) [38]. In ICN, the distance between cache and user was further shortened, and a new feature, *popularity of contents*, was incorporated in the content placement techniques [39]. Now-a-days, users' created video files, and their delivery creating additional network traffic load over wireless channels. Establishing a wireless caching network is more challenging than a wired caching network, because of unpredictable movement of the users and uncertain channel gain quality. Since 2009, the researchers are showing their interest in wireless caching for reducing traffic load, and quality of communication [36], [48]. In 2010, the authors confirmed that the caching technique enhances the system throughput by as much as 400–500% [49]. 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 are combining cache technology with other technologies, and analysing their performances in various wireless networks. The major challenges of wireless caching are given below.

- 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 gain.
- Limited battery life restricts increase of transmit power.
- Limited cache memory demands efficient cache placement and access strategies.
- The mobile devices do not know if cache contents are updated unless it is informed by the network.

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

### 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 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 [50]. The popularity of the randomly requested contents is widely modelled by the Zipf distribution [51]–[53].  $P(f_l)$ , the popularity of  $l$ th requested file  $f_l$  is expressed as

$$P(f_l) = \frac{l^{-\gamma}}{\sum_{f=1}^F p^{-\gamma}}, \quad 1 \leq l \leq F \quad (1)$$

where  $\gamma$  is the exponent of the Zipf distribution that expresses the reuse probability of a requested file, and  $F$  is the total number of files. The widely used Zipf distribution is proficient for measuring the polarity of video files [52]. In [54], based on the varying degree of popularity, the authors have considered multiple levels of non-uniform content popularity in their research work.

Two types of content placement strategy broadly found in literature- *coded placement strategy* [55]–[59] and *uncoded placement strategy* [51], [60]–[62]. 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 [63] and fountain codes [64] 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.

Let a cache-aided system with  $K$  users connected with server through a shared link, and individually can request any one file from the server which has stored  $N$  files of equal size. The cache memory, accessible by each user, can store a maximum  $F$  files. For the uncoded cache placement strategy,  $U_u$ , the load of the shared link is expressed as  $U_u = K(1 - F/N)$ . On the other hand,  $U_c$ , the load of shared link for the coded placement strategy is expressed as [56]

$$U_c = K \left( 1 - \frac{F}{N} \right) \frac{1}{1 + \frac{KF}{N}} \quad (2)$$

The coded caching achieves a caching gain of  $\frac{1}{1 + \frac{KF}{N}}$ . 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 [55]. In [65], 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 authors in [66], explore the advantages of coded caching strategies when cache-enabled access points (APs) like BS, SBS, MBS, etc.) are randomly distributed. The

coded caching strategies exploit coded multicast opportunities that further decrease the backhaul traffic, particularly when cache-enabled APs are densely deployed [67]. Based on the random caching placement and multiple groupcast index coding, the authors have proposed an order-optimal coded caching placement [68]. Niesen *et al.* [69] 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 [70]. Jinbei *et al.* [67] present 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 that defines a process to replace/update the cached content when the cache memory is full. 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 [46], [71]–[73]. 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. For example, combinations of PER with LRU make a PER-LRU caching method. Various cache access and placement strategies and their performances are analyzed in [74]. A gain-based cache replacement policy named, Min-SAUD [75] and a hotspot-based caching scheme [76] are adopted to satisfy the caching replacement requirements. In [77], 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. The delivery strategy decides on a suitable transmission process so that files arrive at the requested user successfully and quickly. In addition, caching delivery strategy determines where the requested file will be transmitted from, the transmitting frequency band, the transmission power, and the encoding process. 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.

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 every time 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. Hennessy *et al.* [87] have measured the advantages of caching using the average access time (AAT) metric as given below.

TABLE III: Caching strategies and performance metrics.

Ref.	Network Model	Proposed Method	Performance Metric	Merit
[78]	Information Centric Networking (ICN)	Proposed Universal Caching algorithm based on discrete time Markov Chain model	Cache hits, access delay and cost of the link	Distance from content source, frequency of fetching the content, number of outgoing links have been considered in the algorithm
[79]	ICN	Proposed cache replacement scheme- Popularity Prediction Cache	Cache hits, access delay and throughput	Capable to predict future popularity and caches the video
[76]	ICN in a mobile ad-hoc network	Proposed caching strategy based on request probability and transition probability analysis, called cache rebalancing	Cache efficiency	Location-sensitive contents are given weightage in the caching placement
[80]	Broadcast network with uncoded caching	Proposed centralized joint cache and channel coding strategy	Minimum required transmit power	Utilizes user's local caches and exploits correlation among the contents in the database
[75]	Wireless network	Proposed a gain-based cache replacement policy named as Min-SAUD	Cache hit ratio, access delay	Considered cost of cache validation in the cache replacement policy
[81]	Cellular network with edge caching	Proposed two online prediction methods popularity prediction model (PPM) and Grassmannian prediction model (GPM) to forecast content popularity	Average successful decoding probability and mean squared error	Can predict the future content popularity
[82]	D2D mobile Network	Proposed maximum distance separable (MDS) coded edge caching	Minimize the network load	Considered device mobility, caching devices distribution and cache size
[83]	D2D mobile Network	Proposed mobility-aware greedy coded caching strategy	Reduce the backhaul traffic	Considered pattern of user mobility
[84]	Data networking	Proposed a cache placement strategy based content popularity and node popularity	Cache hit ratio, start latency, and link load	Distribution of consumers in different regions has been incorporated in the caching method
[85]	Heterogeneous wireless network with edge caching	Investigated the proactive caching strategy for layered video streaming	Average download delay	Cache strategies designed based on download delay model
[77]	heterogeneous cellular system	Resource allocation for concurrent caching replacement and content delivery	minimizes power and and resource utilization	optimal solutions for replacement and content delivery consumes less power
[86]	Heterogeneous D2D network	Proposed optimal cooperative content caching and delivery scheme	Minimize delivery latency, cache hit rate	Considered cache storage capacity and bandwidth capacity constraint

$$AAT = \text{hit time} + (\text{miss rate} \times \text{miss penalty}) \quad (3)$$

where *hit time* indicates when the cache memory provides the requested file, and *miss penalty* is the time that takes to access the internet/cloud.

In [88] 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 [89]. The backhaul traffic load and content delivery latency of cache-aided networks are subjected to the cache memory size. The fundamental trade-off between the advantage of caching and the cache storage capacity has been studied in [58], [90] for coded and uncoded cache systems. They have analyzed the trade-off from an information-theoretic perspective, and the investigation is carried out 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 [91] have optimized the cache memory size subjected to channel statis-

tics, backhaul capacity, and distribution of file popularity in a cellular network.

Table III presents various types of caching strategies and metrics adopted to evaluate their performance. Various types of caching strategies like Markov Chain model-based universal caching [78], maximum distance separable (MDS) coded edge caching [82], mobility-aware greedy coded caching [83], proactive caching [85], and cooperative content caching [86] for NOMA systems have been proposed, and the performance of these methods is analyzed mostly in terms of a cache hit rate. Efficient caching techniques improve systems performance by reducing the cost of the link [78], content access delay [79], required transmit power [80], backhaul traffic load [82], [83], delivery latency [86], and enhancing successful decoding probability [81]. To deal with the sudden change in the popularity of the content and replace the cached content, the online content popularity prediction techniques like the popularity prediction model (PPM) and Grassmannian prediction model (GPM) [81], and popularity prediction cache model [79], [92] are proposed.

### 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

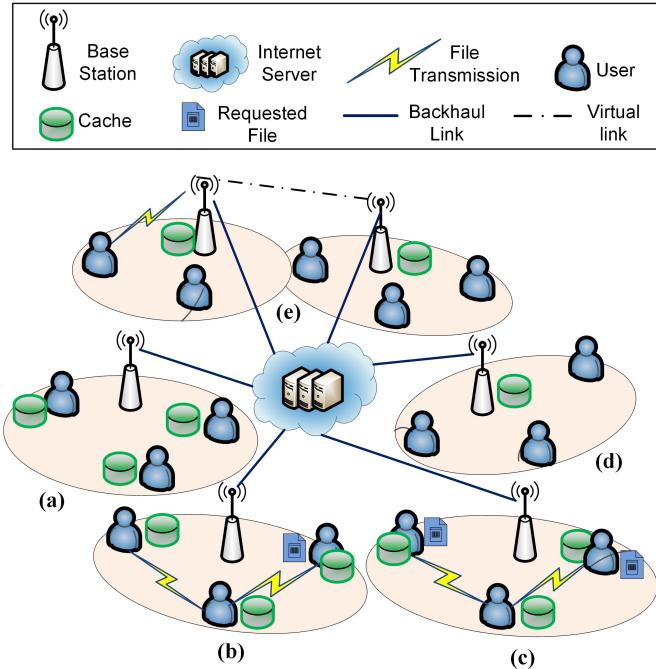


Fig. 2: Various architecture of cache networks (a) D2D caching technique, (b) D2D multihop relay, (c) Cooperative D2D, (d) edge caching technique, and (e) Cooperative edge caching technique

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 [83], [93], [94]. 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 this 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 [95], [96].

*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 [97], 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 [98]–[100], 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 contents. 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 own BS. Figure 2(e) shows the cooperative BSs delivery. Shan *et al.*, in [101] 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 hybrid wireless caching networks, where cache infrastructure is available at both the transmitter (BS, SBS, and access points) and receiver (user) end [58], [102]. 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 [51]. 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 a 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 found as the most common approach for modeling the popularity of randomly requested content. The popularity of the content, distribution of content popularity, correlation among the contents, location of users, and finite storage availability constraint are required to be taken into account in the caching strategy, which introduces challenges in the caching strategies. Furthermore, a wide range of variety in the content and sudden change of their popularity intensify difficulties and show the drawback of content placement during *off-hours*. Two online prediction methods named the popularity prediction model (PPM) and Grassmannian prediction model (GPM) have been proposed to predict the popularity in advance [81]. 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 algorithm [103]. Although coded 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 [55]. Most of the researchers assume error-free channel during content pushing, which is questionable in real time communication scenarios. Therefore,

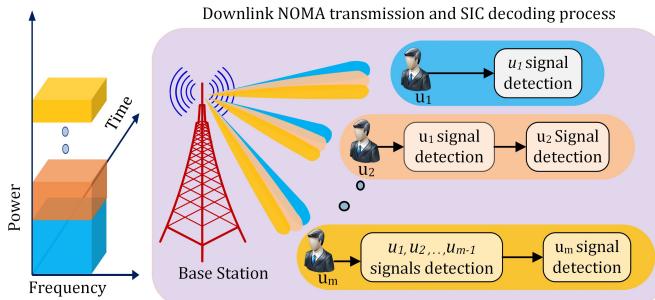


Fig. 3: Power distribution strategy for PD-NOMA. The transmitter allocates maximum power to user  $u_1$  with the weakest channel and minimum power to user  $u_m$  with the strongest channel.  $u_1$  decodes its own signal directly, and  $u_m$  first decodes signal of  $u_1, u_2, \dots, u_{m-1}$  then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

more practical channels need 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 NOMA operation and challenges associated with the cache-aided NOMA.

#### A. Fundamental of 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, the resources like time, frequency, and code are orthogonal which reduces 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 [104].

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. In PD-NOMA, a transmitter combines signals of multiple users with different power levels by applying superposition coding. Unlike the water-filling algorithm for power allocation in OMA [105], [106], 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) stronger user achieves a higher data transmission rate with low transmit power, and (ii) 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 [107], [108]. At the receiver end, the receiver applies the SIC 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 [109]. The power allocation and decoding process of PD-NOMA is pictorially shown in Fig. 3.

The CD-NOMA assigns different codes to users and superimposes over the same time and frequency. The multiuser shared access (MUSA) [110], sparse code multiple access (SCMA) [111], low density spreading (LDS) [112] are the main three types of CD-NOMA. Though the CD-NOMA can remarkably enhance the 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 up-gradation to the present communication networks nor a high transmission bandwidth [113]. In addition to it, PD-NOMA has a low complexity system compared to CD-NOMA from a design perspective. This survey focuses on PD-NOMA, and NOMA indicates a PD-NOMA.

It is necessary to know the order of average channel gains or instantaneous channel state information (CSI) for fixed power allocation [114] or dynamic channel allocation [115] respectively. The users' ordering is accomplished mostly based on the instantaneous value of CSI [109], [116]. It is to note that a wrong user ordering leads to an incorrect choice of power distribution, which may lead to a situation where a few users are always in outage [117]. Therefore, it is necessary to acquire the knowledge of perfect channel gain, which is much more challenging, particularly when users are moving. 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 [118]–[121]. A second-order statistics (SOS)-based CSI model achieves superior system performance than the imperfect CSI based model [122].

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 IV-E). The imperfect SIC is widely modeled as Gaussian distribution [123], [124]. However, there may be some typical scenarios where an error does not obey the Gaussian distribution [125]. 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 SCI on the performance of the weakest user.

#### B. Fundamental of Cache-aided NOMA

The cache-aided NOMA has been established as an advanced communication concept for next-generation communications. According to International Mobile Telecommunications (IMT) [126], 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% [127]). The OMA techniques cannot meet the above requirements. NOMA technique efficiently improves downlink and uplink spectral-efficient by 30% and 100% respectively in eMBB compared to OMA [128]. NOMA-supported mMTC and URLLC applications can serve 5 and 9 times more users, respectively [127]. In the OMA technique, users with better channel conditions get higher priority, and the users with poorer channel conditions need to wait for access; that initiates problems of fairness and high latency. On the other hand, NOMA serves multiple users with various channel gains simultaneously, which provides improved fairness with lower latency [27]. The superiority of NOMA over OMA has motivated researchers to select the NOMA-based cache strategy.

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?

In comparison to 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, BS pushes only the content with 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, OMA-based content placement requires sophisticated methods which efficiently schedule the files based on popularity [129]. Unlike OMA, during a single time slot, applying the NOMA principle, 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, NOMA serves multiple users simultaneously. However, two-user NOMA downlink transmission is proposed for the LTE system [130]. The cache-aided NOMA scheme efficiently improves the cache hit probability and reduces the delivery outage probability compared to conventional OMA-based caching.

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 to cite fu2019mode. The SIC decoding process of a cache-aided NOMA network is slightly different from conventional NOMA

[131]. To understand this, consider a D2D cache-aided network with two users  $u_1$  and  $u_2$ , but they are not neighbors. Let,  $u_1$  requested file  $f_1$  is cached at  $u_2$  but  $u_2$  requested file  $f_2$  is not cached. BS transmits a signal comprising both the  $f_1$  and  $f_2$ . As  $f_1$  is already available in the associate cache of  $u_2$ , without applying SIC,  $u_2$  can remove the  $f_1$  from the superimposed signal and recover own  $f_2$ . Notably, caching helps in the decoding process even when the requested file is not cached (detail discussed in Sec IV-C).

A few research works focus on the cache placement issues of NOMA systems. In [132], 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 [133]. 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 [88]. It is observed that a cache-aided network conventionally stores cache contents during an off-peak time [60]–[62], [88], [132], [133] which may not be an efficient approach, particularly when a network frequently needs to update the content of the cache. Ding *et al.* proposed two algorithms, *push-then-deliver strategy* and the *push-and-deliver strategy* algorithms, that are capable of efficiently storing cache contents during on-peak hours in D2D networks [129], [134]. 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.

**Summary:-** Cache-aided NOMA individually helps each other. NOMA helps in content delivery maintaining the fairness of users. On the other hand, caching helps NOMA in the SIC decoding process even when requested files are not cached. Optimal cache content placement strategies in the NOMA system significantly reduce outage probability [129], transmitting power [133], content delivery latency [88], and network cost [132]. However, the problems of cache content placement optimization in NOMA systems have not been explored enough yet. The conventional cache strategies push contents into the cache storage unit 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 is a promising approach that can push multiple contents within a short duration. This feature of NOMA has made it the best candidate for content pushing during peak-time [129]. Push-then-deliver and the push-and-deliver strategies are capable of content placement during on-peak hours [129]. In the previous section, we notice PPM and GPM can forecast content popularity [81]. A combination of content popularity prediction method and online content placement strategy could be a breakthrough approach for next-generation communication systems.

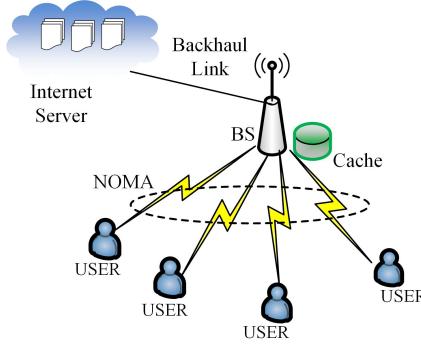


Fig. 4: Basic cache-aided NOMA network, where BS serves multiple users simultaneously other than time and frequency domain, and cache reduces use of backhaul link.

#### IV. KEY PERFORMANCE INDICATORS

This section presents fundamental analysis of key performance indicators (KPIs) of cache-aided NOMA networks. It reviewed the various techniques formulated to enhance the KPIs of systems.

##### A. Sum-Rate Maximization

Sum-Rate measures the successful data transmission rate over the communication channel of unit bandwidth. 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 [22], [135]. Hence, NOMA is exploited to maintain user fairness [25], [28], [136], [137]. 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 [108]. Wireless caching strategy is an efficient approach implemented to increase the sum-rate of 5G networks [138].

In [139], 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 [140]. The performance of the proposed network paradigm was evaluated in terms of the sum rate. Xinyue *et al.* [141] 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 the requested files depends on various network resources like transmit power, available bandwidth, cache status, etc. To analyse the delay reduction technique, consider a cache-aided NOMA network, as shown in

Fig.4, where a cache-enabled BS provides  $I$  contents to  $K$  users. The index sets  $\mathcal{K} = \{1, 2, \dots, K\}$  and  $\mathcal{I} = \{1, 2, \dots, I\}$  are used to denote the indices of user and content respectively. Let,  $L_i, i \in \mathcal{I}$  is the file size of  $i$ th file. BS uses backhaul links during content placement and when users' requested files are not cached. Transmission delay of the cache-aided NOMA system is associated with the delay of both backhaul link and BS-to-user link.

**Transmission Delay of Backhaul Link-** Transmission over the backhaul link is subjected to 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 data transmission rate of backhaul link,  $\mathcal{T}_B$ , transmission time required for un-cached content is given by

$$\mathcal{T}_B = \sum_{i=1}^I (1 - C_i) \frac{L_i}{R_B} \quad (4)$$

**Transmission Delay of BS-to-User link-** The BS delivers requested files either from cache or after downloading from the internet server to the users by NOMA. Considering  $R_k^D$  as the data transmission rate over the BS-to-user link, the delivery delay of the  $k$ th user is  $\mathcal{T}_k = L_i / R_k^D, \forall k \in \mathcal{K}$ . Considering,  $L_{i,k}$  as the size of the  $i$ th file requested by the  $k$ th user,  $\mathcal{T}_{max}$ , the maximum time that the BS takes to deliver is given by

$$\mathcal{T}_{max} = \max \left\{ \frac{L_{i,1}}{R_1^D}, \frac{L_{i,2}}{R_2^D}, \dots, \frac{L_{i,K}}{R_K^D} \right\} \quad (5)$$

End-to-end transmission delay is  $\mathcal{T}_k^t = \mathcal{T}_B + \mathcal{T}_k, k \in \mathcal{K}$ , and maximum network delay is  $\mathcal{T}_{max}^t = \mathcal{T}_B + \mathcal{T}_{max}$ .

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 [142]. In [142], 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 [143]. 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 [145], 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 [89], 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 [89].

##### C. Interference Cancellation

The interference cancellation capability is another feature of cache-aided NOMA systems. During the off-peak time, the BS stores popular content in the cache using the split file caching

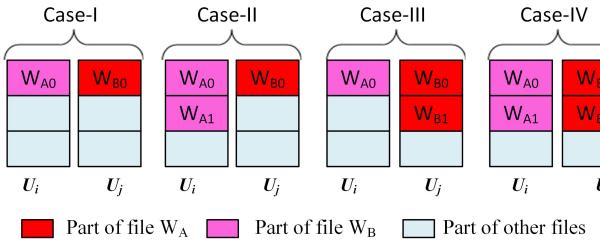


Fig. 5: Different cache status of  $U_i$  and  $U_j$  requested file  $W_{fn}$ ,  $f \in \{A, B\}$ ,  $n \in \{0, 1, 2\}$ . Case-I is the scenario when each user caches a smaller portion of the requested file. This unfavorable cache status can not be avoided in practice. Case-IV is favorable status as each user cache a maximum portion of the requested files.

technique, where each file is split into multiple segments and only a few portions are cached. Perfect knowledge of the content in the cache is available to the BS. A user gets a few portions of the requested file from the associated cache memory and the rest of the portions from the BS. Cache-enabled interference cancellation (CIC) helps BS to remove a few segments from the requested file, which are available in the cache. 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, and CIC eliminates the segments from the superposed signal [88]. As a consequence, the interference power is reduced.

To understand the CIC, consider two cache-enabled users  $U_i$  and  $U_j$  requesting files  $W_A$  and  $W_B$  respectively. The split caching technique has been considered, where files are divided into three segments and sequentially stored fully or partially. The segments are denoted by  $W_{fn}$ ,  $f \in \{A, B\}$ ,  $n \in \{0, 1, 2\}$ . The cache status  $\underline{k}_f$  and  $\bar{k}_f$ ,  $f \in \{A, B\}$  denotes that maximum and minimum portions of the file  $W_f$  is cached respectively. Based on the cache configurations four possible cases:

**Case-I:**  $i = \underline{k}_A$  and  $j = \bar{k}_B$ ,

**Case-II:**  $i = \bar{k}_A$  and  $j = \underline{k}_B$ ,

**Case-III:**  $i = \underline{k}_A$  and  $j = \bar{k}_B$ ,

**Case-IV:**  $i = \bar{k}_A$  and  $j = \bar{k}_B$ .

The file partition and cache status for all the four cache configurations are shown in Fig. 5. Xiang *et al.* have considered only case-I [144] and all the four cases in [88]. Considering all the cache configurations into account,  $x$ , the BS transmitted signals is given by

$$x = \begin{cases} \sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} \\ \quad + \sqrt{p_{j,2}}x_{B2}, & \text{Case-I,} \\ \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} + \sqrt{p_{j,2}}x_{B2}, & \text{Case-II,} \\ \sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}, & \text{Case-III,} \\ \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1}, & \text{Case-IV.} \end{cases} \quad (6)$$

where  $x_{fn}$  is the codeword corresponding to subfile  $W_{fn}$ ,  $f \in \{A, B\}$ ,  $n \in \{0, 1, 2\}$ , and  $p_{kn}$ ,  $k \in \{i, j\}$ ,  $n \in \{0, 1, 2\}$  is the transmit power to  $x_{fn}$ . As the BS has the perfect knowledge of cached content, thus BS does not transmit those subfiles.

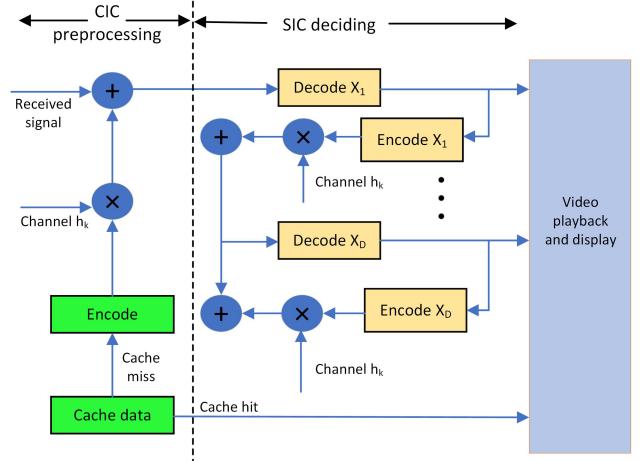


Fig. 6: Joint CIC and SIC decoding technique. The cached portions are removed from the received signal by the CIC processing. The residual signals  $X_1, X_2, \dots, X_D$  which are not cached but sequentially decoded by applying traditional SIC process.

The joint CIC and SIC decoding process is employed at the receiver end. The block diagram of the joint CIC and SIC technique is shown in Fig 6. This technique performs the CIC process before applying SCI decoding. The signal after CIC processing is expressed as

$$y_i^{CIC} = \begin{cases} h_i (\sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} \\ \quad + z_i), & \text{Case-I \& III,} \\ h_i (\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B2}) \\ \quad + z_i, & \text{Case-II \& IV} \end{cases} \quad (7)$$

$$y_j^{CIC} = \begin{cases} h_j (\sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,1}}x_{B1} + \sqrt{p_{j,2}}x_{B2}) \\ \quad + z_j, & \text{Case-I \& II,} \\ h_j (\sqrt{p_{i,1}}x_{A1} + \sqrt{p_{i,2}}x_{A2} + \sqrt{p_{j,2}}x_{B2}) \\ \quad + z_j, & \text{Case-III \& IV} \end{cases} \quad (8)$$

It is to note that the BS does not transmit the file segments that are available in the cache of the requested users, and the receiver using the CIC process discards the file segments (all/a few) belong to other users if those are available in its cache. The caching technique not only reduces traffic load but also helps in the SIC decoding technique. The joint CIC and SIC decoding process in cache-aided NOMA systems significantly increases the sum rate for downlink transmission and reduces the file delivery times [88], [144]. Employing CIC, 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 [140] that enhances the sum rate of the systems.

#### D. 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 [133].

In [99], 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 [146], 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 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 [147]. In [148], 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 [149]. The authors in [150] studied resource allocation for enhancing the energy efficiency by optimizing subchannel allocation and power allocation in a NOMA hierarchical network.

#### E. Probability of Successful Decoding

In NOMA, a receiver applies the 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 their 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 successful decoding probability. To understand the decoding process, consider a simple edge cache-aided NOMA system with  $K$  users,  $\mathcal{K} \in \{1, 2, \dots, K\}$  and a cache-enabled BS, as shown in Fig.4. The channel coefficient of BS-to- $k$ th 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 successful decoding probability of a cache-aided NOMA network rather to illustrate the impact of cache 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 cache. Let,  $\mathcal{S}_1^{(1)}$  is the SINR of the signal for  $u_1$  at  $u_1$ . Being a stronger user,  $u_2$  first decodes the signal of  $u_1$  then decodes own signal using SIC. Consider,  $\mathcal{S}_1^{(2)}$  and  $\mathcal{S}_2^{(2)}$  are the SINR

of the  $u_1$ 's signal at  $u_2$  and  $u_2$ 's signal at  $u_2$  respectively. Now,  $\mathcal{D}^{(1)}$ , the overall successful decoding probability of the system for scenario-1 can be given as

$$\mathcal{D}^{(1)} = \mathcal{P}_r \left( \mathcal{S}_1^{(1)} \geq \gamma \right) \mathcal{P}_r \left( \min(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}) \geq \gamma \right) \quad (9)$$

where  $\gamma$  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 (9) are the successful decoding probability of  $f_1$  and  $f_2$  respectively.

*Scenario-2:* In this scenario,  $f_2$  is not cached. BS need 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 successful decoding probability for the scenario-2 can be expressed as

$$\mathcal{D}^{(2)} = \mathcal{P}_r \left( \mathcal{S}_1^{(1)} \geq \gamma \right) \mathcal{P}_r \left( \min(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}, \mathcal{S}_{BS}^{(2)}) \geq \gamma \right) \quad (10)$$

The first term and second term of (10) are the successful decoding probability of  $f_1$  and  $f_2$  respectively.

*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 successful decoding probability for the scenario-3 is given by

$$\mathcal{D}^{(3)} = \mathcal{P}_r \left( \min(\mathcal{S}_1^{(1)}, \mathcal{S}_{BS}^{(1)}) \geq \gamma \right) \mathcal{P}_r \left( \min(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}) \geq \gamma \right) \quad (11)$$

*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 successful decoding probability is expressed as

$$\begin{aligned} \mathcal{D}^{(4)} = & \mathcal{P}_r \left( \min(\mathcal{S}_1^{(1)}, \mathcal{S}_{BS}^{(1)}) \geq \gamma \right) \\ & \mathcal{P}_r \left( \min(\mathcal{S}_1^{(2)}, \mathcal{S}_2^{(2)}, \mathcal{S}_{BS}^{(1)}, \mathcal{S}_{BS}^{(2)}) \geq \gamma \right) \end{aligned} \quad (12)$$

The successful decoding probability of cache-aided NOMA systems is much better than that of NOMA and cache-aided OMA systems [131], [151]. The optimal power distribution is the most popular approach for enriching the successful decoding probability. The authors formulated optimal power distribution problems over the Rayleigh fading channel [131], Weibull, Nakagami-m, and Rician downlink fading channels [151] in a cache-aided cellular network. 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 [152]. Doan *et al.*

in [153] propose divide-and-conquer-based and deep-learning-based power allocation methods to maximize the successful decoding probability 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. *Cache Hit Probability-* The hit rate is an important parameter that tells how professionally popular files are selected for placing in the cache memory. It is a ratio of the number of times cache successfully delivers requested files and the total number of times users request files. The hit rate primarily used to evaluate the efficiency of cache placement techniques. A successful cache hit reduces communication costs by avoiding use of the backhaul link, and also reduces outage probability significantly.

*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 [41]. 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.

## V. VERTICALS AND USE CASES

In this section, our focus is to review the various practical application scenarios of cache-aided NOMA networks. The challenges associated with these applications and their solutions are discussed.

### A. Non Terrestrial Network

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 form the third layer beyond 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<sup>2</sup> area. Various airborne platforms such as Balloons [154], Helikites [155], and UAVs [156] are recently emerging as potential approaches to meet wireless traffic demands, specifically for mobile users. Based

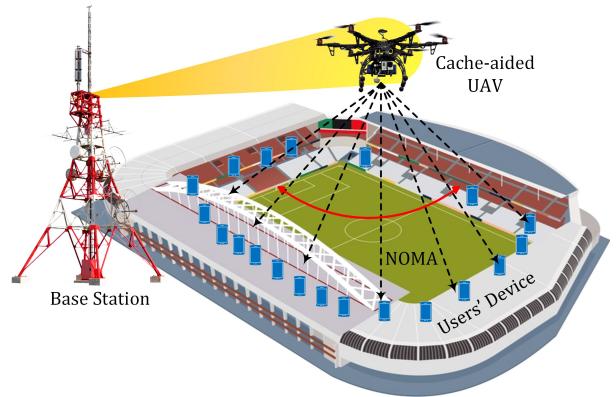


Fig. 7: One of the popular application scenarios of cache-aided NOMA-based UAV-assisted communication systems. The UAV was deployed temporarily to assist the BS to meet thousands of users' requests.

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 of 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) [157] and can meet the latency requirement of Internet of Things (IoT), smart grid, and vehicular communication applications [158]. 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 [159], [160]. 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 [161], [162]. 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 [163]. Google Loon project is one of the industry projects where Google have installed Internet-delivery drone for providing global massive connectivity [164]. The promising research application for 5G communications are as follows: establishing temporal communication infrastructure, MEC for IoT devices, traffic offloading for dense cellular systems, and so on. 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 [165]. In LEOComm, the NOMA has been identified as a potential candidate to provide enhanced spectral efficiency, system capacity [166]–

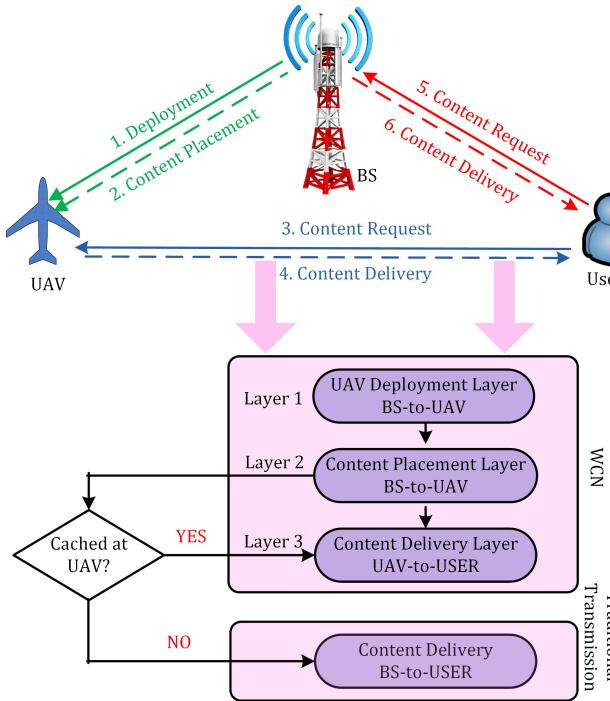


Fig. 8: Cache content delivery strategy and different layers of a UAV-assisted wireless caching system. User requests for desire file to BS only when the file is not cached.

[168]. However, the cache-aided NOMA technique has not been explored in satellite networks and could be a promising research domain.

UAVs deployed in the atmospheric orbit are the most popular and efficient commercial approach 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 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 [169], [170]. 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. 7, where a large number of mobile users in a hot-spot area are under the coverage of a ground macro base station (MBS). The MBS 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 [171], as shown in Fig. 8. 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 [171]. A deep reinforcement learning (DRL) algorithm is proposed for content placement and delivering in a cache-enabling UAV-assisted cellular network [172]. 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 [171] or via UAV [173].

The resource allocation in UAV with cache-aided NOMA system has been studied in [173]–[176]. In [173], 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 [174], 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 [175] 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 [176] 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 [174] [175].

In [173] and [175], 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 [177], 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 [172], 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 the consumption of total energy in [149].

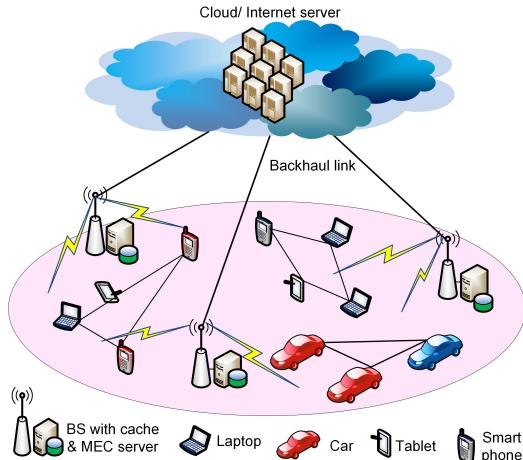


Fig. 9: Application scenario of cache-aided MEC networks. Users are connected and cooperatively sharing files. Devices off-load their computational task to the MEC-cache-aided BS.

### 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 [178]. 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. A simple cache-aided MEC network is depicted in Fig.9.

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. Hao *et al.* have studied the challenges related to the joint optimization of task caching and offloading in [178]. The Authors have proposed a long-short-term memory (LSTM) algorithm for predicting the task popularity of a cache-aided MEC system [99]. Based on the popularity of the historical tasks, the LSTM 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 [179]. Depending on the predicted popularity of the task, a multi-agent Deep-Q-network (MADQN) algorithm was applied to deal with the problem associated with the caching and offloading. A

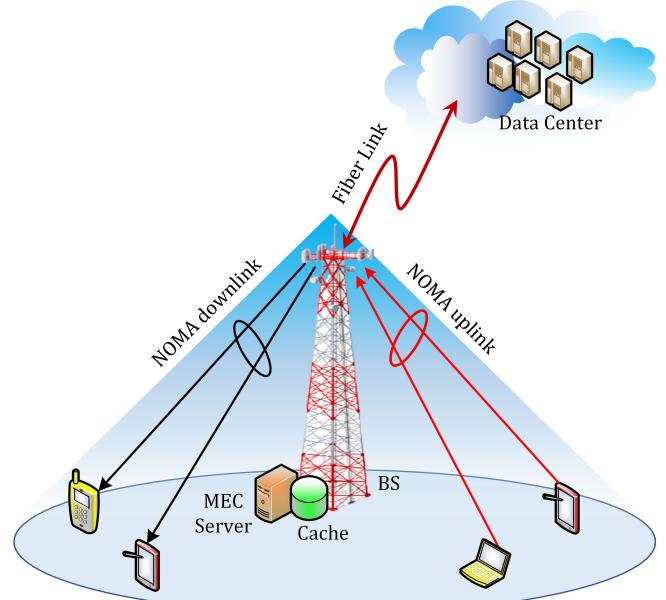


Fig. 10: Basic architecture of a cache-aided MEC network. User's devices off-load computational task to cache and MEC-aided BS.

new collaborative task offloading scheme proposed in [180] is capable of reducing task execution delay up to 42.83% a for single-user caching-enhancement scheme.

To explain the total task completion latency, we consider a simple edge caching NOMA-MEC network with a BS,  $N$  mobile users and a remote cloud, as shown in Fig. 10. The BS is equipped with a MEC server with finite storage capability. Let,  $L_n$  is the size of 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 BS-to- $n$ th user link respectively.  $\mathcal{T}_n^{Lo}$ , the total computational latency for executing task locally includes backhaul latency ( $\mathcal{T}_n^{BL} = L_n/R_n^{BH}$ ), downlink latency ( $\mathcal{T}_n^{DL} = L_n/R_n^{DL}$ ) and local processing time ( $\mathcal{T}_n^{LP} = W_n/f_n^l$ ) which is given by

$$\mathcal{T}_n^{Lo} = (1 - C_n)\mathcal{T}_n^{BH} + \mathcal{T}_n^{DL} + \mathcal{T}_n^{LP} \quad (13)$$

where  $C_n$  is the status of the cache content,  $C_n = 1$  if requested content is available in the cache else it is 0.

The primary objective of MEC is to offload the computational task to the BS as much as possible for remote execution. The total computational latency for edge offloading includes the uplink transmission time, processing time of the MEC server, and backhaul delay. If  $U_n$  is the data of the offloaded task to BS by  $n$ th user and  $R_n^{UL}$  average uplink data transmission rate from  $n$ th user-to-BS, the uplink transmission time 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 MEC server for executing the computational task. Now, the total edge offloading latency i.e., the time needs to execute the offloaded task expressed as

$$\mathcal{T}_n^{EO} = (1 - C_n)\mathcal{T}_n^{BH} + \mathcal{T}_n^{UL} + \mathcal{T}_n^{EP} \quad (14)$$

The  $n$ th user experiences both local task execution latency and edge offloading latency. Let,  $a_n \in \{0, 1\}$  is offloading decision taken by  $n$ th user.  $a_n = 1$  when the user is offloading its computational task to the BS and  $a_n = 0$  otherwise. Hence, the total task completion latency is expressed as

$$\mathcal{T}_n = (1 - a_n)\mathcal{T}_n^{Lo} + \mathcal{T}_n^{EO} \quad (15)$$

Various types of resource allocation methods for cache-aided MEC with NOMA are found in the literature [99], [149], [181]–[183]. 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 [99]. In [181] 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 [149], 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 [182]. The delivery-aware cache placement strategy (DACPSS) 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 [182]. 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 [183].

The authors in [184] 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 [185], the authors aimed to reduce the total completion latency for all users of a cache-aided NOMA-MEC. In [185], the authors formulate 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 [186] 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 onboard information, and entertainment facilities. The IEEE 802.11p technology-based communication for vehicular ad hoc network (VANET) provides 6 - 27 Mbps data rate for a short distance communication [187]. LTE-based vehicle-to-vehicle (V2V) communication supported by the Third-Generation Partnership Project (3GPP) also emerges as an

efficient approach [188]. 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 [189]. In [190], 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 [191]. The NOMA technique is well recognized in vehicular communication for the capability to handle massive connectivity and outperforms the traditional OMA-based system [192]. The authors in [193] investigated the spectral efficiency and resource allocation of a NOMA-based vehicular system. The caching technique showed its proficiency to reduce backhaul overhead traffic [194].

In V2V communication, cache-enabled vehicles stores 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 implied to transmit both the 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 do 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 [194] in 2019 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 [195]. A hybrid multicast/unicast scheme has been investigated in cache-aided NOMA-based vehicular networks [141]. 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 [133] 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 [196]. 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 [197]. 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 [198]. To deal with this an adaptive NOMA/OMA mode-switching method is proposed in [199], [200]. 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 [201]. 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 [202], [203].

CFmMIMO is viewed as one of the arisen technologies for 5G and B5G communications due to its uniform service quality, robust diversity, and interference management ability [196], [204]. 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.* [205] 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 [206]. 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 [207]. 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.* [205] 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 [208]. 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 [209]. Wu *et al.* designed a transceiver for NOMA-based SWIPT-enabled cooperative full-duplex relaying systems [210]. 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 [211]. 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 [212]–[214].

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 [215], 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 (SBS) has been investigated in [216] 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 [217], 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) [218]. The authors have proposed a joint content push and transmission scheme in a cache-enabled SWIPT-based relying network [219]. 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 [218]. 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 [220]. 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 [220], [221].

While dealing with mmWave communications, one of the drawbacks is multiple access because of high power consumption and costly hardware [222]. 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 [223]. In [224], 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 [225]. [226] 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 [227]. All the before-mentioned literature used the Nakagami-m Fading model to represent small-scale fading, whereas [228] 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 [222] 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 [229]. 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 [229]. For a large number of users scenario, NOMA can further reduce the problem of pilot contamination using power domain NOMA and SIC [230]. NOMA also reduces the outage probability and improves spectral efficiency and energy efficiency of the mmWave massive MIMO systems [231]. 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 [232]. 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 [233], [234]. 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. [235], [236] has discussed the security aspects

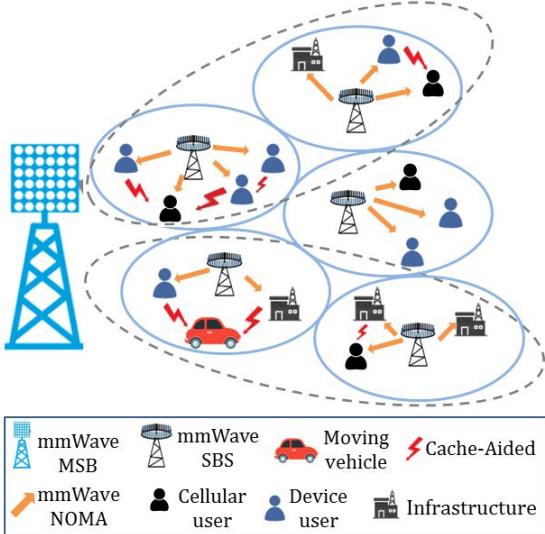


Fig. 11: Cache-Aided NOMA for mmWave.

of NOMA-based Cognitive mmWave Networks, but other aspects are still unexplored.

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 [237], 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. 11 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.

#### G. Intelligent Reflecting Surfaces

Intelligent Reflecting Surfaces (IRS) is an intelligent meta-surface 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 [238], [239]. 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 [240], 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 [241], energy efficiency [242], sum rate [243], [244]. Ding *et al.* proposed an IRS-assisted NOMA transmission such that the network can serve more users compared with spatial division multiple access [241]. Article [245] 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.

#### H. Miscellaneous Applications

Apart from the above-mentioned applications, a few applications of cache-aided NOMA are also found. Haijun *et al.* propose a cache-aided NOMA-based Fog-computing radio access network (F-RAN) architecture for 5G networks and study the power and subchannel allocation problems to achieve high net utility [246]. A joint cache content popularity prediction and access mode selection problem are formulated as the Stackelberg game in cache-aided NOMA-based F-RANs [247]. In [248], 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. The caching techniques are implemented in MIMO-based wireless networks for canceling interference [249], maximization of energy efficiency [20], reducing communication costs [250], [251], and improve transmission reliability [251]. The caching is also applied to counter the severe fading and impulsive noise in MIMO-based power line communications [251]. The outage probability of a cache-aided NOMA MIMO network is derived in [252]. The cache-aided NOMA recently implemented in hybrid satellite-aerial-terrestrial networks (HSATNs) [253] [254] and derived the outage probability and cache hit probability [253].

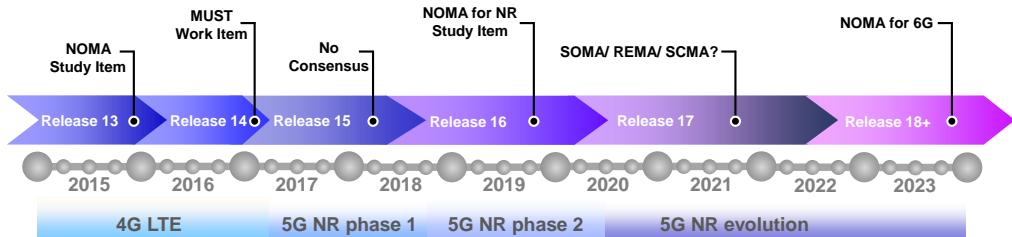


Fig. 12: Timeline of 3GPP releases and different NOMA related standardization activities. NOMA was included in one of the work items in release 16. In release 13 and release 16, discussions were limited to study items only.

## VI. THE ROAD AHEAD

Cache-aided NOMA networks are constantly evolving and is getting diversified with advancement in other domains. In this section we would 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 till date, 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) [255] inspired a NOMA prototype based on software defined radio (SDR) for Wi-Fi [256].

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) [257]. LDM is a two-layer technique [258]. Unlike OMA, both 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 [259] and for terrestrial DVB-T [260]. A comprehensive account of NOMA based broadcast services may be found in Table 2 of the recent article by Shariatzadeh *et al.* [261].

Within the 3rd generation Partnership Project (3GPP), NOMA was first included in LTE Release 13 as a study item (SI) [262] and later, in Release 14, NOMA was included as a work item (WI) [263] under the name of multi-user superposition transmission (MUST) [264]. 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 had been included in Release 16 SI [265]. A complete summary of NOMA based standardization activities within 3GPP is available in the article by Chen *et al.* [266] and later by Yuan *et al.* [267]. On the other hand, the article by Cirik *et al.* [268] 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) [269].

Despite the interest, NOMA was not adopted in any of the work items for 5G NR, as seen in Fig. 12. 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 [270]. However, as pointed out earlier, some variations of contention-based grant-free NOMA is being proposed for 6G. In [271], 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 company whitepapers [272]. The first in the league is NTT DOCOMO, which published a series of technical documents [273], [274]. Huawei, in addition to NOMA and SOMA, proposed a third variant, rate-adaptive constellation expansion multiple access (REMA) [275]. The interest in NOMA has been manifested by other large telcos (ZTE Corporation, SK Telecom) [276], chip suppliers (Intel, Qualcomm) [277], OEMs (LG Electronics, Samsung, Nokia) [278] and equipment vendors (Anritsu) [279].

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 [280]. There is a related two-step random access channel (RACH) standardization within 3GPP as well [281].

## 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 [282]. 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 [282]. 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 [283]. 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 [284].

**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 [285]. However, the NOMA and cache-aided 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 [286]. 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 [287].

**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 [288]. 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, the 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 [289], the authors proposed OTFS-NOMA to improve the spectral efficiency and delivery latency of users 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 [290]. 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 [291]. MIMO-NOMA-based VLC is also becoming a popular approach to enhance the sum rate [292]. 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 a 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 [293] 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 designing, 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. This paper 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 identified 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. Hajaghayi, 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] 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.
- [12] 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.
- [13] O. Semiahi, 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.
- [14] ——, “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.
- [15] 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.
- [16] 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.
- [17] 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.
- [18] 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.
- [19] T. X. Tran and D. Pompili, “Adaptive bitrate video caching and processing in mobile-edge computing networks,” *IEEE Transactions on Mobile Computing*, vol. 18, no. 9, pp. 1965–1978, Sep. 2019.

- [20] 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.
- [21] 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.
- [22] D. Tse and P. Viswanath, *Fundamentals of wireless communication*. Cambridge university press, 2005.
- [23] 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.
- [24] 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.
- [25] 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.
- [26] 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.
- [27] 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.
- [28] 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.
- [29] 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,” *CoRR*, vol. abs/1609.01856, 2016. [Online]. Available: <http://arxiv.org/abs/1609.01856>
- [30] M. Aldababsa, M. Toka, S. Gökçeli, G. K. Kurt, and O. Kucur, “A tutorial on nonorthogonal multiple access for 5G and beyond,” *wireless communications and mobile computing*, vol. 2018, 2018.
- [31] 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.
- [32] 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.
- [33] B. Makki, K. Chitti, A. Behravan, and M.-S. Alouini, “A survey of NOMA: Current status and open research challenges,” *IEEE Open Journal of the Communications Society*, vol. 1, pp. 179–189, Jan. 2020.
- [34] 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.
- [35] H. Yahya, E. Alsusa, A. Al-Dweik *et al.*, “Error rate analysis of NOMA: Principles, survey and future directions,” Jan. 2022.
- [36] U. Niesen, D. Shah, and G. Wornell, “Caching in wireless networks,” in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, Jun./Jul. 2009.
- [37] 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.
- [38] 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.
- [39] 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.
- [40] 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.
- [41] 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.
- [42] 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. 212709–212737, Nov. 2020.
- [43] 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,” *Transactions on Emerging Telecommunications Technologies*, vol. 31, no. 2, p. e3702, 2020.
- [44] 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.
- [45] 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.
- [46] J. Wang, “A survey of web caching schemes for the internet,” *ACM SIGCOMM Computer Communication Review*, vol. 29, 10 1999.
- [47] 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.
- [48] 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.
- [49] “Video aware wireless networks,” <https://software.intel.com/content/www/us/en/develop/articles/video-aware-wireless-networks.html>, accessed: 2012-07-30.
- [50] P. Nuggehalli, V. Srinivasan, C.-F. Chiasseroni, and R. Rao, “Efficient cache placement in multi-hop wireless networks,” *IEEE/ACM Trans. Networking*, vol. 14, no. 5, pp. 1045–1055, Oct. 2006.
- [51] 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.
- [52] 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.
- [53] 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.
- [54] 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.
- [55] 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.
- [56] M. A. Maddah-Ali and U. Niesen, “Fundamental limits of caching,” *IEEE Trans. Inf. Theory*, vol. 60, no. 5, pp. 2856–2867, May 2014.
- [57] 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.
- [58] 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.
- [59] J. Pedersen, A. Graell i Amat, I. Andriyanova, and F. Bränström, “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.
- [60] 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.
- [61] 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.
- [62] 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.
- [63] A. Shokrollahi, “Raptor codes,” *IEEE Trans. Inf. Theory*, vol. 52, no. 6, pp. 2551–2567, Jun. 2006.
- [64] D. MacKay, “Fountain codes,” *IEE Proc. Commun.*, vol. 152, Dec. 2005.
- [65] P. Ostovari, A. Khreichah, and J. Wu, “Cache content placement using triangular network coding,” in *Proc. WCNC*, Apr. 2013, pp. 1375–1380.
- [66] E. Altman, K. Avrachenkov, and J. Goseling, “Coding for caches in the plane,” *arXiv preprint arXiv:1309.0604*, 2013.
- [67] 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.
- [68] 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.

- [69] 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.
- [70] 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.
- [71] L. Podlipnig, Stefan; Bösörmenyi, "A survey of web cache replacement strategies," *ACM Computing Surveys*, vol. 35, 12 2003.
- [72] M. Balamash, Abdullah; Krunz, "An overview of web caching replacement algorithms," *IEEE Commun. Surveys Tuts.*, vol. 6, 2004.
- [73] Y.-B. L. . W.-R. L. . J.-J. Chen, "Effects of cache mechanism on wireless data access," *IEEE Trans. Wireless Commun.*, vol. 2, 11 2003.
- [74] 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.
- [75] 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.
- [76] 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.
- [77] 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.
- [78] B. Panigrahi, S. Shailendra, H. K. Rath, and A. Simha, "Universal caching model and markov-based cache analysis for information centric networks," in *2014 IEEE International Conference on Advanced Networks and Telecommunications Systems (+ANTS)*, Mar. 2014, pp. 1–6.
- [79] Y. Zhang, X. Tan, and W. Li, "PPC: Popularity prediction caching in ICN," *IEEE Commun. Lett.*, vol. 22, no. 1, pp. 5–8, Jan. 2018.
- [80] Q. Yang, P. Hassanzadeh, D. Gündüz, and E. Erkip, "Centralized caching and delivery of correlated contents over gaussian broadcast channels," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 122–136, Jan. 2020.
- [81] 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.
- [82] J. Pedersen, A. Graell i Amat, I. Andriyanova, and F. Bränström, "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.
- [83] 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.
- [84] Y. Gui and Y. Chen, "A cache placement strategy based on compound popularity in named data networking," *IEEE Access*, vol. 8, pp. 196 002–196 012, 2020.
- [85] C. Zhan and Z. Wen, "Content cache placement for scalable video in heterogeneous wireless network," *IEEE Commun. Lett.*, vol. 21, no. 12, pp. 2714–2717, Sep. 2017.
- [86] W. Jiang, G. Feng, and S. Qin, "Optimal cooperative content caching and delivery policy for heterogeneous cellular networks," *IEEE Transactions on Mobile Computing*, vol. 16, no. 5, pp. 1382–1393, May 2017.
- [87] J. L. Hennessy and D. A. Patterson, *Computer architecture: A quantitative approach*, 5th ed. Elsevier Science, 2012.
- [88] 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.
- [89] 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.
- [90] 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.
- [91] 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.
- [92] W.-X. Liu, J. Zhang, Z.-W. Liang, L.-X. Peng, and J. Cai, "Content popularity prediction and caching for ICN: A deep learning approach with SDN," *IEEE Access*, vol. 6, pp. 5075–5089, Dec. 2018.
- [93] 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.
- [94] 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.
- [95] M. Ji, G. Caire, and A. F. Molisch, "Optimal throughput-outage trade-off in wireless one-hop caching networks," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, July 2013, pp. 1461–1465.
- [96] 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.
- [97] 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.
- [98] 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.
- [99] 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.
- [100] 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.
- [101] 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.
- [102] 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.
- [103] 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.
- [104] 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.
- [105] 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.
- [106] 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.
- [107] 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.
- [108] 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.
- [109] 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.
- [110] 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.
- [111] H. Nikopour and H. Baligh, "Sparse code multiple access," in *Proc. PIMRC*. IEEE, Sep. 2013, pp. 332–336.
- [112] 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.
- [113] 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.
- [114] 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.
- [115] 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.

- [116] 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.
- [117] 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.
- [118] 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.
- [119] W. Sun, D. Yuan, E. G. Ström, and F. Bränström, "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.
- [120] 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.
- [121] 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.
- [122] 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.
- [123] 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.
- [124] 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.
- [125] 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.
- [126] "Framework and overall objectives of the future development of IMT for 2020 and beyond," *Tech. Rep. ITU-R M.2083-0*, 2015.
- [127] "Multiple access for 5G new radio interface," *Tech. Rep. 3GPP RI-162305*, CATT, Busan, Korea, Apr. 2016.
- [128] "Candidate solution for new multiple access," *Tech. Rep. 3GPP RI-162306*, CATT, Busan, Korea, Apr. 2016.
- [129] 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.
- [130] J. M. Meredith, "Study on downlink multiuser superposition transmission for LTE," in *document RP150496, TSG RAN Meeting*, vol. 67, Mar. 2015.
- [131] 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.
- [132] M. Moghim, 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.*, Oct. 2020.
- [133] 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.
- [134] 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.
- [135] A. El Gamal and Y.-H. Kim, *Network information theory*. Cambridge university press, 2011.
- [136] 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.
- [137] 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.
- [138] V. W. Wong, R. Schober, D. W. K. Ng, and L.-C. Wang, *Key technologies for 5G wireless systems*. Cambridge university press, Apr. 2017.
- [139] 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.
- [140] 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.
- [141] X. Pei, H. Yu, Y. Chen, M. Wen, and G. Chen, "Hybrid multi-cast/unicast design in NOMA-based vehicular caching system," *IEEE Trans. Veh. Technol.*, vol. 69, no. 12, pp. 16 304–16 308, Dec. 2020.
- [142] 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.
- [143] 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.
- [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] 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.
- [146] 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.
- [147] 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.
- [148] Y. Li, H. Zhang, K. Long, S. Choi, and A. Naillathan, "Resource allocation for optimizing energy efficiency in NOMA-based fog UAV wireless networks," *IEEE Netw.*, vol. 34, no. 2, pp. 158–163, Mar. 2019.
- [149] 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.*, Early access 2021.
- [150] 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.
- [151] S. Mohan, S. Morgansgate, P. Basket, S. Gurugopinath, and S. Muhamidat, "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.
- [152] 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.
- [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] "Project loon. accessed: Jul. 15, 2017." [Online]. Available: <https://x.company/projects/loon/>.
- [155] 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.
- [156] 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.
- [157] 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.
- [158] 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.
- [159] 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.
- [160] ———, "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 vol.1.
- [161] 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.

- [162] 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 Transactions on Aerospace and Electronic Systems*, pp. 1–11, Early Access 2021.
- [163] 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.
- [164] Facebook, "Connecting the world from the sky," Facebook Technical Report, Tech. Rep., 2014.
- [165] E. Kulu, "Nanosatellite & cubesat database," [Online]. Available: <https://www.nanosats.eu/>, 28 Aug. 2021.
- [166] Z. Gao, A. Liu, C. Han, and X. Liang, "Files delivery and share optimization in LEO satellite–terrestrial integrated networks: A noma based coalition formation game approach," *IEEE Trans. Veh. Technol.*, pp. 1–1, Early Access 2021.
- [167] Z. Gao, A. Liu, and X. Liang, "The performance analysis of downlink noma in LEO satellite communication system," *IEEE Access*, vol. 8, pp. 93 723–93 732, May 2020.
- [168] X. Yan, H. Xiao, K. An, G. Zheng, and S. Chatzinotas, "Ergodic capacity of NOMA-based uplink satellite networks with randomly deployed users," *IEEE Syst. J.*, vol. 14, no. 3, pp. 3343–3350, Sep. 2020.
- [169] 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.
- [170] 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.
- [171] 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.
- [172] 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.
- [173] 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.
- [174] 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.
- [175] 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.
- [176] 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.
- [177] 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.
- [178] 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.
- [179] 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.
- [180] 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.
- [181] 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.
- [182] 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, Jul. 2019.
- [183] 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.
- [184] 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.
- [185] 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.
- [186] 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.
- [187] 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.
- [188] 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.
- [189] L. Liang, H. Peng, G. Y. Li, and X. Shen, "IEEE trans. veh. technol." *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 10 647–10 659, Dec. 2017.
- [190] 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.
- [191] 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.
- [192] 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.
- [193] 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.
- [194] S. Gurugopinath, P. C. Sofotasios, Y. Al-Hammadi, and S. Muhandat, "Cache-aided non-orthogonal multiple access for 5G-enabled vehicular networks," *IEEE Trans. Veh. Technol.*, vol. 68, no. 9, pp. 8359–8371, Sep. 2019.
- [195] S. Gurugopinath, Y. Al-Hammadi, P. C. Sofotasios, S. Muhandat, 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.
- [196] 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.
- [197] 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.
- [198] 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.
- [199] 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.
- [200] —, "NOMA/OMA mode selection-based cell-free massive MIMO," in *Proc. ICC*, May 2019, pp. 1–6.
- [201] 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.
- [202] 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.
- [203] 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.
- [204] 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.
- [205] 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.
- [206] 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.

- [207] 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.
- [208] 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.
- [209] 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.
- [210] 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.
- [211] 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.
- [212] 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.
- [213] 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.
- [214] 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.
- [215] 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.
- [216] 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.
- [217] 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.
- [218] 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.
- [219] 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.
- [220] 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, Thirdquarter 2018.
- [221] 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.
- [222] 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.
- [223] S. A. R. Naqvi and S. A. Hassan, "Combining noma and mmwave technology for cellular communication," in *Proc. VTC (Fall)*, 2016, pp. 1–5.
- [224] 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, 2017.
- [225] 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, 2018.
- [226] 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.
- [227] 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.
- [228] 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, 2020.
- [229] 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, 2018.
- [230] 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, 2018.
- [231] 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.
- [232] 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.
- [233] 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.
- [234] 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.
- [235] 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.
- [236] 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.
- [237] 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.
- [238] 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.
- [239] 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.
- [240] 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.
- [241] 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.
- [242] 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.
- [243] 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.
- [244] 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.
- [245] 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.
- [246] H. Zhang, Y. Qiu, K. Long, G. K. Karagiannidis, X. Wang, and A. Nallanathan, "Resource allocation in NOMA-based fog radio access networks," *IEEE Wirel. Commun.*, vol. 25, no. 3, pp. 110–115, Jun. 2018.
- [247] 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.
- [248] 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.
- [249] 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.
- [250] A. Liu and V. Lau, "Cache-induced opportunistic MIMO cooperation: A new paradigm for future wireless content access networks," in *Proc. IEEE Int. Symp. Inf. Theory*, Aug. 2014, pp. 46–50.

- [251] Y. Qian, S. Li, L. Shi, J. Li, F. Shu, D. N. K. Jayakody, and J. Yuan, “Cache-enabled MIMO power line communications with precoding design in smart grid,” *IEEE Trans. Cognitive Commun. Networking*, vol. 4, no. 1, pp. 315–325, Mar. 2020.
- [252] 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.
- [253] 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.
- [254] 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.
- [255] 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).
- [256] 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.
- [257] ATSC A/322:2021, “ATSC standard: Physical layer protocol,” ATSC 3.0, Tech. Rep., Jan. 2021.
- [258] 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.
- [259] 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.
- [260] P. Vanichchanunt, P. La-aidee, 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.
- [261] H. Sharifi-zadeh, 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.
- [262] 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.
- [263] A. Benjebbour, “An overview of non-orthogonal multiple access,” *ZTE Commun.*, vol. 15, no. S1, pp. 21–30, Jun. 2017.
- [264] 3GPP TR 36.859 V13.0.0, “Study on downlink multiuser superposition transmission (MUST) for LTE,” 3GPP, Tech. Rep., Dec. 2015.
- [265] 3GPP TR 38.812 V16.0.0, “Study on non-orthogonal multiple access (NOMA) for NR,” 3GPP, Tech. Rep., Dec. 2018.
- [266] 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.
- [267] 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.
- [268] 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.
- [269] 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.
- [270] 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.
- [271] 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.
- [272] 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).
- [273] DOCOMO 5G Whitepaper, “5G radio access: Requirements, concepts and technologies,” NTT DOCOMO, Tech. Rep., Jul. 2014.
- [274] 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.
- [275] 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).
- [276] SK Telecom 5G Whitepaper, “SK Telecom’s view on 5G vision, architecture, technology, and spectrum,” SK Telecom, Tech. Rep., Oct. 2014.
- [277] 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).
- [278] 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.
- [279] M. Fuse and K. Shioiri, “Non-orthogonal multiple access and massive MIMO for improved spectrum efficiency,” Anritsu, Tech. Rep. 91, Mar. 2016.
- [280] 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.
- [281] 3GPP RP-190711, “Revised work item proposal: 2-step RACH for NR,” 3GPP, Tech. Rep., Mar. 2019.
- [282] 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, 2007.
- [283] 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.
- [284] 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.
- [285] X. Mu, Y. Liu, L. Guo, J. Lin, and R. Schober, “Simultaneously transmitting and reflecting (STAR) RIS aided wireless communications,” *IEEE Trans. Wireless Commun.*, Oct. 2021.
- [286] 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.
- [287] 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.
- [288] 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.
- [289] 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.
- [290] 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.
- [291] 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.
- [292] 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.
- [293] 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.

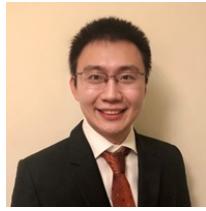
He joined Electronics and Communication Engineering Department, National Institute of Technology, Durgapur, India in 2005. He is currently an Associate Professor there. In 2011, he was a Visiting Lecturer at Asian Institute of Technology, Bangkok. From 2014 to 2016, he worked as a Marie Curie fellow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

Dr. Chandra has published about 100 research papers in refereed journals and peer-reviewed conferences. He is a co-recipient of best short paper award at IEEE VNC 2014 held in Paderborn, Germany and delivered a keynote lecture in IEEE MNCAppls 2012 held in Bangalore, India. He is the Secretary of the IEEE P2982 Standard WG since 2020. 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 Chhatrapati 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.