Noname manuscript No.
(will be inserted by the editor)

Energy and Latency Efficient Caching in Mobile Edge Networks: Survey, Solutions, and Challenges

Lubna B. Mohammed · Alagan Anpalagan · Muhammad Jaseemuddin

Received: date / Accepted: date

Abstract Future wireless networks provide research challenges with many fold increase of smart devices and the exponential growth in mobile data traffic. The advent of highly computational and real-time applications cause huge expansion in traffic volume. The emerging need to bring data closer to users and minimizing the traffic off the macrocell base station (MBS) introduces the use of caches at the edge of the networks. Storing most popular files at the edge of mobile edge networks (MENs) in user terminals (UTs) and small base stations (SBSs) caches is a promising approach to the challenges that face data-rich wireless networks. Caching at the mobile UT allows to obtain requested contents directly from its nearby UTs caches through the device-to-device (D2D) communication.

In this survey article, solutions for mobile edge computing and caching challenges in terms of energy and latency are presented. Caching in MENs and comparisons between different caching techniques in MENs are presented. An illustration of the research in cache development for wireless networks that apply intelligent and learning techniques (ILTs) in a specific domain in their design is presented. We summarize the challenges that face the design of caching system in MENs. Finally, some future research directions are discussed for the development of cache placement and cache access and delivery in MENs.

Keywords Mobile edge network · edge caching · energy efficient · latency efficient · learning technique.

Department of Electrical, Computer, and Biomedical Engineering, Ryerson University, Canada
E-mail: (lubnabadri, alagan, jaseem)@ryerson.ca.

1 Introduction

In recent years, an exponential increase in traffic load has been noticed in wireless networks due to multimedia streaming applications and services, mobile video streaming, web browsing, and social network inter-connections. Wireless devices are expected to generate much higher traffic than wired network devices in the future [1]. To handle these traffic explosions, mobile wireless networks require continuous evolution and improve the performance in terms of power consumption, data throughput, and utilization of network resources such as backhaul network capacity and bandwidth [2]. Mobile edge networks (MENs) is one of the candidate solutions in future wireless networks. Despite the developments of wireless network architecture, the demand for contents by connected devices and many different applications and services on their mobiles, results in constraints put on latency, energy consumption, and quality of services (QoS).

Considering these problems, researchers investigated the possibility of caching content items locally and proactively at the edge of the mobile networks (i.e., caching in SBS and user terminal (UT)) before users request them. Local caching is a promising approach to improve the network bottleneck [3], [4] by providing faster connectivity, lower latency, and less power consumption. SBSs which are also called (sometimes) femto caches, caching helpers, or simply helpers, have normaly high storage capabilities and are used to build a wireless distributed caching infrastructure [5]. Utilizing the advantages of storing contents closer to UTs at the edge of the mobile edge networks allows users to download requested contents from neighbouring SBS or UT caches in SBS for user communication or device-to-device (D2D) communication, respectively, which can boost QoS and reduce the latency while saving power consumption and the network backhaul resources.

Future services and applications are highly bounded by user location, data, and network. Internet architecture with a huge amount of mobile traffic and having mobile users with different moving speed will suffer from poor support for such services. Thus, user mobility patterns should be considered while designing caching in MENs. Due to dynamic updates of user demands, content popularity, and user mobility, it is difficult to decide which contents to cache, where to cache them, and from where to deliver them using traditional decision making techniques. Moreover, the large amount of data needed to develop algorithms for cache system, makes the estimation of cache contents, access, and delivery, a complex and difficult task. In order to meet all these challenges, researchers explore learning and decision techniques for storing, accessing, and delivering the huge amount of data generated within the MENs. A summary of existing survey articles on mobile computing and caching is shown in Table 1. Table 2 lists the acronyms used in this paper.

The remaining of the paper is organized as follows: Sect. 2 discusses mobile edge computing and caching. In Sect. 3, the literature survey of solutions for mobile edge computing and caching challenges in terms of energy and latency are presented. Sect. 4 explains and compares different caching techniques in

mobile edge networks. In Sect. 5, discussion and comparison between different caching techniques in MEN is explored. Sect. 6 summarizes the challenges that faces the design of caching system in MENs. Sect. 7 discusses future research direction followed by conclusions and future work in Sect. 8.

2 Mobile Edge Networks

The increasing growth in mobile data traffic and new mobile applications leads to limitations on end users demands and communications at mobile devices. End users require service availability, service reliability, lower latency and efficient energy usage. To overcome limitations such as computation capabilities, storage capacity, latency, and energy consumption, new wireless network paradigm is needed [13]. Mobile edge network (MEN) architecture has been presented as a promising solution for future wireless networks. Proposals for MEN architecture are presented from industry and academia. They are evolved from the mobile cloud computing by utilizing the computing power and data storage away from the mobile devices into the cloud. MEN architectures are summarized in four main models depending on their services and operations. They are mobile cloud computing (MCC), mobile edge computing (MEC), fog computing, and caching [10].

The fundamental concept of MEN is to make network contents, services, and resources closer to the network edge. This can be implemented through the architecture design of MEN that deploys flexible computing and utilizes storage resources at the mobile network edges. MEC is a network architecture concept that was standardized by European Telecommunications Standards Institute (ETSI) and Industry Specification Group (ISG). MEC was acknowledged as a prime emerging technology for 5G networks [14]. At the edge of the network, IT service environment and cloud-computing capabilities are provided within the Radio Access Network (RAN) [15]. Cisco proposed fog computing as an extension of the cloud computing to wireless network edges. The aim is to accommodate the Internet of Thing (IoT) applications closer to users. At the same time, fog computing nodes are distributed in a wide area and collaborate among multiple end users to provide processing and storage [16]. Researchers at Carnegie Mellon University proposed cloudlet which is a new element that extends the mobile device-cloud architecture. Cloudlet is defined as resource-rich computer or cluster of computers that are connected to the Internet and nearby mobile devices [17]. Both Wi-Fi networks and mobile networks are deployed to provide near-real-time provisioning of applications and handoff of virtual machine images among edge nodes when a device moves. Cloudlet can reduce the end-to-end latency between the mobile device and the cloud [18].

Some of key technologies to enable MEN to be flexible and easy to maintain are software defined networks (SDN), network function virtualization (NFV), and information centric network (ICN)[12]. SDN separates the control plane from the data plane by allowing logical centralization of control and enabling

Table 1 Summary of Survey Articles on Caching in Wireless Networks.

Survey	Focuses on	Contributions
[6]	Popularity based video caching	- An overview of caching in wireless networks, - A comparison of traditional and popularity-based caching. - An overview of the attributes of videos and the evaluation criteria of caching policies. - A review of proactive caching, focusing on prediction strategies, challenges, and open research problems in popularity-based caching. - A comparison of traditional and popularity-based caching. - An overview of the attributes of videos and the evaluation criteria of caching policies. - A review of proactive caching, focusing on prediction strategies, challenges, and open research problems in popularity-based caching.
[7]	Caching strategies	- A survey on caching techniques in macro-cellular networks, heterogeneous networks, device-to-device networks, cloud-radio access, and fog-radio access networks. - A tutorial on caching techniques and caching algorithms. -A comparisons among different algorithms in different performance metrics. - A summary of the main research achievements, challenges, and research directions.
[8]	Achieve low latency communications	- A survey on the technologies to achieve low latency communications - An overview of 5G cellular network caching and mobile edge computing and other 5G requirements.
[9]	Deployment, strategies, edge caches, and network performances	- A survey on edge cache in radio access networks, the deployment location of content placement strategy, and coded caching. - A summary of the impacts on high spectral efficiency, high energy efficiency, and low latency. - Challenges in the joint optimization of radio and cache resources.
[10]	Research efforts made on the MENs	- A review on convergence of computing, caching and communications. - A survey on cloud technology, software defined network, and network function virtualization. - A review of the open research challenges and future directions.
[11]	Caching mechanism in information-centric networking.	- An overview of the in-network caching mechanisms - An illustration of how it works, its benefits and drawbacks. - A comparison of some typical in-network caching mechanisms through.
[12]	Device-enhanced multi-access edge computing (MEC)	- A survey on the device-enhancement of MEC services for end devices through the resources of other end devices. - A survey on computation offloading and caching to device enhanced MEC - A review of limitations of existing device-enhanced MEC mechanisms
Our paper	Energy and latency efficient caching in MENs.	- A survey of solutions for mobile edge computing and caching in terms of energy and latency. - Comparison of caching in MENs and between different caching techniques. - An illustration of the research in cache developments that apply intelligent and learning techniques. - A summary of the challenges that faces the design of caching system. - A discussion of future research directions for the development of cache placement and cache access and delivery in MENs.

Table 2 Table of Acronyms.

Acronym	Meaning
AI	Artificial Intelligence
ANN	Artificial Neural Networks
AR	Augmented Reality
CRP	Chinese Restaurant Process
CUs	Cloud Units
D2D	Device-to-Device
DCNNs	Deep Convolutional Neural Networks
DNN	Deep Neural Network
ELM	Extreme Learning Machine
ENs	Edge Nodes
ETSI	European Telecommunications Standards Institute
FSS	Fuzzy Soft Set
IIoT	Industrial Internet of Things
ILTs	Intelligent and Learning Techniques
IoT	Internet of Thing
ISG	Industry Specification Group
LFU	Least Frequently Used
LRU	Least Recently Used
MBS	Macrocell Base Station
MCC	Mobile Cloud Computing
MDP	Markov Decision Process
MEC	Mobile Edge Computing
MENs	Mobile Edge Networks
MILP	Mixed-Integer Linear Program
ML	Machine Learning
PP	Pricing Problem
QoS	Quality of Services
RAN	Radio Access Network
RL	Reinforcement Learning
RMP	Restricted Master Problem
RNN	Recurrent Neural Networks
SBSs	Small Base Stations
SCNs	Small Cell Networks
SVC	Scalable Video Coding
SVD	Singular Value Decomposition
UTs	User Terminals
VR	Virtual Reality

direct programming of wireless network controls with improved energy efficiency. ICN is used to speedup content distribution and utilize network resources [10]. ICN serves requests from closer content nodes along the path which enables content caching in both the air and the mobile devices. NFV enables flexible design and management of network functions independent of the underlying physical network equipment [19]. Integrating the programming control principle in SDN with information centrality in ICN leads to dynamic networking, caching, and computing resources to meet the requirements of different applications [20], [21]. Also, NFV-based caching solutions offers caching of personalized and secure contents isolated from other content providers and from other participants [12]. By utilizing these technologies, functions, con-

tents and resources are moved closer to end users. This enables the MEN that exploits a large number of low-cost storage devices at different places in network edges, to proactively cache popular contents during off-peak periods. Caching can be deployed at different levels in mobile networks instead of fetching them from the core network [10]. Reducing the number of network hops between the location of the contents and the user requesting the contents will reduce the latency for retrieving the contents [22]. MENs bridge the gap between the capability limitations of storage and computation in user terminals and their increasing demands. It is done by placing storage and computation resources at the edge of the network closer to user terminals. MEN can reduce latency and energy consumption. There are various techniques in the literature that are proposed to process data locally at the edge of the network and accelerate data streams, which will reduce the traffic bottleneck toward the coding network [23]. The caching locations which are considered as caching levels in MEN architectures are macro base station (MBS), small base station (SBS), and user devices allowing for device to device (D2D) communications. The places that can be used to cache most popular contents within MENs are shown in Figure 1 and described below:

1. **UT caching:**
   Exploiting the storage resources in UTs is one of the key technology in 5G networks [10]. Caching in user devices allows the improvements of caching strategies to allow D2D communications.
2. **SBS caching:**
   Each cell in MEN employs a large number of SBSs. SBS includes higher storage capacity than UT cache capacity. They are closer to the end users and usually provide higher data rates [10]. Therefore, utilizing caching in SBSs is a promising solution to improve QoS in next-generation heterogeneous networks.
3. **MBS caching:** MBS covers a larger area within the heterogeneous network and can serve more users. The storage capacity in MBS is higher than other caches within the cell which will lead to a higher probability of finding the requested file (hit).

3 Energy and Latency Efficiency in Mobile Edge Networks

This section presents the benefits of MEN as an emerging technology for the future wireless networks. There are a number of research work that has been done to show the efficiency of MEN in terms of energy consumption, latency requirements, and storage capacity for different applications and services. Table 3 and Table 4 present an overview of the main research work areas in the literature that proposed possible solutions to future mobile edge computing and caching network in terms of energy and latency, respectively. The research areas can be categorized into the following main streams:

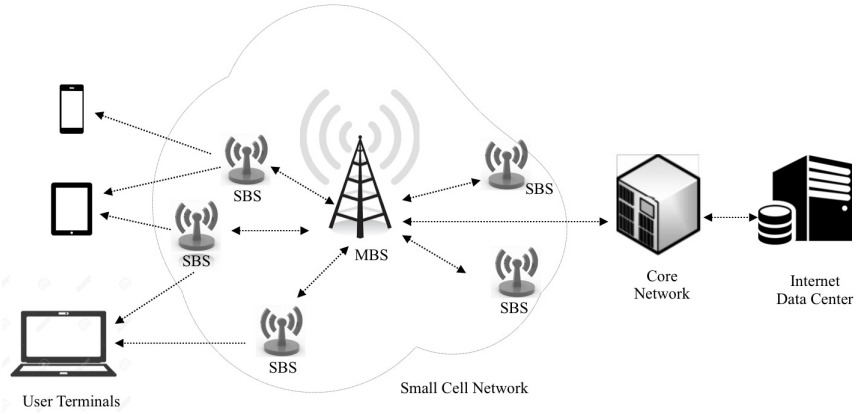

Fig. 1 Architecture of Mobile Edge Network.

1. Computation offloading: The advances in computing technology and various applications that require high computation power and resources to run complex programs have increased lately. These applications use wireless networks and run on mobile devices with limited capabilities to support the needed resources [24]. One of the solutions to solve this problem is computation offloading. In computation offloading, the mobile devices transfer tasks to an external edge cloud and receive the results from the edge cloud. Offloading increases mobile terminals capabilities by migrating the computation to more resourceful computers (servers) at the edge of the network [25].
2. Task caching: Computation offloading considers computing capabilities at the edge of the network by assuming enough hardware and software resources to execute the tasks. However, enough storage capacity for computation offloading is another important challenge that faces future wireless networks. In [26], the authors proposed the task caching which refers to caching the task (application computation task) and its related data at the edge of the network.
3. Content caching: Contents requested by end users in massive multimedia services over mobile network face network capacity limitations and increases backhaul links load. Requesting the same contents by different users also causes network congestion and a waste of network resources .The development of mobile edge caching techniques is another promising solution for wireless networks. Content caching of the most popular files can prevent duplicate transmission of the same contents and improve end-user QoS. Quality of service can be improved, since downloading the contents from network edge (for example, base stations or end user terminals) reduces latency compared to downloading the contents from Internet contents providers (core network) [10].

4. Resource allocation: With multiple user terminals, MEN servers have much less computational resources. One of the main issues in the design of MEN is to consider resource allocation. It is the process of allocating the finite radio and computational resources to multiple mobiles under resource constraints. There are two categories of resource allocation schemes for MEN: centralized and distributed. In the centralized resource allocation, the MEN server is responsible for all mobile information, makes the resource allocation decisions, and sends the decisions to mobile devices. While in distributed resource allocation, many techniques including game theory and decomposition techniques are used to develop a distributed algorithm [27].
5. Multicast caching: To reduce the load of wireless links in traditional unicast connection-based transmission and avoid transmitting the same file multiple times to multiple users, a multicast caching is proposed for base stations in 5G mobile networks [28]. In multicast caching, the popular content is brought close to the users. The optimization objectives are to minimize the average latency for all content requests and minimize the average energy consumption [8].
6. Service chain management: Service chaining policy refers to the term that describes executing multiple service functions in an ordered list to guarantee performance and security requirements [29]. In MEN networks, light weight data centers can be used at the edge of the network. In these centers, operators deploy service chaining as to steer traffic through the management of a set of service functions. Service chaining is realized by using the software-defined network and network function virtualization technologies. Service chain management can reduce network latency by offloading the workload or bandwidth from the core network service [30].

Mobile edge computing and caching are considered as a promising solution that supports many emerging applications and services with specific constraints of latency, energy, and reliability [31]. In the following section, comparison of different caching techniques in mobile edge network are presented.

4 Caching in Mobile Edge Network

Most research work in the literature cache either uncoded contents or uncoded parts of files during the placement phase. The base station broadcasts the coded files (linear combination of multiple files) to user terminals during the content delivery phase. Then, the users can decode their files simultaneously [54]. During the delivery phase, the cache memory contents of the requested user are updated to store new files. There are different algorithms and techniques to implement cache replacement. Researchers have proposed caching algorithms for wireless systems that range from simple algorithms to more advanced intelligent techniques. These algorithms are divided into two main streams. The first is cache replacement algorithms based on prior knowledge about contents popularity, while the second is cache replacement without prior knowledge about contents popularity [55]. Table 5 shows a comparison of

Table 3 Overview of Techniques Based Energy Efficient in Mobile Edge Computing and Caching.

Area	Ref	Approach	Summary
Computation offloading	[24]	Markov decision process (MDP)	Propose a spatial and temporal computation offloading decision algorithm in edge cloud-enabled heterogeneous networks.
	[32]	Minority games theory	Develop distributed server activation mechanism.
	[33]	Multi-label classification and deep learning	Develop a dynamic offloading framework for mobile users.
	[34]	An approximation collaborative computation offloading	Present centralized cloud and MEC over FiWi networks and the cloud-MEC collaborative computation offloading model.
	[35]	Karush Kuhn Tucker Lagrangian multiplier method	Design an energy-efficient autonomic offloading scheme for physical layer design and application latency.
	[36]	Dynamic sequential game theory	Propose an adaptive sequential offloading game approach.
Task caching	[26]	Mixed integer non-linear programming	Caching of complete task application and their related data and design a multi-user computation offloading algorithm in edge cloud.
Content caching	[37]	Poisson point processes modelling of BS power and locations	Analyse energy consumption of cache-enabled wireless network using spatial model based on stochastic geometry.
	[38]	Lagrange multiplier and duality	Exploit statistical information on individual popularity preferences in caching policies.
	[39]	Social-tie factor Modelling	Proposed social-aware cache information processing for future ultra-dense networks.
	[40]	Dual decomposition and a sub gradient algorithm	Propose joint design of the transmission and caching policies and formulate problem that minimize a generic cost function.
Resource Allocation	[41]	Mixed discrete-continuous optimization	Propose a joint caching and offloading mechanism that optimally allocate the storage resource at the BS for caching and the uploading and downloading time durations.
	[42]	Dual-decomposition method and alternating direction method of multipliers	Propose optimization problem to jointly consider bandwidth provisioning and content source selection.
Multicast allocation	[43]	Graph theory	Propose multicast caching in dense small cell networks when a group of users can benefit from multicast caching at a lower energy cost.
	[44]	Distributed potential game model and cloud and wireless resource allocation algorithm	Propose a distributed joint computation offloading and optimization scheme in heterogeneous networks.
	[45]	Clustering method based on game theory	Propose a user attribute aware video distribution mechanism using scalable video coding.

Table 4 Overview of Techniques Based Latency Efficient in Mobile Edge Computing and Caching.

Area	Ref	Approach	Summary
Computation offloading	[46]	Heuristic search, reformulation linearization and semi-definite relaxation	Formulate an optimization problem to jointly minimize the latency and offloading failure probability.
	[47]	Lyapunov stochastic optimization	Propose a dynamic policy for task offloading and resource allocation.
	[48]	Lyapunov optimization	Investigate a green MEC system with energy harvesting devices and propose computation offloading strategy.
	[36]	Dynamic sequential game theory	Propose an adaptive sequential offloading game approach and design a multi-user computation offloading algorithm.
	[49]	Markov decision process	Develop a post-decision state based learning algorithm that learns the optimal joint offloading and auto scaling policy on-the-fly.
Service chain	[30]	Hash-based group Management	Propose a hash-base group table to reduce the computation time for assigning user into groups to reduce the control plane latency.
Content caching	[50]	Transfer learning algorithm	Propose a proactive content caching optimization model.
	[51]	Assessment tests of caching solution	Propose a prototype implementation of a mobile edge cache.
Resource allocation	[31]	Submodular optimization	Propose resource cognitive intelligence based on learning of network contexts and design an optimal caching strategy.
	[52]	Auction theory	Propose a decoupled resource allocation model that manages the allocation of computing resources distributed at the edges independent of the service provisioning management performed at the service provider end.
	[53]	Nash bargaining game	Propose resource optimization problem based users fairness and the global throughput.

different caching techniques in literature in terms of their dependency on content popularity, online learning, training phase, context-aware, socially aware, mobility aware, and prediction ability.

In [56] and [57] least recently used (LRU) and least frequently used (LFU) algorithms are used respectively. These techniques are simple cache replacement algorithms that do not consider future content popularity and update the cache continuously during the delivery phase. In the LRU algorithm, the cache includes an ordered list which is updated to follow the recent access of all cached contents. When the cache is full, the new content is placed in the least

recently accessed cache content. The content of cache can be changed based on prior knowledge of content popularity. The research work in [58] and [59] use popularity statistics of different video files modeled using a Zipf distribution. The cache replacement algorithm by tracking variations in the popularity distribution and updating cache content at user terminals and collaborative device to device communication are combined to increase the efficiency of content delivery. There is a trade-off between having an optimal content replacement that predicts future requests efficiently and the speed of computing the content popularity. Also, in these methods, there is no personalization to user context and preferences.

In [60] and [61], authors exploited storing of video files closer to users in femto caches. They formulate the problem with the aim to increase the throughput by unloading a lot of traffic from the main cellular network. The work presented in [2] proposes caching and multicasting techniques. Caching aims to allow popular content files at network edges in order to shorten the distance between contents and requesters. While multicasting aims to serve identical requests happening at nearby locations through common multicast streaming by sending multiple copies of the same content to different users. The exploiting of proactive caching contents based on file popularity and correlations among users and files patterns are proposed in [3]. Files can be proactively cached during off-peak demands by using a machine learning algorithm and collaborative filtering, with context-awareness. The procedure aims to predict the set of influential users and social structure, and to proactively cache strategic contents on those user terminals to utilize device-to-device communications. This approach requires a training set of known content popularities and can learn during a training phase to decide which content to place in the cache.

In [62], the cache strategy is modeled in a heterogeneous small cell network using a reinforcement learning based coded caching framework. Authors have designed an optimal cache placement policy that uses the learned file popularity to find the optimal cache contents. The cache placement policy takes into account the users' connectivity to the SBS. At regular intervals, the cache pre-fetches segments of the popular files (coded) to serve users' requests. Caching algorithm is presented in [63] based on contextual multi-armed bandits optimization. In this algorithm, the SBS updates its contents regularly by observing the demand of cached files and learns the contexts of popularity profile over time. The objective of the multi-armed bandit optimization is to maximize the number of cache hits. While in [64], a different extension of the multi-armed bandit framework is proposed. In this framework, the authors have exploited the topology of user connections by incorporating coded cache contents. Based on observations of instantaneous demands that assume content popularity distribution, an optimal cache content placement strategy can be achieved. While previous algorithms do not consider future prediction of popular contents in the design of cache replacement algorithm, the work in [65] and [66] aim to learn popularity trends. Their works include the design of context-aware proactive caching. There is no prior knowledge about con-

tent popularity in [65], while in [66] the cache replacement method learns the popularity of contents and uses it to determine which contents to place in the cache and which contents to evict from the cache.

5 Comparison of Different Caching Techniques in MEN

There are different studies that formulate the caching problem at the edge of the network. These proposals examine the problem from different perspectives. In each study, the optimization problem is formulated based on input attributes that are manipulated by the optimization algorithm and the scheme of caching used in the model. The performance indices in these proposals are overall delay, user satisfaction ratio, offloading probability, and total throughput. They have one general common objective which is redirecting user requests from the expensive and limited backhaul links to local cache storages at the edge of MEN networks. Table 6 illustrates a comparison between different caching techniques in MEN networks.

In [67], the authors studied the association between UTs and SBS in small cell networks. Based on file availability in SBS and the backhaul congestion state, the SBSs decide which users they should serve. The problem is formulated using one-to-many matching game theory. In [68], two proactive caching scenarios are examined. The goal in both cases is to keep user satisfaction ratio above the required limit. In the first case, the contents are cached proactively at the SBSs during low-peak demands. The cache procedure is built based on supervised machine learning algorithm using singular value decomposition (SVD). This technique includes two parts. The training of the input matrix that represents the users'-to-files rating association and predictfollowed by conclusions and future work in Secting what files each user will request (file popularity matrix). In the second case, the contents are cached proactively in users' devices. The centrality metric is used to measure the social influence of a node and its connection with other nodes (social community). Then, the k-means clustering method is used to form a set of influential users within a community (users' social ties). Proactive caching procedure considers the number of times each file is downloaded by each user to form a user-to-files history matrix. The beta distribution is assumed to denote the probability that a content is selected by a given user. The popular contents that will be cached in influential users' devices are selected based on Chinese restaurant process (CRP). By caching at UT and enabling D2D communication, the load on SBS and backhaul loads are reduced [68].

In [69], optimal two one-tier caching placement is presented based on the difference of convex programming. The objective of the optimization problem is to maximize the offloading probability. The offloading happened in three cases:

1. Self offloading when the requested contents are found on UT (local cache),
2. D2D-offloading when the requested contents are found from near devices, and

3. SBS-offloading when the requested contents are found in near SBS.

Their results show that popular contents must be cached under relatively low node density while other contents must be cached evenly under relatively high node density.

In [70], the author formulates the caching problem as a video recommendation system. They clustered files and users depending on video file preferences and formulate the cache scheme in two phases: the D2D cooperative phase where files are stored in UTs and caching phase where files are stored in SBS. The optimal caching is designed based on the greedy intra-cluster algorithm to obtain minimum total average file download time. The results show that clustering files before applying the optimal caching algorithm can reduce the computational complexity of the huge number of involved users and files. The work in [71] proposed joint caching, routing, and channel assignment for video files in collaborative small cell cellular network. Their objective is to maximize network throughput by using conflict graph to characterize the communication link interference. The optimization problem is modeled as a large-scale linear programming problem that is solved using column generation method. The algorithm selects a subset of variables that have potential improvements to the objective function in order to minimize the complexity of the optimization problem. The optimization problem is then divided into two sub-problems: restricted master problem (RMP) and pricing problem (PP). Their results show that the overall throughput of the video data that can be delivered to users, is considerably increased over the state-of-the-art femto caching models. In [72], proactive caching is designed based on traditional collaborative filtering by regularized decomposition to estimate the popularity matrix. Then transfer learning is used to improve the estimation accuracy by transferring and learning hidden knowledge from other domain such as social networks. Finally, an optimal caching strategy is implemented as a distributed iterative algorithm to update the cache. Results show that user satisfaction ratio increases with the number of SBS compared to other caching approaches. Authors in [73] investigate proactive caching for service providers to reduce redundant backhaul transmission to edge nodes (ENs). The Stackelberg game is used to formulate the problem and it was decomposed into two sub-games, a storage allocation game and a number of user allocation games. The service provider is modeled as a leader that decides the prices for the storage and backhaul resources on ENs. ENs are modeled as followers. Their results show that lower total backhaul resources can be achieved with proactive caching based game theory compared with centralized popularity based caching and random caching. The authors discussed the complexity and scalability of edge caching in wireless communication networks where there will be a large number of ENs, users, and service demands and will involve a huge amount of data. The complexity is defined as the number of iteration steps of the caching algorithm and the amount of information exchanged between network edges. The performance of the caching algorithm with the increase of network size is addressed with the scalability of the caching algorithm. The popularity estimation used in

Table 5 Comparison of different caching techniques.

Cache Technique	Content Popularity	Online Learning	Training Phase	Context Aware	Social Aware	Mobility Aware	Predict Aware
LRU and LFU [56] and [57]	Yes	No	No	No	No	No	No
PopCaching [66]	Yes	Yes	No	No	No	Yes	No
Femtocaching (coded cache) [60] and [61]	Yes	No	No	No	No	No	No
Femtocaching (uncoded cache)[58]	Yes	No	No	No	No	No	No
QoE-oriented resource efficiency caching [59]	Yes	No	No	No	No	No	No
Multicast aware caching [2]	Yes	No	No	No	No	No	No
Proactive caching based machine learning [3]	No	No	Yes	Yes	Yes	Yes	No
Distributed caching strategies [62]	Yes	Yes	No	No	No	No	No
Context-aware cache [55] and [63]	No	Yes	No	Yes	No	Yes	No
Optimal uncoded cache placement [64]	Yes	Yes	No	No	No	No	No
Decentralized cache [72]	Yes	No	No	No	Yes	Yes	No
Stackelberg game caching [73]	Yes	No	No	No	No	No	No
Popularity prediction caching [74]	Yes	No	Yes	Yes	No	Yes	No
Cache-aware user association [67]	Yes	No	Yes	No	No	No	No
Social and spatial proactive caching [68]	Yes	No	Yes	Yes	Yes	Yes	No
Optimal caching placement [69]	Yes	No	No	No	No	No	No
Clustered D2D caching [70]	Yes	No	No	No	No	No	No
Joint caching [71]	Yes	No	No	No	No	No	Yes
Cell-site-aware caching [75]	Yes	No	No	No	No	No	No
Adaptive caching [76]	Yes	No	Yes	No	No	No	Yes
Mobility-aware caching [77]	Yes	No	No	No	No	Yes	No

Table 6 Comparison of different caching techniques in MEN.

Ref. Location	Methodology Attributes	Caching	Input Index	Objective	Constraints	Performance
[67]	Game Theory	SBS	Content availability Delivery data rate	Maximize UT to SBS association utility	Backhaul capacity Storage size	Overall delay
[68]	Regularized SVD	SBS	User to files rating	Minimize backhaul load	Backhaul rate Cache storage	User satisfaction ratio
[68]	Eigenvector centrality K-means clustering	UT	Users' social ties Users to files history	Minimize small cell load	SBS data rate D2D data rate UT storage	User satisfaction ratio
[69]	Difference of Convex programming	UT SBS	Content popularity D2D transmission range SBS transmission range Users and SBS densities	Minimize total offloading probability	UT storage SBS cache storage	Offloading probability
[70]	Greedy intra-cluster algorithm	UT SBS	D2D download time BS download time	Minimize total average file download time	storage capacity	Delay loss
[71]	Column generation method	SBS	User requests File popularity	Minimize schedule length	Cache capacity Received file fraction Files cached in SBS Channel capacity Channel active time	Total throughput
[72]	Regularized decomposition	SBS	The popularity matrix	Maximize number of users served by neighbouring SBS	Cache size	User satisfaction ratio
[73]	Stackelberg game	SBS	File demand probability EN storage cost Backhaul bandwidth Backhaul cost	Minimize the total backhaul resources Maximize EN utility	Cache capacity SBS coverage	Total backhaul resources and outage
[74]	Intelligence based content-aware	BS	Video popularity	Minimizes the backhaul load	Cache capacity	Offload ratio
[76]	Mixed-Integer linear program	BS	Content popularity Cache deployment	Minimize content downloading delay	Initial file transferring cost Cache deployment budge	Predictive Type-I and Type-II error probability
[77]	Greedy Algorithm	BS UT	Set of files Set of coded segments Cache strategy matrix	Maximize cache hit rate	BS capacity UT capacity	Cache hit rate

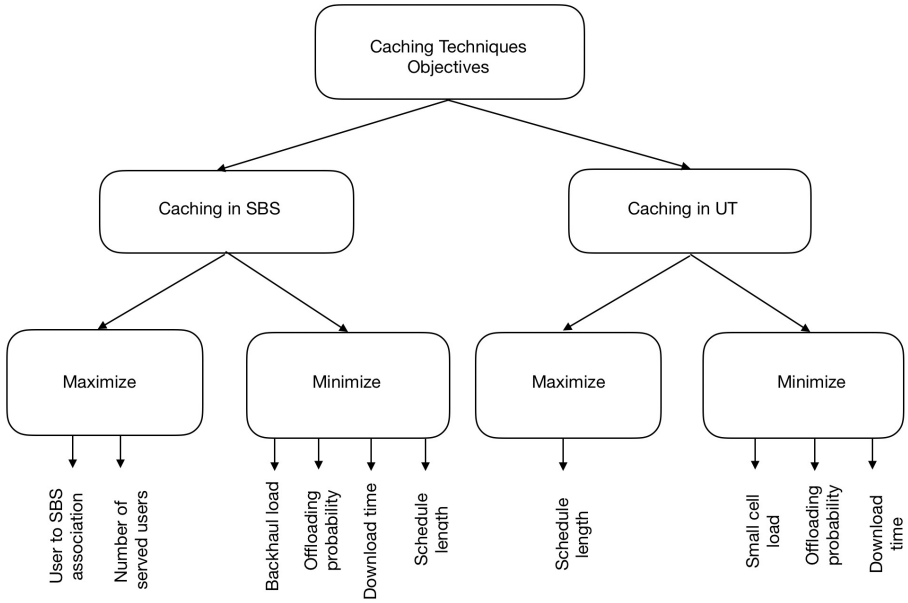

Fig. 2 Different caching techniques objectives in MEN.

caching techniques for video files in the work presented previously is based on user request probability or the number of views of videos. The work in [74] computes video popularity for published and unpublished videos using intelligence based content-awareness. The prediction of video popularity enhances cache placement decision as well as the quality of service in cellular networks.

In [76], an adaptive caching scheme is proposed that takes into account user behavior, content popularity, request statistics from users, and operating characteristics of the cellular network. The network operating characteristics include network topology, link capacity, routing strategy, cache size, and energy usage to read/write files from hardware storages (which is called cache deployment cost). The content popularity is predicted using the extreme learning machine (ELM) based on content features. The features of the content are computed using a combination of human perception models and network parameters. The adaptive caching scheme uses a mixed-integer linear program (MILP) to cache the results of popularity estimators. The caching decision is made while the network is not heavily utilized. Without affecting network quality of service, popular content can be transferred between BSs. The impact of mobility awareness in cache placement algorithm is discussed in [77]. The authors formulate the problem of caching coded segments at BSs and UTs taking into account users mobility and the content amount per transmission. The optimization problem is formulated as an integer programming problem that can be solved by submodular optimization.

6 Challenges in Designing Caching Techniques

As illustrated in Figure 1, mobile edge network (MEN) consists of multiple small base stations (SBSs) and user terminals (UTs). Each macrocell consists of one MBS connected to a gateway of the core network via a high-speed interface, N SBSs which are connected to the MBS through backhaul links, and M UTs connected to neighboring devices through D2D communication and to SBSs. The sets of SBSs and UTs are denoted by s=$\{s_1, s_2, \cdots, s_N\}$ and u=$\{u_1, u_2, \cdots, u_M\}$, respectively. There are cache storage in each MBS, SBS, and UT with different storage capacities. Within one macro cell, cache storage capacities can be defined by two sets c_s=$\{c_{s_1}, c_{s_2}, \cdots, c_{s_N}\}$ and c_u = $\{c_{u_1}, c_{u_2}, \cdots, c_{u_N}\}$ for SBSs and UTs caches, respectively. Assume that MBS has enough cache capacity to store F files defined by the set f=$\{f_1, f_2, \cdots, f_F\}$. Following the study of different caching techniques in literature, we can summarize the following challenges:

6.1 Content Popularity Modelling

In order to improve the performance of caching strategies, it is required to incorporate content popularity in caching decision making [66]. The content library consists of F files and stored at the MBS cache. Each file is f_z for $z = 1, \cdots, F$. The size of file f_z is denoted as f_{l_z}. The files are requested from the main library based on their popularity distribution. The popularity of the F files are denoted by the set p, where p=$\{p_1, p_2, \cdots, p_F\}$. The set p can be characterized by Zipf popularity distribution. If the files are arranged from the highest popular file to the lowest popular file, the popularity of the $i-$ th ranked content can be shown by Eq. (1) [78]:

$$p_i = \frac{\frac{1}{f_i^{\gamma}}}{\sum_{i=1}^{F} \frac{1}{i^{\gamma}}}. \tag{1}$$

The distribution for file f_i is characterized by the exponent factor γ, also called the skewness of the popularity. When $\gamma = 0$, the popularity is uniform over contents. As γ grows, the popularity becomes more skewed. Table 7 shows the methods used to model content popularity in literature. In many work, the popularity of files are generally modelled using Zipf distribution of all files. Zipf distribution gives a fixed popularity profile and it is assumed that content popularity is known in advance. Based on Zipf distribution, a small portion of Internet contents is highly popular while the rest is rarely requested [74]. In reality, content popularity needs to be estimated depending on number of related factors and not only on the content popularity distribution. Examples of these factors are files' preferences, users' preferences, users' context, social network characteristics, users' previous requests, etc. Also, content popularity must not be fixed and it is expected to change continuously with time, date, and location.

Table 7 Content Popularity Modelling in Caching Techniques

Ref.	Content Popularity Method
[56], [57], [60], [61], [2], [3], [64], [73], [67], [71], and [77]	- Zipf distribution.
[79]	- Number of views vs rank of videos in terms of views.
[74]	- Feature extraction and popularity prediction for unpublished videos.
[72], [62], and [76]	- Popularity estimation based on learning methods.

The approach in [79], assumes the popularity of video contents changing slowly and the popularity distribution of all files can be considered as fixed and known prior to the cache placement algorithm. They defined the popularity distribution of video files depending on the number of views vs the rank of videos in terms of views. In [63], context-dependent popularity profiles are learned online while observing connected users' demands and their context information. The placement algorithm does not depend on prior knowledge of content popularity, but it models connected users' context-dependent demands of files following Zipf distribution. The context information used in modeling the content popularity is the maximum number of users that can be served by SBS, the fraction of female users, and the fraction of underage users. The total number of files that are used in Zipf distribution formula is divided according to the context information of connected users.

In [74], the authors proposed a popularity prediction model for video files. The popularity of video contents is estimated from both published (statistical information) and unpublished video (newly uploaded videos). The process consists of the following stages, (1) Feature extraction from unpublished videos based on deep neural network technique, (2) Clustering the features resulting from stage 1 based on collaborative filtering technique, and (3) Fitting a regression model to predict the popularity of unpublished videos while using the statistical information of the published videos as a training set to the regression model. The approach in [72] used Zipf distribution as the training set to design a learning based approach. Their model estimates content popularity using regularized decomposition based collaborative filtering and they improve estimation accuracy using transfer learning technique.

6.2 User Mobility

Modeling user mobility depends on spatial and temporal properties. The spatial properties provide physical location information of user mobility patterns, while the temporal properties provide time-related information [80]. User mobility can be modeled by assuming a pairwise contact process that follows an independent Poisson process. The work in [81] implies that the occurrence time of pairwise contact event can be predicted in large time scale. The Pois-

son process is used for counting the occurrence of contacts between UTs, and between UTs and SBSs occurring at a certain rate.

To establish successful communication between mobile UT u_i and SBS s_j, u_i must be within the communication radius of SBS s_j. For independent Poisson process, the pairwise contact duration $T_{i,j}^{SBS}$ between mobile UT u_i and SBS s_j follows the exponential distribution with parameter $\lambda_{i,j}^{SBS}$. Here, $\lambda_{i,j}^{SBS}$ represents the contact rate between mobile UT u_i and SBS s_j. The contact duration $T_{i,j}^{SBS}$ when mobile UT u_i is within the communication range of SBS s_j can be defined as follows [77]:

$$T_{i,j}^{SBS} = \{(t - t_0) : ||l_j^t - l_i^t|| < d^{SBS}, t > t_0\}, \tag{2}$$

where t_0 represents the most recent time when mobile UT u_i enters the communication range d^{SBS} of SBS s_j. The locations of SBS s_j and mobile UT u_i at time t are represented by l_j^t and l_i^t, respectively.

Similarly, to establish successful communication between mobile UT u_i and mobile UT u_k, the shortest distance between the two devices must be within the communication range d^{D2D}. The contact rate between mobile UT u_i and mobile UT u_k is represnted by $\lambda_{i,k}^{UT}$. The contact duration $T_{i,k}^{UT}$ when mobile UT u_i and mobile UT u_k are within the communication range of d^{D2D} can be defined as follows [77]:

$$T_{i,k}^{UT} = \{(t - t_0) : ||l_i^t - l_k^t|| < d^{D2D}, t > t_0\}, \tag{3}$$

where t_0 represents the most recent time when mobile UT u_i enters the communication range d^{D2D} of mobile UT u_k, and l_i^t and l_k^t represent the locations of UT u_i and UT u_k at time t, respectively. In most of the work discussed in previous sections, it is assumed that users remain stationary while requesting and obtaining files. Research with this assumption does not include mobility as an effective parameter while taking cache placement decisions. A user may be served by any SBS located in the user communication range. Considering mobility on the caching design in future wireless networks caching, can be classified into three categories based on cache location:

1. Cache in SBS: In these research [82], [83], [84], and [85], file caching in SBS while user mobility is considered.
2. Cache in mobile UT: In these research, [86], [87], [80], [88], [89], and [90], user mobility-aware caching design are proposed by utilizing D2D communication links.
3. Cache in SBS and mobile UT: In [84], [80], [85], and [91], the researchers proposed user mobility-aware cache placement in SBS and mobile UT.

Caching efficiency can be improved by exploiting user mobility aware cache placement in SBS and UT [80]. However, few cache placement techniques have taken the impact of user mobility [77]. Most existing approaches which estimate cache contents proactively face the redundancy problem. This has happened in some caching strategies that have neighboring SBSs storing the same

popular contents. Redundancy results in wastage of cache resources and minimizes the cache storage capacity that is available for users. The interactions between network edges should be taken into account while optimizing the caching strategy [92].

6.3 Power Constraint

Energy consumption becomes a more challenging problem in the design of wireless communications due to increase in energy consumption cost, number of broadband wireless network users, and growing demand of the contents in the future networks [93]. Delivering contents from SBS to UTs and from UT to another UT will consume power and drain energy at both the network and UT. Cache system should be designed with an objective to find optimal transmit power and sustain the continuous growth of power consumption [94]. There are two power consumption models presented and discussed in [77], as follows:

1. **Energy consumption for D2D caching:**
   It is assumed that the interference of D2D communication is not considered. When UT u_k transmits the cached contents to UT u_i, the components of power consumption of UT u_k are given as follows:
   - β_{UT} is the inverse of power amplifier efficiency factor of mobile UT u_k,
   - $\mathcal{P}_T^{u_k}$ is the mobile device transmission power of mobile UT u_k,
   - $\mathcal{P}_C^{u_k}$ is the circuit power consumption of mobile UT u_k,
   - $\mathcal{P}_H^{u_k}$ is the energy consumption of caching hardware devices of mobile UT u_k.

   Then, power consumption of UT u_k, $\forall k$ is given by:

$$\mathcal{P}^{u_k} = \beta_{UT} \times \mathcal{P}_T^{u_k} + \mathcal{P}_C^{u_k} + \mathcal{P}_H^{u_k}, \forall k \in \{1, \cdots, M\}.$$

   Neglecting the energy consumed for delivering the cache contents, the power consumption can be written as:

$$\mathcal{P}^{u_k} = \beta_{UT} \times \mathcal{P}_T^{u_k} + \mathcal{P}_C^{u_k}, \forall k \in \{1, \cdots, M\}. \tag{4}$$

   When UT u_k transmits file f_z of length f_{l_z} to UT u_i, the energy consumption can be computed as [77]:

$$E_z^{u_k} = \frac{f_{l_z}}{\mathcal{R}_{i,k}}.(\beta_{u_k} \times \mathcal{P}_T^{u_k} + \mathcal{P}_C^{u_k}), \forall k \in \{1, \cdots, M\}, \tag{5}$$

   where the data rate $\mathcal{R}_{i,k}$ of D2D communication between UT u_i and UT u_k can be calculated as follows:

$$\mathcal{R}_{i,k} = W_{i,k}^{UT} \log_2 \left(1 + \frac{\mathcal{P}_T^{u_k}.d_{i,k}^{-\alpha_{UT}}}{\sigma_{UT}^2}\right), \forall i \in \{1, \cdots, M\}, \quad \forall k \in \{1, \cdots, M\}, \tag{6}$$

   and

- $W_{i,k}^{UT}$ is the channel bandwidth from UT u_k to UT u_i,
- $d_{i,k}$ is the distance between u_i and u_k,
- σ_{UT}^2 is the average noise power for D2D communication,
- α_{UT} is the path loss factor.

2. **Energy consumption for SBS caching**
Similarly, to compute the energy consumption to transfer file f_z from SBS s_j to UT u_i, it is assumed that there is no interference between SBSs. The downlink speed $\mathcal{R}_{i,j}$ is given below:

$$\mathcal{R}_{i,j} = W_{i,j}^{SBS} \log_2 \left(1 + \frac{\mathcal{P}_T^{s_j}.d_{i,j}^{-\alpha_{SBS}}}{\sigma_{SBS}^2}\right), \forall i \in \{1, \cdots, M\}, \quad \forall j \in \{1, \cdots, N\}, \tag{7}$$

where
- $W_{i,j}^{SBS}$ is the downlink transmission bandwidth from SBS s_j to UT u_i,
- $\mathcal{P}_T^{s_j}$ is the SBS transmission power,
- $d_{i,j}$ is the distance between u_i and s_j,
- σ_{SBS}^2 is the average noise power in communication with SBS,
- α_{SBS} is the path loss factor.

Then, the components of power consumption of SBS s_j are given as follows:
- β_{SBS} is the inverse of power amplifier efficiency factor,
- $\mathcal{P}_C^{s_j}$ is the offset of site power.

When SBS s_j transmits file f_z of length f_{l_z} to UT u_i, the energy consumption can be computed as [77]:

$$E_z^{s_j} = \frac{f_{l_z}}{\mathcal{R}_{i,j}}.(\beta_{SBS} \times \mathcal{P}_T^{s_j} + \mathcal{P}_C^{s_j}). \tag{8}$$

Formulating cache system requires involving the trade off between minimizing energy consumption by caching contents at the edge of the network closer to user terminals and maximizing the probability of content popularity to place contents that will be requested by users in the near future. Caching contents requires energy to deliver the contents from MBS to SBS and UT caches. If these contents are not requested by users, and users request other contents which will be delivered again from MBS, then the energy consumed on filling SBS and UT caches was lost. The challenging problem is the adapting of a caching system to reduce power transmission by caching contents that has high probability of popularity. Table 8 illustrates power consumption-aware caching algorithms proposed for wireless networks.

6.4 Quality of Service (QoS)

The quality of service (QoS) is a network performance characteristics that is experienced by the end user. Two critical metrics can be used to refer to the QoS in MENs, they are: latency and throughput. These constraints need to be taken into account while formulating the optimization problem of caching at

Table 8 Power Consumption-Aware Caching Systems

Ref.	Contribution
[95]	- Proposed content caching for smart grid enabled wireless multimedia transmission system with optimal power allocation to users.
[77]	- Proposed an optimal transmit power of SBSs and UT in order to reduce the delivery energy cost.
[96]	- Developed a framework to minimize the total network power consumption by a joint design of adaptive BS selection, backhaul content assignment and multicast beam forming.
[97]	- Proposed optimal allocation cooperative caching scheme for industrial internet of things (IIoT) in 5G heterogeneous energy consumption.
[98]	- Formulate the optimal caching placement at the wireless the energy efficiency of heterogeneous edge that maximize wireless networks.
[99]	- Design a green content caching and mobile user–base station association mechanism in the SCNs.
[100]	- Propose two energy-efficient caching in heterogeneous networks: scalable video coding (SVC) based fractional caching and SVC-based random caching.

the edge of MEN. Table 9 and Table 10 summarize previous work on latency and throughput computation in caching scheme for wireless network.

1. Latency: In caching systems, latency refers to the average content delivery delay experienced by the end users [101]. According to cache types, latency can be classified into three types:
   (a) Average latency of delivering the requested content from another nearby UT cache through D2D communication.
   (b) Average latency of delivering the requested content from nearby SBS cache.
   (c) Average latency of delivering the requested content from nearby MBS cache.

   The latency is also refereed to as : delay, download time, and content delivery deadline. In future wireless networks, new services and applications will appear, such as augmented reality (AR) and virtual reality (VR) that have tight latency requirements than typical video streaming. Caching at the edge of the network promises to reduce latency required for requested data access and delivery. Table 11 illustrates the target requirements for different services and applications [102], [103], [104], and [105]. The reliability can be defined as the probability of successful transmission of a certain amount of data from one peer to another peer within a given deadline or time frame [106]. Additionally, storage indicates if the target service requires storage capacity for its manipulated data and the mobility indicates if the service needs processing of user terminals locations. Based on the requirements given in Table 11, latency is highly critical in most of these applications and services.
2. Throughput: In caching systems, the throughput refers to the data units that can be delivered through the network per unit time interval [101]. In MEN, this metric is used as a joint indicator of network transmission

Table 9 Latency Computation in Caching Schemes

Ref.	Contribution
[107]	- Proposed latency-centric placement and delivery strategies for cloud and cache aided wireless networks.
[88]	- Propose cooperative vehicle-aided content edge caching scheme to minimize the content delivery latency.
[108]	- Proposed hybrid content caching algorithms for joint content caching control in BSs and cloud units (CUs) subject to finite service latency.
[109]	- Proposed a joint caching and association strategy to minimize the average requested content download delay.
[110]	- Proposed an optimal cooperative content caching and delivery policy aiming to minimize the average downloading latency.
[111]	- Proposed two content caching policies: caching popular files and greedy caching in BS and D2D with the aim to minimize transmission delay.
[112]	- Proposed probabilistic caching placement-aided throughput in stochastic wireless D2D caching to measure the density of successfully served requests by local caches.
[113]	- Proposed deterministic caching algorithm and enable D2D connections based on reinforcement learning to minimizing the download latency.

Table 10 Throughput computation in Caching Schemes

Ref.	Contribution
[115]	- Proposed femtocaching and D2D collaboration to improve video throughput.
[111]	- Proposed two content caching policies: caching popular files and greedy caching in BS and D2D and investigate the behaviour of the average throughput per request.
[116]	- Proposed optimal file placement for deterministic and random caching with the aim to increase throughput for high user density wireless video network.
[113]	- Proposed deterministic caching algorithm based on reinforcement learning to maximize system throughput.

capabilities. Authors in [114] discuss throughput capability in decentralized coded and uncoded caching in a multihop D2D communication for next generation cellular networks. They illustrate the effect of using UT cache placement strategy on the increase of throughput capabilities.

6.5 Caching for Emerging Applications and Networks

Recently, new applications and services such as AR/VR, IoT, traffic monitoring, and big data processing with their requirements discussed in Sect. 6.4 have been emerged. In addition to their target requirements, these applications includes various types of sensors, are launched to be used by different types of mobile devices, and produce a wide variety of data. Therefore, MEN has been introduced with the cloud computing capabilities, IT service environment, and caching at the edge of the network to transfer the data processing and caching to the edge of the network. However, there are some points that

Table 11 Target Requirements for Different Services and Applications in Future Wireless Networks

Application/ Services	Bandwidth	Latency	Reliability	Storage	Mobility
AR/VR	1Mbps-16Mbps	$< 1ms \leq 10^-5$	High	High	High
Image/Video Editing	10Mbps	$(5-10)ms$	$\leq 10^-7$	High	High
Gaming	10Mbps	$< 1ms$	$\leq 10^-5$	High	Low
Image/Voice /Image Recognition	1Mbps -1Gbps	$< 1ms$	$\leq 10^-5$	High	High
IoT	$(1-100)Mbps$	< 20 ms	$\leq 10^-9$	High	High
Big Data	$(1-100)Mbps$	$< 20ms$	$\leq 10^-9$	High	Low
Radio/ Backhaul Optimization	$> 1Mbps$	$(100-1000)ms$	$\leq 10^-7$	Medium	Low
Traffic Monitoring/ Shaping	$(1-10)Mbps$	1000ms	$\leq 10^-9$	Medium	High

need to be considered in designing energy and latency efficient caching in MEN to overcome the problems that face emerging applications and networks:

1. Offloading tasks from mobile device with limited capabilities to the nearest mobile edge server may face delay due to congestion in communication environment in mobile edge server. In this case, task requirements (in terms of latency and reliability) will not be met. Therefore, it is important to select mobile edge server that provide communication, processing time, and storage capacity not necessarily the one with the shortest distance [117].
2. Many of these emerging applications are intelligent applications such as personalized shopping recommendation, video surveillance, intelligent personal assistant, and smart applications. Artificial intelligence (AI) applications require big data analysis. Mobile devices running these applications may suffer from limitation in device capabilities to perform high computation, poor performance, efficient energy, and limited data storage. The merge of MEN and AI is required such that MEN collaborate between edge devices and SBSs to serve users requests and AI simulate intelligent human behaviour in mobile devices by learning from previous data [118].
3. In smart industry, unmanned aerial vehicles (UAV) have been deployed to assist MEN infrastructure. UAV is a mobile device that can host SBSs storage and edge computing and has the advantage of being equipped with cameras, sensors, and devices for communication. In emergency communication scenarios that happen in MEN having dense zone characterized by large number of users generating large number of service requests. The number of SBSs in one area may fail to cover and serve users. UAVs can be used to host SBSs storage and computation capabilities. MENs based

UAVs architecture proposed to cover and serve users in challenged network situations [119].

7 Learning and Decision Technique for Optimal Caching Design

In next-generation 5G wireless networks and beyond, ultra-dense heterogeneous networks which are highly dynamic and complex, will add many challenges for network design and management. The wireless network will face huge data consumption from connected users and machines, that adds more complexity and challenges. The design of MENs that includes distribution of computational resources and storage devices in the form of local caches enables the utilization of decision theory, complex machine learning (ML), and AI approaches to providing possible solutions for the growing challenges. Developing an optimal caching system with frequent changes of input parameters (users' mobility, file requests probability, and contents popularity) with an objective to maximize network throughput, minimize power consumption, and minimize content download time, is a highly computational complexity problem. A learning and decision technique based approach allows combining reasoning, learning, prediction, and decision making algorithms to efficiently find solutions for optimal cache design.

In the literature, there are number of research work in cache developments for future wireless networks that applied learning and/or decision approaches in a specific domain in their design. Table 12 illustrates a summary of these research and the solution they provided. A brief review of some of future research directions for the development of cache systems are discussed in the following.

7.1 Decision Theory

When the problem requires to select one action from several possibilities, we will require to formulate a decision-making problem. In statistical theory, the branch that deals with such problems are called statistical decision theory or hypothesis testing [133]. In possibility theory, there are a variety of information fusion operators. Mainly they can be classified into three groups [134] as follows:

1. Conjunctive operators: Can be used for merging agreeing sources and they search for values when all the sources are agreeing.
2. Disjunctive operators: Can be used for merging conflicting sources.
3. Trade-off operators: Can be used for partially in conflict sources.

7.2 Evolutionary Approaches

Evolutionary approaches (soft-computing) can be used to solve NP-hard problems that requires hard computations. The model resulted from evolution-

Table 12 ILTs for Caching System.

Learning Technique	Paper	Technical Solution
Regularized SVD K-means clustering	[4]	- Proactively cache files based on file popularity and correlation among users. They exploit influential users in social structure of the network to cache strategic contents.
RSVD based CF and TL	[72]	- Estimate the content popularity and improve the estimation accuracy.
Deep learning	[120]	- Predict content popularity.
Extreme learning machine (ELM)	[76]	- Estimate the popularity of cache content based on the features of the content.
Deep belief network (DBN)	[121]	- Extract semantic information of user playback pattern.
Cumulative filtering	[122]	- Predict the content popularity distribution.
ML on Hadoop framework	[123]	- Estimate content popularity.
Clustering technique	[124]	- Track the evolution of content popularity over time.
Clustering technique	[125]	- Content popularity based users clustering.
Reinforcement learning	[126]	- Enabling access points to learn the optimal fetching-caching decisions.
Q-learning	[127]	- Learn, track, and adopt optimal policy.
Rank-Directed Sparse Bayesian Learning	[128]	- Estimate content popularity.
Transfer learning	[129]	- Estimate content popularity.
Deep neural network (DNN)	[130]	- Proposed caching placement and content delivering optimization algorithms.
Bayesian learning and RL	[113]	- Propose Bayesian learning method to predict personal preferences and reinforcement learning is proposed for the content placement algorithm.
Genetic algorithm	[131]	- Proposed cache placement algorithm for hierarchical collaborative caching.
Fuzzy soft set (FSS)	[132]	- Proposed fuzzy soft-set decision making for cache placement algorithm.
Q-Learning	[127]	- Proposed an optimal online caching policy.
Deep Learning	[130]	- Proposed a DNN to train an optimization problem for cache placement, user association, and content delivery.

ary approach is able to manipulate uncertainty and incomplete datasets. Two types of evolutionary approaches are used in the design of caching systems:

1. Genetic Algorithm: The most popular evolutionary strategy that can be used to solve multi objective optimization problems. The model is designed by deriving from previous generations. The individuals are allowed to reproduce and cross among themselves with a bias allocated to the fittest members. New generations result from the combinations of the most favourable characteristics of the mating members of the population. New generation is better to fit than previous generations. The control parameters of genetic algorithm are: the number of individuals in the populations, crossover probability, mutation probability, and number of generations [135]. The authors in [131] proposed a hierarchical collaborative caching strategy focusing on

content placement for 5G networks. The objective of the cache placement optimization problem is to maximize the saving in total latency of the system. The optimization problem is formulated as two sub-problems that are proved to be NP-hard. To solve the computational complexity of the problem, they used genetic placement algorithm to find approximate optimal solution.

2. Fuzzy Logic: Fuzzy logic system is used to find solutions for problems with uncertainty under membership degrees perspective. Fuzzy systems allow to represent set membership as a possible distribution. Since fuzzy theory depends on degree of membership rather than probability (likelihood), this makes fuzzy logic more effective in building fuzzy conditional inference to model uncertain information. In [132], we have proposed an algorithm for proactive caching based on fuzzy soft set (FSS) approach for decision making on file caching. The algorithm decides which files to cache and where to cache them depending on file popularity distribution, file to user preferences, file clustering, and helpers to connected users clustering. Cache placement based FSS learns the relationship between the popular files and the preferences of the files with current connected users.

7.3 Machine Learning (ML)

ML techniques model the functional relationship between input datasets and output actions with the aim to optimize some parameters. The resulted model is able to estimate an output as close as possible to the actual value. ML techniques can be categorized in two main groups: supervised and unsupervised learning depending on whether the data is labelled or not. In supervised learning, the aim is to model input and output datasets (labelled data) while unsupervised learning aims to model the hidden structure from unlabelled data sets. In caching systems, ML can be utilized to explore and extract knowledge from connected users and network characteristics to build an intelligent decision making system to make decisions for cache placement, cache access, and cache delivery options.

Some ML techniques have been applied in caching system such as: Reinforcement Learning (RL) and Deep Learning.

1. Reinforcement Learning: In reinforcement learning, the machine interacts with its dynamic environment through trial and error interactions. As a result of the interactions, the agent learns actions by receiving input of the current state of the environment and chooses the next action based on possible actions. The agent action affects (change) the state of the environment. The machine receives a value of the transition state, which can be rewards or penalties. The goal is to learn a trajectory of actions that maximize the rewards (or minimize the penalties) over its lifetime. Reinforcement learning learns the optimal policy that models environment states and actions that will maximize (or minimize) its objectives [136]. In [127], SBSs prefetch popular content during off-peak traffic hours and

send the contents to the edge of the network during peak period. The cache control unit in the SBS is designed to learn, track, and adapt to the work dynamics. The authors proposed an optimal online caching policy by developing Q-learning algorithm. The Q-learning scheme is introduced with a linear function approximation to offer fast convergence, reduce complexity, and obtain scalability over large networks

2. Deep Learning: Deep learning represents the form of learning that creates complex features by using multiple transformation steps. Much larger quantities of data are used during learning steps. Deep learning techniques show the ability to explore information included in massive data sets more effectively than traditional ML techniques. Deep learning implies learning complex artificial neural networks (ANN) that extract progressively patterns in the datasets. In traditional ANN, the three-layer perceptron (input, hidden and output layers) learns by training the hidden and output layers to adapt to the task of interest. In deep learning, more hidden layers are added to the network to subject features to the sequence of transformations. Each layer's transformation represents an inference. Modeling complex inferences can be made easier using the sequence of computational steps. The depth of the ANN represents the complexity of the learning algorithm. Some ANN learning algorithms include feedback loops. There is a number of ANN deep learning techniques such as deep multilayer perceptrons, deep convolutional neural networks (DCNNs), and recurrent neural networks (RNN) [137].
   Authors in [130] proposed a deep neural network (DNN) to train an optimization problem for cache placement, user association, and content delivery in advance and before applying these optimization algorithms in real-time caching.

8 Conclusions and Future Work

In this paper, energy and latency efficient caching in mobile edge networks (MENs) are reviewed. MEN enables the use of caching capabilities at the edge of the network in macro base station, small base stations, and user terminals. Different caching techniques are presented and compared. Then the challenges that face the design of caching system in MEN are discussed. We propose to use decision, evolutionary, and learning theoretical approaches to solve these problems. MENs also enable complex computation to be done which allows deep learning techniques to be adapted in these networks to solve problems related to energy and latency constraints.

Upon review of recent developments in the design of caching in MEN, we noted that there are several challenges in modelling and implementing caching placement, access and delivery at the edge of the network due to continuous changes in content popularity, user mobility, and number of users within the network. More challenges appear in caching at MENs due to high computation requirements of future applications that need to satisfy power and

delivery time constraints with the quality of service requirements, improved network throughput, and reduced end-to-end and backhaul delay costs. Future research work is required to investigate the development of algorithms for cache placement, cache access, and cache delivery by utilizing the data storage and computing capabilities of mobile edge networks. The main focus is on using learning and decision techniques to implement the algorithms.

In future work, we need to investigate the impact of user mobility, user activities, and cell pattern on content caching that can minimize the latency for providing the requested content to the users while on the move. More investigations are required on the impact of previous behaviour (history of file requests, cache contents, user activities, etc) and learn what can minimize latency in future user requests. The aim is to find which files to cache at SBSs and UTs to maximize the cache hit rate taking into consideration users mobility, content popularity, and cache storage capacity. Also, we need to develop cache access and cache delivery algorithms to minimize the download time and energy consumption, respectively. An efficient solution is required to build a model that is able to learn the hidden features in the input data sets, features of system attributes and their relationships, the relationship between cache placement in previous decisions, and cache access and delivery decisions to predict next decisions that may improve overall system performance. The solution approach needs to balance between the computation time and the solution quality.

References

1. Bastug E, Bennis M, Debbah M. Proactive caching in 5G small cell networks. Towards 5G: Applications, Requirements and Candidate Technologies 2016: 7898.
2. Poularakis K, Iosifidis G, Sourlas V, Tassiulas L. Exploiting caching and multicast for 5G wireless networks. IEEE Transactions on Wireless Communications 2016; 15(4): 29953007.
3. Bastug E, Bennis M, Debbah M. Living on the edge: The role of proactive caching in 5G wireless networks. IEEE Communications Magazine 2014; 52(8): 8289.
4. Bastug E,Bennis M,Debbah M. Social ands patial proactive caching for mobile data offloading. 2014IEEEInternational Conference on Communications Workshops, ICC 2014 2014: 581586.
5. KiskaniM K,Sadjadpour H R. Multihop caching-aided coded multicasting for the next generation of cellular networks. IEEE Transactions on Vehicular Technology 2017; 66(3): 25762585.
6. Goian HS, Al-Jarrah OY, Muhaidat S, Al-Hammadi Y, Yoo P, Dianati M. Popularity-based video caching techniques for cache-enabled networks: a survey. IEEE Access 2019; 7: 2769927719.
7. Li L, Zhao G, Blum RS. A survey of caching techniques in cellular networks: Research issues and challenges in content placement and delivery strategies. IEEE Communications Surveys & Tutorials 2018; 20(3): 17101732.
8. Parvez I, Rahmati A, Guvenc I, Sarwat AI, Dai H. A survey on low latency towards 5G: RAN, core network and caching solutions. IEEE Communications Surveys & Tutorials 2018.
9. Piao Z, Peng M, Liu Y, Daneshm and M. Recent advances of edge cache in radio access networks for internet of things: Techniques, performances, and challenges. IEEE Internet of Things Journal 2018; 6(1): 10101028.

10. Wang S, Zhang X, Zhang Y, Wang L, Yang J, Wang W. A survey on mobile edge networks: Convergence of computing, caching and communications. IEEE Access 2017; 5: 67576779.
11. Zhang M, Luo H, Zhang H. A survey of caching mechanisms in information-centric networking. IEEE Communications Surveys & Tutorials 2015; 17(3): 14731499.
12. Mehrabi M, You D, Latzko V, Salah H, Reisslein M, Fitzek FH. Device-Enhanced MEC: Multi-Access Edge Computing (MEC) Aided by End Device Computation and Caching: A Survey. IEEE Access 2019; 7: 166079166108.
13. Rahimi MR, Ren J, Liu CH, Vasilakos AV, Venkatasubramanian N. Mobile Cloud Computing: A survey, state of art and future directions. Mobile Networks and Applications 2014; 19(2): 133143.
14. Abbas N, Zhang Y, Taher kordi A, Skeie T. Mobile edge computing: Asurvey. IEEE Internet of Things Journal 2018; 5(1): 450465.
15. Hu YC, Patel M, Sabella D, Sprecher N, Young V. Mobile edge computingaATA key technology towards 5G. ETSI white paper 2015; 11(11): 116.
16. Chiang M, Zhang T. Fog and IoT: An overview of research opportunities. IEEE Internet of Things Journal 2016; 3(6): 854864.
17. Satyanarayanan M, Bahl P, Caceres R, Davies N. The case for vm-based cloudlets in mobile computing. IEEE pervasive Computing 2009; 8(4).
18. Pang Z, Sun L, Wang Z, Tian E, Yang S. A survey of cloudlet based mobile computing. In: IEEE. ; 2015: 268275.
19. Liu Y,Point JC, Katsaros KV, Glykantzis V, Siddiqui MS, Escalona E. SDN/NFV based caching solution for future mobile network (5G). In: IEEE. ; 2017: 15.
20. Chen Q, Yu FR, Huang T, Xie R, Liu J, Liu Y. Joint resource allocation for software-defined networking, caching, and computing. IEEE/ACM Transactions on Networking 2018; 26(1): 274287.
21. Huo R, Yu FR, Huang T, et al. Software defined networking, caching, and computing for green wireless networks. IEEE Communications Magazine 2016; 54(11): 185193.
22. Zhang X, Zhu Q. Distributed mobile devices caching over edge computing wireless networks. In: IEEE. ; 2017: 127132.
23. Taleb T, Dutta S, Ksentini A, Iqbal M, Flinck H. Mobile edge computing potential in making cities smarter. IEEE Communications Magazine 2017; 55(3): 3843.
24. Ko H, Lee J, Pack S. Spatial and temporal computation offloading decision algorithm in edge cloud-enabled Heterogeneous Networks. IEEE Access 2018; PP(99): 1-1.
25. Kumar K, Liu J, Lu YH, Bhargava B. A survey of computation offloading for mobile systems. Mobile Networks and Applications 2013; 18(1): 129140.
26. Hao Y, Chen M, Hu L, Hossain MS, Ghoniem A. Energy efficient task caching and offloading for mobile edge Computing. IEEE Access 2018.
27. Mao Y, You C, Zhang J, Huang K, Letaief KB. A survey on mobile edge computing: The communication perspective. IEEE Communications Surveys & Tutorials 2017; 19(4): 23222358.
28. Cui Y, Jiang D. Analysis and optimization of caching and multi casting in large-scale cache-enabled heterogeneous wireless networks. IEEE transactions on Wireless Communications 2017; 16(1): 250264.
29. Ding W, Qi W, Wang J, Chen B. OpenSCaaS: an open service chain as a service platform toward the integration of SDN and NFV. IEEE Network 2015; 29(3): 3035.
30. Hung CH, Hsieh YC, Wang LC. Control plane latency reduction for service chaining in mobile edge computing system. In: IEEE. ; 2017: 15.
31. Chen M, Qian Y, Hao Y, Li Y, Song J. Data-driven computing and caching in 5G networks: Architecture and delay analysis. IEEE Wireless Communications 2018; 25(1): 7075.
32. Ranadheera S, Maghsudi S, Hossain E. Computation offloading and activation of mobile edge computing servers: A minority game. IEEE Wireless Communications Letters 2018.
33. Yu S, Wang X, Langar R. Computation off loading for mobile edge computing: A deep learning approach.In:IEEE.;2017: 16.
34. Guo H, Liu J, Qin H, Zhang H. Collaborative computation offloading for mobile-edge computing over fiber-wireless networks. In: IEEE. ; 2017: 16.
35. Luo C, Salinas S, Li M, Li P. Energy-efficient autonomic offloading in mobile edge computing. In: IEEE. ; 2017: 581588.

36. Deng M, Tian H, Lyu X. Adaptive sequential offloading game for multi-cell mobile edge computing. In:IEEE.; 2016: 15.
37. Perabathini B, Bastug E, Kountouris M, Debbah M, Conte A. Caching at the edge: a green perspective for 5G networks. In: IEEE. ; 2015: 28302835.
38. Lee MC, Molisch AF. Individual Preference Aware Caching Policy Design for Energy-Efficient Wireless D2D Communi- cations. In: IEEE. ; 2017: 17.
39. Zhang J, Zhang X, Yan Z, Li Y, Wang W, Zhang Y. Social-aware cache information processing for 5G ultra-dense networks. In: IEEE. ; 2016: 15.
40. Gregori M, Gomez-Vilardebo J, Matamoros J, Gunduz D. Wireless content caching for small cell and D2D networks. IEEE Journal on Selected Areas in Communications 2016; 34(5): 12221234.
41. Cui Y, He W, Ni C, Guo C, Liu Z. Energy-efficient resource allocation for cache-assisted mobile edge computing. arXiv preprint arXiv:1708.04813 2017.
42. Liang C, He Y, Yu FR, Zhao N. Energy-efficient resource allocation in software-defined mobile networks with mobile edge computing and caching. In: IEEE. ; 2017: 121126.
43. Mrad S, Hamouda S, Rezig H Graph Theory based multi cast caching for better energy saving in dense small cell networks. In: IEEE. ; 2017: 20152020.
44. Zhang J, Xia W, Yan F, Shen L. Joint computation offloading and resource allocation optimization in heterogeneous networks with mobile edge computing. IEEE Access 2018; PP(99): 1-1.
45. Wu D, Liu Q, Wang H, Wu D, Wang R. Socially aware energy-efficient mobile edge collaboration for video distribution. IEEE Transactions on Multimedia 2017; 19(10): 21972209.
46. Liu J, Zhang Q. Offloading schemes in mobile edge computing for ultra-reliable low latency communications.IEEE Access 2018(99): 113.
47. Liu CF, Bennis M, Poor HV. Latency and reliability-aware task offloading and resource allocation for mobile edge computing. arXiv preprint arXiv:1710.00590 2017.
48. Mao Y, Zhang J, Letaief KB. Dynamic computation off loading for mobile-edge computing with energy harvesting devices. IEEE Journal on Selected Areas in Communications 2016; 34(12): 35903605.
49. Xu J, Chen L, Ren S. Online learning for offloading and auto scaling in energy harvesting mobile edge computing. IEEE Transactions on Cognitive Communications and Networking 2017; 3(3): 361373.
50. Hou T, Feng G, Qin S, Jiang W. Proactive content caching by exploiting transfer learning for mobile edge computing. In: IEEE. ; 2017: 16.
51. Poderys J, Artuso M, Lensbl CMO, Christiansen HL, Soler J. Caching at the Mobile Edge: A Practical Implementation. IEEE Access 2018; 6: 86308637.
52. Xu J, Palanisamy B, Ludwig H, Wang Q. Zenith: Utility-aware resource allocation for edge computing. In: IEEE. ; 2017: 4754.
53. Zhu Z, Peng J, Gu X, et al. Fair Resource Allocation for System Throughput Maximization in Mobile Edge Computing. IEEE Access 2018; 6: 53325340.
54. Kiskani MK, Vakilinia S, Cheriet M. Popularity based file categorization and coded caching in 5G networks. In: IEEE.; 2017: 15.
55. Muller S, Atan O, Schaar v.dM, Klein A. Context-aware proactive content caching with service differentiation in wireless networks. IEEE Transactions on Wireless Communications 2017; 16(2): 10241036.
56. Leconte M, Paschos G, Gkatzikis L, Draief M, Vassilaras S, Chouvardas S. Placing dynamic content in caches with small population. In: ; 2016: 1-9.
57. MAijller S, Atan O, Schaar v. dM, Klein A. Smart caching in wireless small cell networks via contextual multi-armed bandits. In: ; 2016: 1-7.
58. Shanmugam K, Golrezaei N, Dimakis AG, Molisch AF, Caire G. Femto caching: Wireless content delivery through distributed caching helpers. IEEE Transactions on Information Theory 2013; 59(12): 84028413.
59. Vo NS, Duong TQ, Guizani M. QoE-oriented resource efficiency for 5G two-tier cellular networks: A femtocaching framework. In: IEEE. ; 2016: 16.
60. Golrezaei N, Molisch AF, Dimakis AG, Caire G. Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution. IEEE Communications Magazine 2013; 51(4): 142149.

61. Vo NS, Duong TQ, Guizani M. QoE-oriented resource efficiency for 5G two-tier cellular networks: A femto caching framework. In: ; 2016: 1-6.
62. Sengupta A, Amuru S, Tandon R, Buehrer RM, Clancy TC. Learning distributed caching strategies in small cell networks. In: ; 2014: 917-921.
63. Muller S, Atan O, Schaar v. dM, Klein A. Smart caching in wireless small cell networks via context ual multi-armedbandits. In: IEEE. ; 2016: 17.
64. Sengupta A, Amuru S, Tandon R, Buehrer RM, Clancy TC. Learning distributed caching strategies in small cell networks. In: IEEE. ; 2014: 917921.
65. MAijller S, Atan O, Schaar v. dM, Klein A. Context-aware proactive content caching with service differentiation in wireless networks. IEEE Transactions on Wireless Communications 2017; 16(2): 1024-1036.
66. Li S, Xu J, Van Der Schaar M, Li W. Popularity-driven content caching. In: IEEE. ; 2016: 19.
67. Pantisano F, Bennis M, Saad W, Debbah M. Cache-aware user association in backhaul-constrained small cell networks. In: IEEE. ; 2014: 3742.
68. Bastug E, Bennis M, Debbah M. Social and spatial proactive caching for mobile data offloading. In :IEEE.;2014:581586.
69. Rao J, Feng H, Yang C, Chen Z, Xia B. Optimal caching placement for D2D assisted wireless caching networks. In:IEEE. ; 2016: 16.
70. Zhang X, Wang Y, Sun R, Wang D. Clustered device-to-device caching based on file preferences. In: IEEE: 16.
71. Khreishah A, Chakareski J, Gharaibeh A. Joint caching, routing, and channel assignment for collaborative small-cell cellular networks. IEEE Journal on Selected Areas in Communications 2016; 34(8): 2275-2284.
72. Wang Y, Chen Y, Dai H, Huang Y, Yang L. A learning-based approach for proactive caching in wireless communication networks. In: ; 2017: 1-6.
73. Zheng Z, Song L, Han Z, Li GY, Poor HV. A Stackelberg Game Approach to Proactive Caching in Large-Scale Mobile Edge Networks. IEEE Transactions on Wireless Communications 2018: 1-1.
74. Doan KN, Van Nguyen T, Quek TQ, Shin H. Content-aware proactive caching for backhaul offloading in cellular network. IEEE Transactions on Wireless Communications 2018; 17(5): 31283140.
75. Ahlehagh H, Dey S. Video-aware scheduling and caching in the radio access network. IEEE/ACM Transactions on Networking (TON) 2014; 22(5): 14441462.
76. Tanzil SS, Hoiles W, Krishnamurthy V. Adaptive scheme for caching youtube content in a cellular network: Machine learning approach. Ieee Access 2017; 5: 58705881.
77. Chen M, Hao Y, Hu L, Huang K, Lau VK. Green and mobility-aware caching in 5G networks. IEEE Transactions on Wireless Communications 2017; 16(12): 83478361.
78. Tan Y, Yuan Y, Yang T, Xu Y, Hu B. Femtocaching in wireless video networks: Distributed framework based on exact potential game. In: IEEE; 2016: 16.
79. Shanmugam K, Golrezaei N, Dimakis AG, Molisch AF, Caire G. Femtocaching: Wireless content delivery through distributed caching helpers. IEEE Transactions on Information Theory 2013; 59(12): 84028413.
80. Wang R, Peng X, Zhang J, Letaief KB. Mobility-aware caching for content-centric wireless networks: Modeling and methodology. IEEE Communications Magazine 2016; 54(8): 7783.
81. Yang L, LiuT,Hu Q, Liu S, Huang H. Empirical analysis on temporal statistics of pairwise contact patterns in dynamic human networks. In: IEEE. ; 2017: 916.
82. Wang T, Song L, Han Z. Dynamic femtocaching for mobile users. In: IEEE. ; 2015: 861865.
83. Poularakis K, Tassiulas L. Code, cache and deliver on the move: A novel caching paradigm in hyper-dense small-cell networks. IEEE Transactions on Mobile Computing 2017; 16(3): 675687.
84. Guan Y, Xiao Y, Feng H, Shen CC, Cimini LJ. MobiCacher: Mobility-aware content caching in small-cell networks. In: IEEE. ; 2014: 45374542.
85. Liu X, Zhang J, Zhang X, Wang W. Mobility-aware coded probabilistic caching scheme for MEC-enabled small cell networks. IEEE Access 2017; 5: 1782417833.
86. Lan R, Wang W, Huang A, Shan H. Device-to-device offloading with proactive caching in mobile cellular networks. In: IEEE. ; 2015: 16.

87. Wang R, Zhang J, Song S, Letaief KB. Mobility-aware caching in D2D networks. IEEE Transactions on Wireless Communications 2017; 16(8): 50015015.
88. Zhang K, Leng S, He Y, Maharjan S, Zhang Y. Cooperative Content Caching in 5G Networks with Mobile Edge Computing. IEEE Wireless Communications 2018; 25(3): 8087.
89. Wang R, Zhang J, Song S, Letaief KB. Exploiting mobility in cache-assisted D2D networks: Performance analysis and optimization. arXiv preprint arXiv:1806.04069 2018.
90. Deng T, Ahani G, Fan P, Yuan D. Cost-optimal caching for D2D networks with user mobility: Modeling, analysis, and computational approaches. IEEE Transactions on Wireless Communications 2018; 17(5): 30823094.
91. Chen M, Hao Y, Qiu M, Song J, Wu D, Humar I. Mobility-aware caching and computation offloading in 5G ultra-dense cellular networks. Sensors 2016; 16(7): 974.
92. Wang Y, Chen Y, Dai H, Huang Y, Yang L. A learning-based approach for proactive caching in wireless communication networks. In: IEEE. ; 2017: 16.
93. Nguyen QN, Arifuzzaman M, Yu K, Sato T. A Context-Aware Green Information-Centric Networking Model for Future Wireless Communications. IEEE Access 2018; 6: 2280422816.
94. Abou-Zeid H, Hassanein H. Toward green media delivery: location-aware opportunities and approaches. IEEE Wireless Communications 2014; 21(4): 3846.
95. Huang X, Ansari N. Content caching and distribution in smart grid enabled wireless networks. IEEE Internet of Things Journal 2017; 4(2): 513520.
96. Peng X, Shi Y, Zhang J, Letaief KB. Layered Group Sparse Beamforming for Cache-Enabled Green Wireless Networks. IEEE Transactions on Communications 2017; 65(12): 55895603.
97. Duan P,Jia Y, Liang L, Rodriguez J, Huq KMS, Li G. Space-reserved cooperative caching in 5G heterogeneous networks for Industrial IoT. IEEE Transactions on Industrial Informatics 2018; 14(6): 2715.
98. Gabry F, Bioglio V, Land I. On energy-efficient edge caching in heterogeneous networks. IEEE Journal on Selected Areas in Communications 2016; 34(12): 32883298.
99. Guo F, Zhang H, Li X, Ji H, Leung VC. Joint optimization of caching and association in energy harvesting powered small cell networks. IEEE Transactions on Vehicular Technology 2018.
100. Zhang X, Lv T, Ni W, Cioffi JM, Beaulieu NC, Guo YJ. Energy-efficient caching for scalable videos in heterogeneous networks. 2018.
101. Yu FR, Huang T, Liu Y. Integrated networking, caching, and computing. CRC Press . 2018.
102. Intel . Intel 5G: A network transformation imperative. 2015.
103. Parvez I, Rahmati A, Guvenc I, Sarwat AI, Dai H. A survey on low latency towards 5G: RAN, core network and caching solutions. CoRR 2017.
104. Suppliers association mG, others . The road to 5G: Drivers, applications, requirements and technical development. 2015.
105. Qi Y, Hunukumbure M, Nekovee M, Lorca J, Sgardoni V. Quantifying data rate and bandwidth requirements for immersive 5G experience. In: IEEE. ; 2016: 455461.
106. Popovski P. Ultra-reliable communication in 5G wireless systems. In: IEEE. ; 2014: 146151.
107. Sengupta A, Tandon R, Simeone O. Fog-aided wireless networks for content delivery: Fundamental latency trade-offs. 2016. 2017.
108. Kwak J, Kim Y, Le LB, Chong S. Hybrid content caching in 5G wireless networks: Cloud versus edge caching. IEEE Transactions on Wireless Communications 2018; 17(5): 30303045.
109. Wang Y, Tao X, Zhang X, Mao G. Joint caching placement and user association for minimizing user download delay. IEEE Access 2016; 4: 86258633.
110. Jiang W, Feng G, Qin S. Optimal cooperative content caching and delivery policy for heterogeneous cellular networks. IEEE Transactions on Mobile Computing 2017(1): 11.
111. Amer R, Butt MM, Bennis M, Marchetti N. Inter-cluster cooperation for wireless D2D caching networks. IEEE Transactions on Wireless Communications 2018; 17(9): 61086121.
112. Chen Z, Pappas N, Kountouris M. Probabilistic caching in wireless D2D networks: Cache hit optimal versus throughput optimal. IEEE Communications Letters 2017; 21(3): 584587.

113. Cheng P, Ma C, Ding M, etal. Localized small cell caching: A machine learning approach based on rating data. IEEE Transactions on Communications 2018.
114. Kiskani MK, Sadjadpour HR. Throughput analysis of decentralized coded content caching in cellular networks. IEEE Transactions on Wireless Communications 2017; 16(1): 663672.
115. Golrezaei N, Molisch AF, Dimakis AG, Caire G. Femtocaching and device-to-device collaboration: A new architecture for wireless video distribution. IEEE Communications Magazine 2013; 51(4): 142-149.
116. Golrezaei N, Mansourifard P, Molisch AF, Dimakis AG. Base-station assisted device-to-device communications for high- throughput wireless video networks. IEEE Transactions on Wireless Communications 2014; 13(7): 36653676.
117. Kim K, Hong CS. Optimal task-UAV-edge matching for computation offloading in UAV assisted mobile edge computing. In: IEEE. ; 2019: 14.
118. Zhou Z, Chen X, Li E, Zeng L, Luo K, Zhang J. Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE 2019; 107(8): 17381762.
119. Narang M, Xiang S, Liu W, et al. UAV-assisted edge infrastructure for challenged networks. In: IEEE. ; 2017: 6065.
120. Liu WX, Zhang J, Liang ZW, Peng LX, Cai J. Content popularity prediction and caching for ICN: A deep learning approach with SDN. IEEE access 2018; 6: 50755089.
121. Hao H, Xu C, Wang M, Xie H, Liu Y, Wu DO. Knowledge-centric proactive edge caching over mobile content distribution network. In: IEEE. ; 2018: 450455.
122. Bastug E, Bennis M, Zeydan E, et al. Big data meets telcos: A proactive caching perspective. Journal of Communications and Networks 2015; 17(6): 549557.
123. Kader MA, Bastug E, Bennis M, et al. Leveraging big data analytics for cache-enabled wireless networks. In: IEEE. ; 2015: 16.
124. Paschos G, Bastug E, Land I, Caire G, Debbah M. Wireless caching: Technical misconceptions and business barriers. IEEE Communications Magazine 2016; 54(8): 1622.
125. Hajri SE, Assaad M. Energy efficiency in cache-enabled small cell networks with adaptive user clustering. IEEE Transactions on Wireless Communications 2018; 17(2): 955968.
126. Sadeghi A, Sheikholeslami F, Matrques AG, Giannakis GB. Reinforcement learning for 5G caching with dynamic cost. In: IEEE. ; 2018: 66536657.
127. Sadeghi A, Sheikholeslami F, Giannakis GB. Optimal and scalable caching for 5G using reinforcement learning of space-time popularities. IEEE Journal of Selected Topics in Signal Processing 2018; 12(1): 180190.
128. Mishra SK, Pandey P, Arya P, Jain A. Efficient proactive caching in storage constrained 5G small cells. In: IEEE. ; 2018: 291296.
129. Hou T, Feng G, Qin S, Jiang W. Proactive content caching by exploiting transfer learning for mobile edge computing. International Journal of Communication Systems 2018; 31(11): e3706.
130. Lei L, You L, Dai G, Vu TX, Yuan D, Chatzinotas S. A deep learning approach for optimizing content delivering in cache-enabled HetNet. In: IEEE. ; 2017: 449453.
131. Tang Q, Xie R, Huang T, Liu Y. Hierarchical collaborative caching in 5G networks. IET Communications 2018; 12(18): 23572365.
132. Mohammed L, Jaseemuddin M, Anpalagan A. Fuzzy soft-set based approach for femtocaching in wireless networks. In: IEEE. ; 2018.
133. Cheng Q, Niu R, Sundaresan A, VARSHNEY PK. Distributed detection and decision fusion with applications to wireless sensor networks. INTEGRATED TRACKING, CLASSIFICATION, AND SENSOR MANAGEMENT 2013: 619.
134. Roux L. An application of possibility theory information fusion to satellite image classification. In: Springer. ; 1997: 166179.
135. Lobato FS, Steffen Jr V. Multi-objective optimization problems: Concepts and self-adaptive parameters with mathematical and engineering applications. Springer. 2017.
136. Chen Z, Liu B. Life long machine learning. Synthesis Lectures on Artificial Intelligence and Machine Learning2016;10(3): 1145.
137. Witten IH, Frank E, Hall MA, Pal CJ. Data mining: Practical machine learning tools and techniques. Morgan Kaufmann . 2016.