Jin et al. EURASIP Journal on Wireless Communications and Networking (2017) 2017:33
DOI 10.1186/s13638-017-0806-6

EURASIP Journal on Wireless Communications and Networking

REVIEW Open Access

Information-centric mobile caching network frameworks and caching optimization: a survey

Hao Jin*, Dan Xu, Chenglin Zhao and Dong Liang

Abstract

The demand for content oriented service and compute-intensive service stimulates the shift of current cellular networks to deal with the explosive growth in mobile traffic. Information centric mobile caching network architectures have emerged in Information-Centric Networking as well as mobile cellular and ad-hoc networks deployed with caches. Caching optimization based on information centric mobile caching has become the key issue, and several significant research challenges remain to be addressed before its widespread adoption. In this paper, a brief survey on Information centric mobile caching network architecture and caching optimization is presented, including cache placement in different mobile wireless network architectures, the taxonomy of cache insertion and eviction policies, the modeling behavior of caching networks as well as caching optimization based on network centric and user centric metrics, and typical applications based on mobile caching. Finally, the research directions and open challenges are investigated.

Keywords: Mobile Caching, Information-Centric Networking, Mobile wireless network architecture, Cache policy

1 Introduction

With tremendous numbers of smart phones, laptops and tablets, more and more users demand for content oriented service and compute-intensive service. Such an explosive growth in mobile traffic really leads to a significant paradigm shift in current cellular networks. Among the new service paradigms, contents based mobile multimedia videos which would be implemented by duplicating and distributing a few popular large size contents to mobile devices become an important portion of the mobile traffic. Since most users are mainly interested in accessing vast amount of information instead of physical location, the paradigm shift in the usage model of the Internet leads to the investigation of new networking paradigm, namely Information-Centric Networking (ICN) [1].

ICN shifts internet usage from a sender-driven end-to-end communication paradigm to a receiver-driven content retrieval [2]. The architecture of ICN supports transparent and ubiquitous in-network caching to speed up content distribution and provides users with mobility and flexibility in accessing and generating information.

The quality of experience of users would be affected during the time when they are downloading and caching contents due to mobility. As a result, integration of the ICN paradigm and mobile wireless networks is a promising direction for future mobile networks [3]. In order to incorporate ICN with mobile networks, in-network caching strategy is one of the crucial issues since it would influence the performance and leverage the cost of mobile wireless networks. On the one hand, caching decreases the average user-perceived delay and the redundant traffic load by caching duplicates of popular contents in the local routers or access points (APs), and in turn relieves the pressure on remote servers. On the other hand, according to the prediction of Moore's law (and recently by Kryder's law), the capacity of storage units has increased exponentially over the past thirty years with consistently declining costs per stored bit [4]. The cheap price of caches also contributes to the deployment of information centric caches on the edge of networks, thus the deployment of caching resources on edge equipments (e.g. smart mobile devices) gains

* Correspondence: hjin@bupt.edu.cn
The Key Laboratory of Universal Wireless Communications for Ministry of Education, Beijing University of Posts and Telecommunications, Beijing 100876, China

attention, and information centric mobile caching becomes a significant challenge for providing content oriented services.

In order to enable mobile users to access nearby network caching elements such as servers or mobile devices for popular contents, effective mobile caching strategies are provided to reduce duplicating content transmissions by adopting intelligent caching strategies inside mobile networks. These mobile caching strategies can be refined on three key issues: namely cache placement selection, cache policy design and cache content selection.

Cache placement selection refers to where to cache contents in the mobile wireless networks. In fact, caches can be deployed in various network elements depending on network architectures. Base stations, APs and mobile nodes can be equipped with caches to support content exchange. That is to say, the issue of cache placement would affect the framework of mobile wireless networks.

Cache policy design focuses on how to cache contents in cache-enabled mobile networks. It refers to different policies on caching insertion and caching eviction. Caching insertion policies decide whether to cooperate with other caching nodes and how contents or chunks can be stored in the caches, and caching eviction policies concern the dynamics of contents in the caches.

Cache content selection involves what contents to be cached, which are to be updated. Content popularity is often used as an important factor for accelerating content retrieval, while content diversity for increasing the types of contents cached locally means that even popular contents should not be cached in multiple local caching nodes.

Since the resources in the network are constrained, which include cache, computing, energy and transmission bandwidth resources, it is of great significance to optimize what contents to cache and how to insert and evict contents from the caches considering both content popularity, content diversity and node mobility in order to achieve ideal performances with various optimization objectives. Caching optimization deals with the problems from the point of view of optimization on network/user centric performances based on different network architectures, analytical modeling methods and content caching policies.

This paper aims to present a survey on information-centric mobile caching and its optimization. The main content is summarized as follows: (1) Focusing on cache placement selection, the cache placement in different mobile wireless network architectures are analyzed; (2) On cache policy design, taxonomy of cache insertion and eviction policies is provided, and state-of-the-art methods are investigated; (3) On caching optimization, caching behavior modeling is illustrated, then network centric and user centric optimization issues are surveyed.

The remainder of the paper is organized as follows. In section 2, ICN and some mobile network caching architectures are introduced, and the comparison of caching placement methods in different mobile wireless network architectures is given. In section 3, content cache policies are illustrated in detail, including caching insertion policies and caching eviction policies. In section 4, research issues on network centric and user centric caching optimization are introduced, including modeling behavior of the caching system, related analytical modeling methods as well as simulation. In section 5, some content based applications are given. In section 6, several research challenges and open issues are provided. Section 7 concludes the paper.

2 Mobile wireless caching frameworks

ICN supports in-network caching and decouples information from its sources as an appropriate approach to achieve information distribution and mobility support. The ubiquitous in-network caching opens up many opportunities for exploiting content awareness in order to place information closer to the user due to its explicit naming of content rather than communication endpoints. Along with the dramatic developments of mobile wireless networks for information centric services, evolutional architectures and procedures for future access networks based on ICN become a promising direction to promote the integration of cellular and wireless access technologies with ICN [3, 5].

In this section, different caching frameworks are investigated, including caching frameworks based on ICN, mobile cellular network, mobile ad hoc network and hybrid network.

2.1 Caching Framework based on Information-Centric Networking

The networking paradigm of current Internet is host based, and the information exchanges are realized by network address based routing. The host-centric Internet can hardly meet the increasing demands of content based services supported by mobile end users since smart end devices and various multimedia applications are deployed. ICN has shifted the current complex Internet model to a simple one based on named contents and it enables named content routing through publisher-subscriber driven communication. In ICN, end-users express their interests for a given content by sending content requests, and the entire network is responsible for routing the requests based only on the content name towards the best content caches and delivering the content through the reverse paths to the end-users [6]. The key technologies of ICN include information naming, name-based routing, in-networking caching and related caching strategies, native multicast, self-secured content as

well as mobility management of consumers and publishers. The main elements in ICN are content publishers, end users (content subscribers) and cache-enabled content routers (CRs).

For information naming in ICN, the content name is the only identifier for content objects, which permits either the end-user or the intermediate networking unit to locate the best content holder. Besides the fundamental property of identifying the different content, information naming is globally unique, location independent, self-defined and security intergraded. Hierarchical naming and flat naming are two categories of ICN naming structures which are mainly supported in existing ICN solutions [6].

The name-based ICN routing can be realized by the undirected naming resolution service and the direct name based routing. The naming resolution requires one or several centralized servers (e.g. RendezVous points, register servers, trigger points, etc.) in the networking topology. The content publications are collected in these servers, which have a global view of all the published content objects and the networking topology. When an ICN router wants to forward a request message, the routing path is calculated in the centralized server by existing routing protocol. The name based routing is directly performed in the ICN routers. Each router has a local forwarding information base which is filled by the content publication messages. The request forwarding paths are calculated in the local routers by their own forwarding strategies [6].

Mobility and caching are realized by information naming, name resolution and data routing in ICN. In ICN, caching is ubiquitous. On-path and off-path caching are supported by name resolution request process. In on-path caching, the network places information cached along the path taken by a name resolution request, while in off-path caching, the network exploits information cached outside that path. For mobility support, subscriber mobility and publisher mobility are distinguished. Subscriber mobility is intrinsically supported in ICN architectures, since mobile subscribers can just send new subscriptions for information after a handoff. Publisher mobility is more difficult to support compared to subscriber mobility, because the name resolution system (in the coupled approach) or the routing tables (in the decoupled approach) need to be updated [1].

Figure 1 gives the abstract architecture of caching framework based on ICN, in which caches are usually placed at every CR.

Some ICN architectures are constructed since 2007 [1, 7] including DONA, 4WARD, PSIRP, COMET, CONVERGENCE, NDN/CCN, PURSUIT, SAIL, MobilityFirst, ANR Connect, CBMEN, NDN-NP, MobilityFirst-NP, Green-ICN, I-CAN, COMIT, POINT, UMOBILE, RIFE, BONVOYAGE, etc. [8–14].

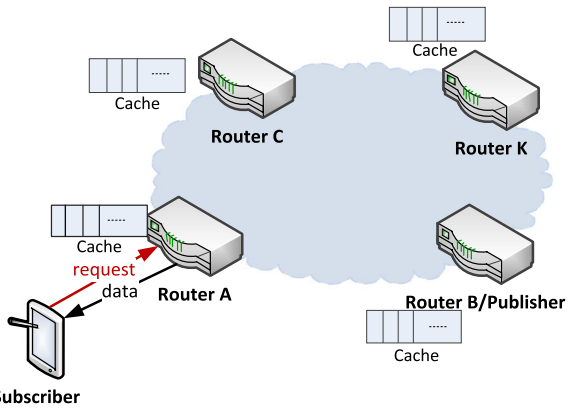

Fig. 1 Caching Framework based on Information-Centric Networking

From the perspective of relevant projects, most of the projects which have launched from 2007 to 2012 focus on theoretical research such as architecture, naming, routing, forwarding and security. While the majority of the projects launched since 2013 are concentrated on specific application scenarios to enhance the architectures, such as Internet access, opportunistic communications as well as multimodal transport. The business models for ICN and the migration strategies are also concentrated which are compatible with TCP/IP, such as IP over ICN. Consequently, the study on ICN has been shifted from theoretical research to some specific applications practice [7]. Since the main topic of this paper is on mobile caching framework and caching optimization, and information centric caching and its mobility support mechanisms are distinct depending on various ICN architectures, in this section, caching and mobility support are surveyed based on several typical ICN architectures.

DONA [1, 15] is one of the first complete ICN architectures with flat names proposed by UC Berkeley. DONA supports on-path caching by Resolution Handler (RH) infrastructure. A RH which decides to cache a requested data object can replace the source IP address of an incoming FIND request with its own IP address before it forwards the message to the next RH. Once the response traverses the current RH, the data returned would be cached. If path-labels are used, the data always return via the intermediate RHs, which can make a caching decision or not. If a subsequent FIND message requesting the same object reaches a caching RH, the RH can directly return the data to the subscriber. Information may also be replicated off-path if the information through its local RH is registered. A RH receiving multiple REGISTER messages for the same information maintains (and propagates upwards) only the pointers to the best available copy. Mobile subscribers can simply send new FIND messages from their current location relying on the RH infrastructure to provide them with

the closest copy of the information. Mobile publishers can also unregister and re-register their information when changing their network location [1].

CCN [1, 16] architecture is the other fully-fledged ICN architecture presented by PARC. NDN [17] project funded by the US Future Internet Architecture project develops the CCN architecture. NDN supports on-path caching, since each CR first consults its Content Store (CS) whenever it receives an INTEREST message and caches all information objects carried by DATA messages. A CS just caches every incoming Data packets [18, 19] temporarily in order to avoid packet losses and satisfy requests for the same data subsequently. NDN supports Off-path caching by delivering an INTEREST to any data source that may be hosting the requested information object. Subscriber mobility is implemented by sending new INTEREST messages from its current location for the information objects it has not received yet. Publisher mobility is enabled by Forwarding Information Base which requires advertising the name prefixes for the information it is hosting via the routing protocol again [1].

As one of the important architectures proposed by EU Framework 7 Programme, PURSUIT constructs architecture with a publisher-subscriber protocol stack. PURSUIT supports both on-path and off-path caching. In the on-path caching, forwarded packets are cached at Forwarding Nodes to serve subsequent requests. However, on-path caching may not be very effective because name resolution is decoupled from data routing. In the off-path caching, caches act as publishers by advertising the available information to the Rendezvous Network. Mobility is facilitated by multicast and caching [20]. Four types of mobility cases are considered. Local subscriber mobility can be handled via multicast and caching, while Publisher mobility is harder to support since the publisher's new position information in the network need to be notified to the topology management function. In PURSUIT, Publish-Subscribe systems are organized as a collection of autonomous components [21], so clients act either as publishers who publish new events in the network or as subscribers who subscribe events they are interested in.

The SAIL architecture supports on-path caching at the CRs. In SAIL, the caches are regarded as publishers by envisaging the deployment of large scale information object caching and replication mechanisms cooperative with the Name Resolution System (NRS). SAIL considers a hierarchy of caches in which local caches are part of a tree which includes a small number of caching servers at the root. The caches which are in the higher level of the hierarchy have larger storage space in order to store popular objects, which would have been evicted by local caches because of their limited storage size. Cache replacement policies are also investigated on which popular objects are dynamically moved to caches that are closer to the consumers. Host mobility is enabled by maintaining topological information for each registered host by the NRS. Upon a change of location, the moving host updates the topological information in the NRS where it is registered and a notification is issued to the nodes which are communicating with the mobile host [1, 22].

COMET supports on-path and off-path caching. On-path caching originates from name resolution. Registering cached copies with the Content Resolution System is required in Off-path caching. In COMET, two novel schemes are proposed which are different from NDN's "cache everything" on-path caching, one is the Prob-Cache, which is a probabilistic caching scheme [18], the other is the Centrality scheme based on the observation that Content-aware Routers (CaRs) locating on shortest paths are more likely to obtain a cache hit, thus an information object should only be cached by the CaR with the highest centrality in its path. Mobility-aware CaRs are deployed at the edge of the access networks to support user mobility, and track the mobility of users and context information and can predict their future locations [1].

In CONVERGENCE, on-path caching is supported in a manner similar to NDN. Off-path caching and replication are facilitated by registering additional copies of an information object stored at Internal Nodes to the NRS. CONVERGENCE supports subscriber mobility via new requests as in NDN, and Publisher mobility relies on unspecified name resolution system [1, 23].

MobilityFirst supports on-path caching passing messages at intermediate CRs opportunistically, and allows subsequent requests for the same Globally Unique Identifier (GUID) to be replied with the locally cached copy. In addition, once an information object is cached off-path or replicated, the Global Name Resolution Service (GNRS) is notified of the change in order to update the corresponding GUID entry with the additional network addresses. Despite of the "slow-path" operation which the GNRS can be repeatedly consulted as a message traverses the network, each CR can adopt its own policy and decide when to consult with the GNRS for additional cached copies. Addressing host, information, and entire network mobility is another important objective in MobilityFirst. Host mobility is primarily handled by GNRS when a network attached object changes its point of attachment. Network mobility is also supported. Mobility causes disconnections and variable link conditions in networks, which can be solved by a storage-aware routing mechanism exploited at the intra-domain level by deploying local storage as in delay-tolerant networking [1, 24].

The method of information centric caching and mobility support of the typical ICN architectures are

summarized. Besides, some performances of the architectures are investigated based on the ICN architectures surveyed above in the aspect of scalability of control plane, content based caching protocol, information centric context networking and domain clustering based networking. In the following, some improved architectures and mechanisms are presented based on the aforementioned ICN architectures.

A scalable area-based hierarchical architecture (SAHA) for intra-domain communication is proposed to address the control plane scalability problem in Software defined information centric networking (SD-ICN) in [25]. The SAHA supports scalable awareness of network resources and content resources, and it also guarantees efficient interest matching and resource adaptation.

A decentralized content-based publish/subscribe (CBPS) network for large-scale content distribution is proposed in [26]. The fundamental idea for CBPS networks to reach the large-scale is to convert the current exhaustive filtering service model into a service model capturing the quantitative and qualitative heterogeneity of information consumer requirements, where a subscription has to select every relevant publication. A service model is designed addressing the consumers' requirements for content-based information retrieval and the relevant protocols are provided. The proposed approach is evaluated, and different content and interest forwarding strategies as well as caching policies are compared in terms of resource efficiency and QoS metrics in realistic workload scenarios.

In [27], an information and context oriented networking framework (ICON) is presented to support the deployment of pervasive networks by combining two networking paradigms that are highly correlated to the efficiency of data sharing: namely data-centric networking and opportunistic networking. ICON incorporates four components, namely Decision Engine, Data Engine, Context Engine and Network Engine. In ICON, data is shared taking into account users' social affinities instead of the capability of devices by using a data-centric forwarding algorithm, which is based on a utility function that reflects the probability of encountering nodes with a certain interest among the ones that have similar daily social habits. The reason to use social proximity with content knowledge is that nodes with similar daily habits have higher probability of having similar content interests; and social proximity metrics allow a faster dissemination of data by taking advantage of more frequent and longer contacts between socially closer nodes. However, ICON lacks of evaluation considering node mobility and traffic models.

In order to deal with extensive delivery latency, an ICN based networking method integrated with hash-routing and domain clustering techniques is presented in [28]. The network is represented as a graph which consists of the sets of routers and communication links and each node of the graph enables to cache and forward a request to corresponding caching nodes. Domain clustering technique is used to partition the set of routers of the large domain into clusters containing the subset of routers, and a hash function is implemented at the nodes of each cluster to determine both the content placement and the request-to-cache routing process, thus reducing the retrieval delay.

2.2 Caching Framework based on Mobile Cellular Networks

In general, mobile clients access the network by connecting to one or multiple APs to obtain mobile multimedia services (e.g. audio and video), when their position changes, they change the AP which they connect to. In mobile cellular networks, the APs contain NodeBs in 3G networks and eNBs in 4G LTE networks [29, 30], macro base stations(MBSs)and small cell base stations(SBSs) which are proposed in heterogeneous cellular network(HCN) [4, 30–35], as well as Wi-Fi access points and worldwide interoperability for microwave access (WiMAX) base stations [30].

Figure 2 gives the typical architectures and scenarios of caching framework in mobile cellular networks. In mobile cellular network, caches can be deployed either in the core network or in the access networks, and also can be deployed both in the core network and access networks. The MBS and the SBSs are cooperative to cache contents in order to relieve the backhaul traffic load and improve hit rate of contents of mobile users. In cache-based Cloud Radio Access Networks(C-RAN), caches are deployed not only in base-band units (BBU) pool but also in some edge elements such as content access points(C-APs) and user equipments (UEs).

In this section, the research issues focusing on the caching frameworks based on mobile cellular network are addressed.

In [4], a two-layer cellular caching framework is constructed to investigate cache aware user association policies. The architecture consisting of a single MBS and a set of SBSs is provided as the cache-enabled architecture based on SBSs, such as picocells, microcells or femtocells overlaid on existing macro-cellular wireless systems. The request is sent to the SBS at first, and if the SBS owns a copy of the requested item, the request can be satisfied locally; otherwise, the request that cannot be satisfied by the SBS is routed to the MBS [32, 34].

In [31], a three-tier heterogeneous wireless network is put forwarded consisting of a number of macro base stations, relays and mobile users. BSs, relays and users are cooperative in transmitting content data.

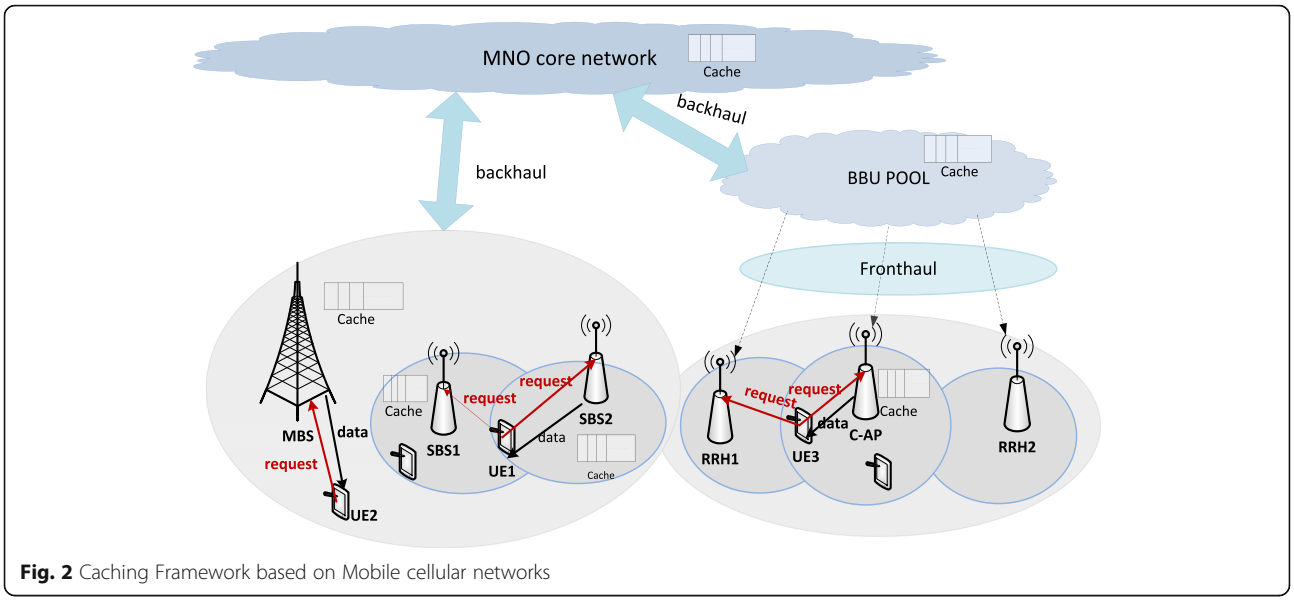

Fig. 2 Caching Framework based on Mobile cellular networks

In [29], a cooperative cell caching system consisting of some content providers(CPs) and a great number of cells is presented, CPs are outside the mobile network operator(MNO), while cells are inside the MNO network and covers the whole service area. Caching cooperation among cells are used to reduce the transmission latency due to short distance between the cells and mobile users, and redundant data streams are offloaded from the CPs at the same time, therefore heavy burdens on the backhaul channels are alleviated.

The authors in [35] propose a mobile caching framework based on the underlying HCN topology. A factor graph is used to represent the framework, where vertex set consists of factor nodes and variable nodes. Each factor node is related to a mobile user and each variable node is related to a SBS.

With the development of cloud computing and virtualization, C-RAN has become a new solution in cellular access network. In C-RAN, distributed remote radio heads (RRHs) connect to the BBU pool via fronthauls, while the BBU pool connects to the content cloud through backhaul. According to the C-RAN architecture, a cooperative caching framework is proposed in [36] based on Distributed Cloud Service Network architecture, in which several distributed cloudlets with local caches cooperate and share contents with each other. [37, 38] proposes a cluster content caching structure in Edge Cloud-Radio Access Networks (EC-RANs), in which baseband signal processing and control functions are partially decentralized in the edge equipments, such as base stations and UEs with cache, and a common local content cache is deployed in BBU so that cache can be shared by the RRHs within the cluster. A fog computing based RAN (F-RAN) architecture is proposed in [39] which user plane and control plane are decoupled. In F-RAN, MBSs are provided for control signal interaction on the control plane, while RRHs are deployed to enhance high speed data transmission on the user plane. Besides, F-RAN explores the potential of caching in the edge caching APs as well as the edge equipments, and UEs supports cooperative caching through device to device (D2D), therefore, the burden of the fronthaul and BBU pool can be alleviated. In [40], a survey of C-RAN and fog network is presented to investigate architecture harmonization of C-RAN with fog computing network.

Caching-as-a-Service (CaaS) is proposed in [41] as a caching virtualization framework which makes caching more flexible. CaaS instances can be created, migrated, scaled (up or down), shared and released adaptively in the mobile clouds depending on the user demands and requirements from 3rd-party service providers.

2.3 Caching Framework based on wireless ad hoc networks

Wireless Ad Hoc Networks (WAHNs) consist of autonomous mobile nodes, these nodes cooperate with each other to exchange information by multiple-hop communication. Although each node has limited transmitting range, some nodes behave as ad-hoc routers and forward information (e.g. requests) from other nodes. Cache can be deployed either on each node or on some selected nodes to leverage cooperation [42]. Figure 3 is a typical illustration of caching framework based on WAHNs, in which caching nodes cooperate to cache contents and retrieve requested contents in a multiple-hop fashion.

In this section, some research issues are investigated based on information-centric caching frameworks in wireless ad hoc network.

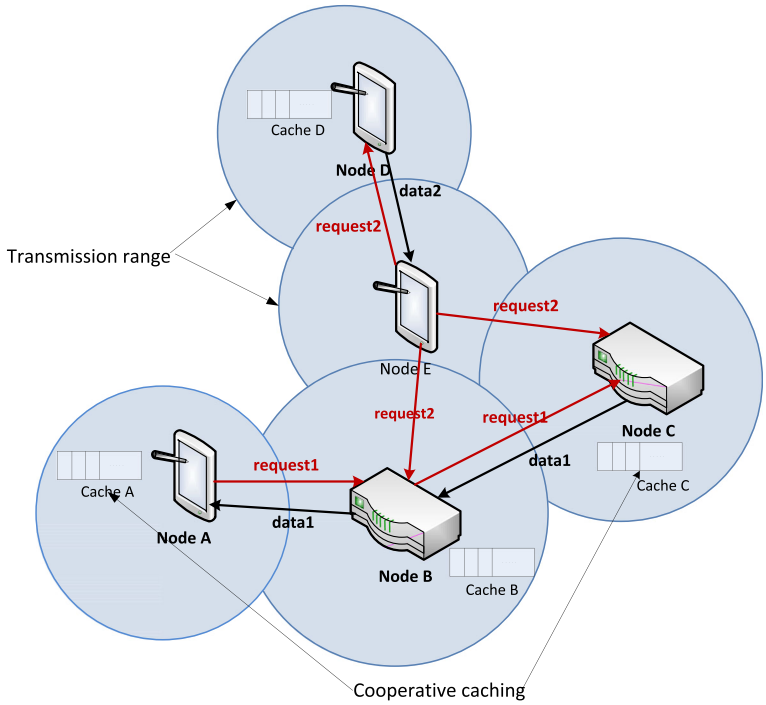

Fig. 3 Caching Framework based on Wireless ad hoc networks

A cooperative caching framework for mobile nodes is proposed in [43], in which the buffer storage of each mobile node is divided into three parts: namely self, friends and strangers. Each cache node stores its most frequently accessed data items in the self part. The mobile nodes with higher contact frequency are considered as friends to the cache node. Each cache node helps its friends to store data items according to friends' preference in the friends' part, in addition, the cache node randomly caches a subset of the remaining data items in the strangers' part.

In [44], a novel cooperative caching strategy is proposed for mobile ad-hoc networks (MANETs) named Administrative Cluster-Based Cooperative Caching (ACCC). The network is divided into a set of overlapping clusters, and each cluster is managed by a Cluster Manager and a ClusterBackup. In order to improve the reliability of the cooperative caching system, at most two copies of the cached data items in each cluster are kept.

In [45], a holistic cooperative-caching mechanism in WAHNs is proposed to reduce average query-solving time and network traffic, and improve query-solving rate considering data request-frequency, presence-index, and cache-splitting (popular and non-popular).

In [46], an interest-based cooperative insertion policy is presented in a multi-hop wireless network adopting CCN-like cooperative caching. In a wireless network consisting of a server and several nodes interconnected through a multi-hop linear topology, each node is associated with a unique user, and every user device acts as a caching node and only stores the contents according to the interest. The social relationship among users is quantitated as degree of similarity, and the user interest for a specific content is modeled by degree of interest based on a social space model.

In wireless multi-hop networks, content servers are generally located outside a wireless multi-hop network and a user accesses these servers through a gateway node (GW). As popular contents have a tendency to be stored in cache storage, and when a content request encounters cached content on the way to the GW, content download is launched from this encountered node. So wireless cache networks are expected to reduce concentration of content request and content traffic of popular contents around GW. Since caching networks (not only wired network but also wireless network) have a bandwidth limitation along default-path, two solutions are provided to make cached contents not along default-path, namely explicit cache coordination and implicit cache coordination. In [47], a mechanism guide for cached contents called Breadcrumbs is evaluated in WAHNs as an implicit cache coordination approach. The results show that Breadcrumbs in WAHNs is efficient in leveraging the limitation of cache availability on a default-path in caching network.

An intelligent group caching scheme called Dynamic Group Caching is presented which allows grouping of mobile hosts at one hop distance [48], in which Head and Group Master are elected to manage the group.

2.4 Caching Framework based on Hybrid Network

WAHN is appealing in many places such as battlefields and disaster areas, because mobile nodes are available to communicate in a multi-hop manner without infrastructure. However, with the aim to cover a wider area with less fixed infrastructure, sometimes WAHN combining with infrastructural network is a useful approach to form a hybrid network [49]. A typical solution of the hybrid networks is composed of MANET nodes and cellular infrastructural networks, which constructs hybrid networks supporting both mobile cellular communication and ad hoc communication. Besides the hybrid network of cellular and ad hoc network mentioned above, usually, the typical hybrid networks involve hybrid network based on cellular network and WLAN/D2D, hybrid network based on ICN and ad hoc network (ICMANET), etc. Projects related include I-CAN, UMOBILE, RIFE, and, GreenICN etc. [8, 11, 12, 50]. Figure 4 presents the typical caching framework based on hybrid networks, in which caches are placed in MBSs, SBSs, APs and user devices in radio access networks. In this section, research issues on caching enabled hybrid network are addressed.

I-CAN is presented to advance the integration of cellular (licensed spectrum) and wireless (license-exempt) access technologies in [3, 5]. In I-CAN, proactive caching exploiting mobility and content prediction, in-network caching and replication, multipath/multisource transport and traffic engineering are investigated in integrated cellular and Wi-Fi networks.

A cache-enabled hybrid network is formed by cellular network and autonomous nodes in military network

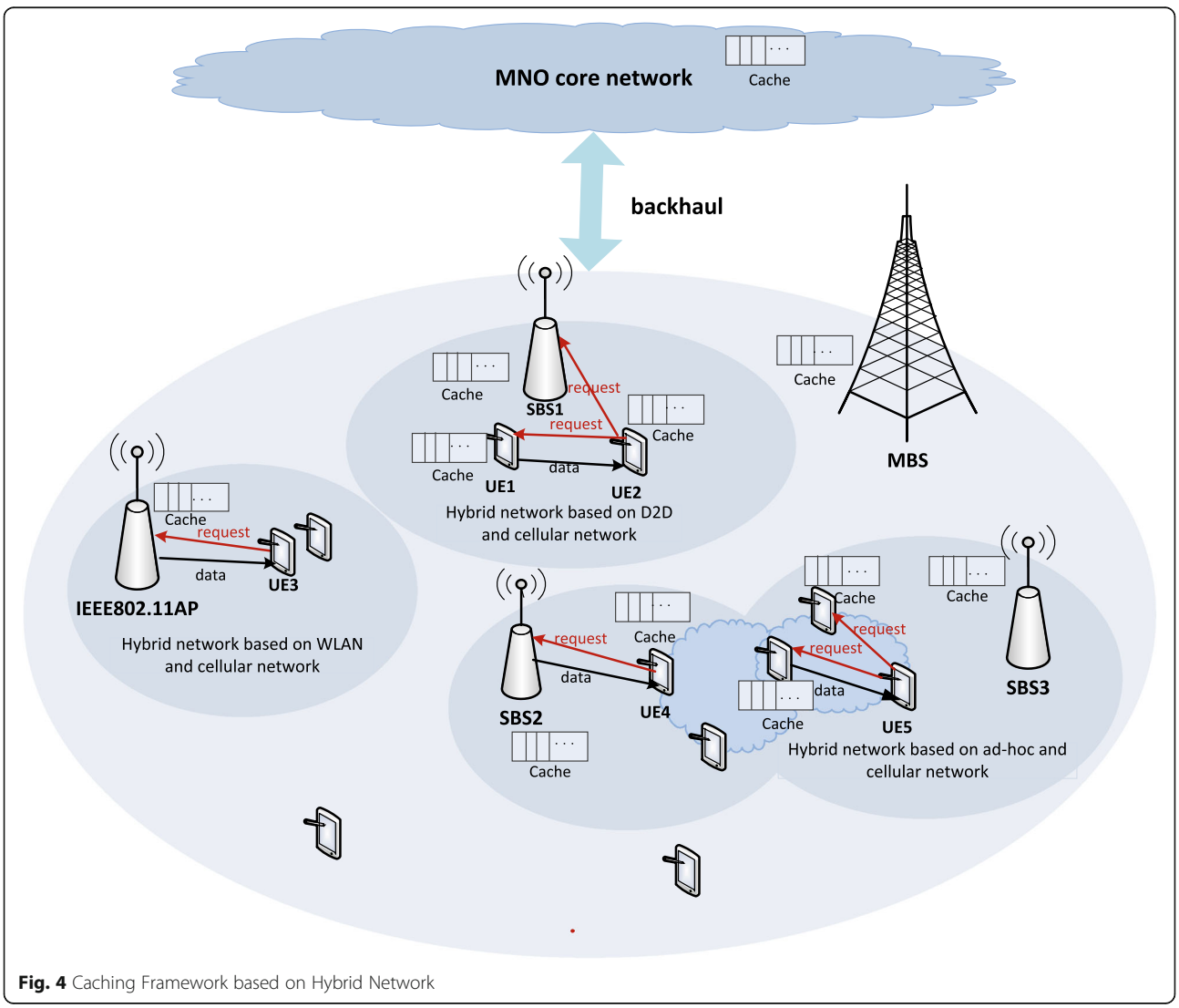

Fig. 4 Caching Framework based on Hybrid Network

[51]. In the hybrid network, the back-end server cache all the needed content, the in-MANET cluster nodes are equipped with caches and decide which content to cache. On the one hand, back-end server can provide the requested contents via the cellular network; on the other hand, cluster nodes may offer the contents.

As IEEE802.11 supports massive throughput and high transmission rate, and IEEE802.11 APs are widely deployed without additional infrastructure, most mobile devices are equipped with IEEE 802.11 series. The idea of ICN-enabled IEEE 802.11 wireless APs as nano data centers is introduced to reduce traffic load and operational cost [52].

In [53], a caching framework is addressed combining a cellular network and an opportunistic network of D2D communication. The cellular network provides pervasive network coverage and a global communication channel for all mobile nodes and it is mainly used for control message exchange. Content delivery traffic is offloaded through D2D communication between nearby mobile nodes to reduce content delivery cost. The links used for D2D communication can be Wi-Fi Direct, Bluetooth, LTE/LTE-A, etc. Seeds and relays are chosen as helpers to contribute their storage as content caches and allow other nodes to download content from their caches through D2D links.

In [54], an ad hoc based caching framework integrated with NDN is proposed in the highway VANET scenario by using ad hoc Wi-Fi 802.11a, and NDN stack is installed in each vehicle and RSU. A highway-customized ICN-based cooperative caching mechanism named ICoC is provided aiming at maximizing video playback quality without excessive startup delay and minimizing the playback freezing ratio.

In [55], content-centric data transmission is investigated in the context of short opportunistic contacts based on CCNx framework content-centric networking architecture in wireless mesh network.

All of the literatures above indicate that mobile caching network frameworks can be distinguished in network type and cache location. Based on the research on information centric mobile caching frameworks mentioned, a summary on the main research issues is provided in Table 1.

3 Content Cache Policy

Content cache policy determines which objects are to be placed at which cache nodes, and it can be classified into two categories based on the operations to caches which include cache insertion policies and cache eviction policies. Content cache policy can be used in wired communication and wireless communication networks. Because of the constrained transmission bandwidth and mobility of users in mobile wireless communication networks, the main factors affecting the performance of mobile content cache policies include caching granularity, content popularity, caching redundancy, policy complexity, mobility of caching networks and whether caching nodes cooperate with each other or not.

The performances of caching policy involve two aspects, one is to improve quality of experience of users, and the evaluation metrics are usually hit ratio and access delay; the other is to improve network performance, the involved criteria usually include reducing network traffic load and alleviating the burden on the network server. In mobile caching network, the network traffic load is usually transformed to the traffic load of backhaul, fronthaul as well as link traffic among end users. Intelligent caching policies are developed in order to optimize network centric or user centric performance metrics.

Since some research issues focus either on the caching insertion or caching eviction policies, while some of the caching schemes involve both of the caching insertion and caching eviction policy, in this section, cache insertion policies, cache eviction policies and cache management schemes including content insertion and eviction are analyzed respectively.

3.1 Cache insertion policies

Cache insertion policy refers to the strategy of cache enabled equipments to insert contents according to some factors related to routing, content popularity and social relation of the users, etc.

The state-of-the-art cache insertion policies mainly consists of Caching Everything Everywhere (CEE) [56](refer to “ALL-CACHE” in[46],“leave copy everywhere (LCE)”in [57]), Leave Copy Down (LCD) [46, 56, 57], Caching with Probability (Prob(p)) [46, 56, 57], ProbCache [18], WAVE [58], Breadcrumbs [47], Ditto

Table 1 Mobile wireless caching frameworks and related research issues

Network	Cache location	Main Works
ICN based network	Every router(on path and off path)	[1, 2, 6, 8–28, 58, 60–62, 65, 68, 70, 80–82, 88, 90, 94–102, 104, 105]
Cellular/wireless based network	Edge based caching (in MBSs, SBSs, APs and user devices)	[4, 29–41, 93, 103, 106, 107]
Ad hoc based network	Edge based caching(in every node and in the server)	[42–48, 59, 89, 92]
Hybrid network	Edge based caching (in APs, BSs and mobile devices)	[3, 5, 7, 49, 51–55, 85, 86, 122]

[59], Partial Cache [60], Interest-based cooperative caching (ICC) [46], Proactive Content Pushing Scheme (PCP) [61], Hierarchical Cooperative Caching(HCC) [43] and Mobility/Popularity-Based Caching Strategy(MPCS)[62],etc. The principles of the cache insertion policies are summarized below.

- CEE: CEE is the default cache decision policy in CCN/NDN, and it aims to reduce user access latency and frequent downloading from the original content servers. In CEE, a requester sends out an Interest packet which is forwarded to the server, and Data packet flows back along the reverse path and it is cached at every CR along the path. It is clear that this approach is simple but has the disadvantage of caching redundancy because every CR holds the copy of the Data packet, which causes the low utilization of caching resource.
- LCD:LCD is designed to reduce caching redundancy in ICN. In LCD, a caching suggestion flag bit is added at the Data packet header whose state determines whether CRs along the downloading path should cache the Data packet. When a cache hit occurs in a cache node, only its direct downstream node caches the content, which prevents other CRs along the downloading path from caching the same object. The frequently requested contents tend to be cached in CRs close to end users. Therefore, LCD is a cooperative policy which considers the contents' access frequencies.
- Prob(p): Prob(p) is used to reduce caching redundancy and improve cache efficiency. In Prob(p), cache decision is made for each CR along the downloading path with probability p, which means each CR does not cache the requested content with probability 1-p, then if a content is requested for many times, the probability that all CRs do not cache the content is low, namely, popular contents are more likely to be cached in routers in this scheme.
- ProbCache:ProbCache aims to reduce caching redundancy and achieve efficient utilization of available cache resources along a delivery path. For each CR, the probability of caching incoming chunks is influenced by caching capability of paths and the distance between the CR and the user. The probability is proportional to the caching capability of the CR and inversely proportional to the distance between the CR and the user. If a CR with a big cache is close to the user, the probability of caching contents is high, while if a CR with a small cache is far from the user, the chance to cache incoming contents is little. The scheme guarantees fair load distribution and fair content flow multiplexing between contents that travel to different destinations in terms of path length.
- WAVE: In WAVE, contents are segmented into chunks. The number of chunks to be cached is adjusted based on the popularity of the content. An upstream node recommends the number of chunks to be cached to its downstream CRs by adding caching suggestion flag bit in data packets in order to help them make caching decisions.
- Breadcrumbs: Breadcrumbs aims at efficiently utilizing content caches located outside of the default path. When a content request arrives at a content server and a content is downloaded, each router along the download path stores a pointer called Breadcrumbs. This pointer points to the direction in which the content was sent. When a content request eventually encounters Breadcrumbs for the corresponding content along the default path, it is redirected towards Breadcrumbs direction.
- Ditto: It is an on-path opportunistic caching policy which caches overheard data at the granularity of small chunks to improve subsequent transfer throughput and reduce traffic load on the gateways in wireless mesh networks. If user i requests a video object but it can not be hit by nearby routers, then the request is sent to the gateway node, the gateway node returns the video object on the reverse path. Ditto caches the object at routers on the path, and it also overhears and caches part of the video at routers which are off the path within radio range of routers on the path. If user j subsequently requests the same video file, the request can be satisfied by the routers in radio range which overhears the transfer instead of traveling multiple hops to the gateway node.
- Partial Cache: It is also a chunk based caching policy in NDN. In partial cache policy, the requested content is divided into several data chunks and transmitted along the routing path, and when a user requests a multimedia named content published by a provider, the multimedia content can be partially cached in an intermediate node or a proxy node. That is to say, each neighboring node partially caches different chunks of named multimedia content. In fact, the grouped neighboring NDN nodes will keep a whole cache of named multimedia content in order to serve the next same content request from the neighboring region.
- ICC: It is a caching insertion policy used in multi-hop wireless networks which only stores those contents that would be of possible interest for the corresponding user according to the social distance which is defined as the Euclidean distance of the user u and the content c. Since online recommendation systems provide

models to evaluate similarity of interests among users and the user's interest for a content based on the behavior of the user, the insertion policy of ICC at user u's cache simply stores the content c if their social distance is below a threshold.

- PCP: It is a caching insertion policy in NDN to reduce the service disruption time due to the content provider mobility. In PCP, interest and data packets are exchanged between the content providers and content consumers. When a handover trigger is detected by the content provider, the content provider sends an unsolicited DATA packet containing the selected contents for pushing. Then, the NDN routers receive the pushed contents and store the contents in their content caches. After completing the handover, the content consumer sends an interest packet. Since the requested content has been already pushed to NDN routers, the content consumer can retrieve the content from the NDN routers instead of the content provider.
- HCC: It is a caching insertion policy based on the social relations with the mobile user. The cache space is divided into three levels: namely self cache space, friends' cache space and strangers' cache space. The self cache space is used for the mobile user to store the data items according to his/her preference. The friends' cache space is remained to help his/her friends to cache some data items, and the strangers' cache space is for the mobile user to store cache data items randomly.
- MPCS: It often happens in the real world that users leave one site without finishing the current download due to the large size of contents (e.g. a video clip). On the movements between sites, users usually turn on cellular network interface to continue downloading, and may switch to Wi-Fi networks when arriving at the next hotspot. For the unfinished download, a request for the same content will be generated at the new site. In this case, these requests are dependent on the download at the previous sites, thus the content request rate at each site is affected by other sites due to the user mobility. By taking advantage of both the popularity of contents and the mobility pattern of users for ICN, MPCS consists of two caching strategies, one is called MOC(Mobility-Oblivious Caching), in which the most popular contents at any site will be cached locally to get an optimized cache hit rate, and the other is called MAC(Mobility-Aware Caching), in which the cached contents are chosen from the popularity rank list considering both the popularity ranks and the site transition matrix of mobile users.

From the schemes of the caching insertion policies surveyed above, the features of the policies differentiate in content granularity, content popularity, mobility support, complexity and cooperation among cache-enabled network elements. Table 2 illustrates the comparison of different cache insertion policies.

3.2 Cache eviction policies

The cache eviction policy refers to the policy to remove contents from cache due to caching space limitation. The state-of-the-art cache eviction policies mainly include FIFO [63], simple Random Replacement (RR) [64], LRU[65], LFU[63], Least Recently/Frequently Used (LFRU)[66], LRU-k[67], Time Aware Least Recent Used (TLRU) [68], Frequency-Based-FIFO(FB-FIFO)[69], Aging popularity-based caching scheme (APC)[70] and Adaptive Replacement Cache(ARC)[71],etc. The processes of the cache eviction policies are analyzed as following.

- FIFO: FIFO is the simplest policy for content eviction. The content item which has been in the

Table 2 The comparison of different cache insertion policies

Policy type	Cooperative	Content granularity	Content popularity	Mobility support	Complexity	Main works
CEE	no	Content level	no	no	low	[46, 56, 57]
LCD	yes	Content level	yes	no	medium	[46, 56, 57]
Prob(p)	no	Content level	yes	no	low	[46, 56, 57]
ProbCache	no	Chunk level	yes	no	medium	[18]
WAVE	yes	Chunk level	yes	no	high	[58]
Breadcrumbs	yes	Content level	yes	no	medium	[47]
Ditto	yes	Chunk level	yes	yes	high	[59]
Partial Cache	yes	Chunk level	yes	yes	medium	[60]
PCP	yes	Content level	yes	yes	medium	[61]
ICC	yes	Content level	yes	no	medium	[46]
HCC	yes	Content level	yes	no	low	[43]
MPCS	yes	Content level	yes	yes	high	[62]

cache for longest time is evicted to make room for a new content item. The popularity of contents and expiration time of content availability are not considered.

- RR: It is designed to evict a content item randomly in the cache to make room for a new content. In RR, N contents already in the cache are sampled and from the samples the content with least usefulness content is replaced which is mainly determined according to its recentness, frequency of use and size. The next $M < N$ samples are retained for the succeeding iteration. When the next new content arrives, N-M new contents in the cache are sampled, and the least useful content from the N - M new samples and the M old samples are replaced.
- LRU: It is a typical cache eviction policy often used in caching system. Upon arrival of a request, if the requested content k is not the one already in the cache(namely, a new content), the least recently used content(the bottom content) in the cache is evicted to make room for the new content item. If the requested content k is already in the cache, then it is brought to the first position and contents previously cached above the content k are moved one position down.
- LFU: Assuming the content popularity is known a-priori, LFU statically stores the C most popular contents in the cache. When the cached content is requested, the counter value is added by one; when a content item is needed to be evicted, the content with the minimum counter value is selected.
- LFRU: the LFRU policy is proposed by associating a value with each lock in the buffer, which is called the Combined Recency and Frequency value and quantifies the likelihood that the block will be referenced in the near future. Each reference to a block in the past contributes to this value and a reference's contribution is determined by a weighing function $F(x)$, where x is the time span from the reference in the past to the current time. There exists a spectrum of implementations of the LRFU that again subsumes the LRU and LFU implementations. This spectrum is dictated by how much weight is given to recent and older histories and the time complexity of the implementations lies between $O(1)$ (the time complexity of LRU) and $O(log_2 n)$ (the time complexity of LFU), where n is the number of blocks in the buffer.
- LRU-K: LRU-K algorithm keeps track of the times of the last K references to popular database pages, and uses this information to statistically estimate the interarrival times of references on a page by page basis. The LRU-1 (classical LRU) algorithm can be thought of as taking such a statistical approach by keeping in memory only those pages that seem to have the shortest interarrival time.
- TLRU: TLRU is an extension of simple LRU. In TLRU, TTU is defined as a time stamp of a content which stipulate the usability time for content based upon the locality of content and content publisher announcement. TTU_j^i is calculated for arriving content based on composite function, and if the average request time t_{ij} is smaller than TTU_j^i, then save the arriving content in cache; otherwise, apply LRU on a content set which the less popular contents are contracted.
- FB-FIFO: In FB-FIFO, a variable-size protected segment S_p is created in the cache for content items which are requested more than once within a short time span, and the remaining cache is considered as an unprotected segment S_U. Assuming that both cache segments are managed separately with the FIFO algorithm, when a new content item is requested, the item is moved to S_U as the newest content item. If S_U is full, the content item that was brought into earliest will be evicted from the cache; If an item in S_U experiences a cache hit, the item is moved to S_p as the newest item; If S_p is full, the item that was brought into earliest will be moved back to S_U as the newest item.
- APC: It is an online caching eviction policy. In APC, each node maintains a table of interested content objects, where each content object is attached by an aging key which is used to indicate the popularity of the object. The value of the key is updated periodically in order to track the change of content popularity effectively. When all the cache space of the CR has been occupied, the tagged object with the least value of aging key in the cache will be evicted.
- ARC: ARC is presented as a self-tuning, low-overhead algorithm that responds to online changing access patterns. ARC maintains two LRU page lists: L1 and L2. L1 maintains pages that have been seen only once recently, while L2 maintains pages that have been seen at least twice recently. The algorithm actually caches only a fraction of the pages on these lists. The pages that have been seen twice within a short time may be thought of as having high frequency or as having longer term reuse potential. Hence, L1 captures recency, while L2 captures frequency. If the cache can hold c pages, and these two lists are kept to roughly the same size c, together, the two lists comprise a cache directory that holds at most $2c$ pages. ARC caches a variable number of most recent pages from both L1 and L2 such that the total number of cached pages is c. ARC outperforms the LRU algorithm by dynamically responding to changing access patterns

and continually balancing between workload recency and frequency features.

Some cache management schemes include content insertion and eviction, such as q-LRU, k-LRU, k-Random [63] and Split-Cache [45].

Both q-LRU and k-LRU differ from LRU for the insertion policy, but are identical to LRU for eviction policy. In q-LRU, upon arrival of a request, a new content item is inserted into the cache with probability q. In k-LRU, the cache is managed by a virtual cache. The requests have to traverse a chain of $k-1$ virtual caches, which store only content item hashes and perform meta-cache operations before the content item is stored in the physical cache indexed by k. When a cache request arrives, the request enters cache i if its hash is already stored in cache i-1 (or if $i = 1$).

k-Random works like k-LRU, the only difference is that the eviction policy of each cache is random.

In Split-Cache scheme, the cache C is splitted into two parts C1 and C2 for popular and less popular items respectively. Some percent of the most frequently requested items are cached in C1, and the rest are cached in C2. Every node u keeps track of the number of unique requests for a data item i (request frequency). Information of the presence of item i in its neighborhood is also kept, and a presence index for each data item is computed. Merit Function is formulated by using estimated cumulative presence index and request frequency counter. If a new data item n comes to node u and the cache is full, then u first uses the request frequency of n to identify which cache (C1 or C2) item n should be put in. Then u performs the eviction procedure using the Merit function value and the corresponding merit values of the other items in the cache, and removes the item with the highest merit value in the process.

From the procedures of cache eviction policies, it is necessary to categorize the policies according to content granularity, content popularity, mobility support and complexity. Table 3 gives the comparison of the cache eviction policies described aforesaid.

4 Research issues based on mobile Caching network

ICN enables caching of content pieces that can be reused by other end users requesting the same content. Ubiquitous in-network caching means caching all content fragments that traverse it, that is to say, if a matching request is received while a fragment is still in its cache store, it will be forwarded to the requester from that network element, avoiding the requester going all the way to the hosting server. As one of the typical architectures in ICN, CCN [16] advocates such content caching. Such a universal caching strategy is unnecessarily costly and sub-optimal. Since the motivation of caching in networks aims to realize the following goals: (1) lower the content delivery latency whereby a cached content near the client can be fetched faster than from the server, (2) reduce traffic and congestion since content traverses fewer links when there is a cache hit locally and (3) alleviate server load as every cache hit means serving one less request, it is significant to model the behavior of the caching system, understand the fundamentals of caching strategies pertinent to caching network topology and network scalability, and design the optimal caching policies based on the different caching frameworks, then make optimization analysis on the caching network based on the specific optimization objectives in order to enhance the overall content delivery performance in terms of network bandwidth cost, retrieval delay reduction and server burden alleviation [19].

In this section, we focus on research issues analyzing the performance and optimization of information-centric mobile caching networks.

The research methods include mathematic modeling and simulation. In research issues based on mathematic modeling, optimization theory, Markov chain and game

Table 3 The comparison of cache eviction policies

Policy type	Content granularity	Content popularity	Mobility support	Complexity	Main works
FIFO	Content level	no	no	low	[63]
RR	Content level	no	no	low	[64]
LRU	Content level /Chunk level	yes	no	low	[65]
LFU	Content level/chunk level	yes	no	low	[63]
LFRU	Content level	yes	no	medium	[66]
LRU-k	Content-level	yes	no	medium	[67]
TLRU	Content level	yes	no	medium	[68]
FB-FIFO	Content level	yes	no	high	[69]
APC	Content level	yes	no	high	[70]
ARC	Content level	yes	no	medium	[71]

theory are usually used. In research papers based on simulation, discrete event simulation, trace-based simulation and test bed/experiment are used. The taxonomy of research methods for information centric mobile caching network is summarized in Fig. 5.

The main contributions of research issues can be classified into two respects, one is modeling behaviors and evaluating performances of caching system, the other is the optimization on information centric mobile caching network. The optimization objective is either network centric objective or user centric objective. Modeling behavior and performance analysis of caching system is surveyed in 4.1, and optimization of mobile caching network based on network centric and user centric objectives is analyzed in 4.2.

4.1 Modeling behavior and performance analysis of caching system

When a caching system is working, the content objects are inserted and evicted by the system, and the contents are required by users. Since the contents may be stored in any of the caching locations in the network, the contents are retrieved in the whole caching network and can be found at a low cost. According to the behavior of the caching system during the time when it is working, the performance of a caching system is related to its caching architecture, caching behavior, content request pattern and content features. Figure 6 illustrates a typical caching system based on caching behavior.

Caching architecture depends on the caching network topology. The typical topology includes hierarchy [72, 73], tandem [74] and arbitrary topology [75].

Caching behavior is often defined as the caching state concerning cache node, cache space, the total number of contents and content location in the cache, caching content insertion policy and content eviction policy. The content which is inserted or evicted from one of the cache nodes would cause the state transition of the caching system. Caching behavior involves steady state behavior and transient behavior.

For content request pattern, content requests are usually assumed to arrive in the network exogenously and the content request arrival process usually follows the Poisson process. The independent reference model (IRM) is often used to capture the notion of page reference frequencies [76]. Under IRM, requests received at different times are stochastically independent when modeling content request pattern.

The Zipf discrete distribution is often used in the literature to represent the popularity of Internet contents since it is shown that it is an adequate model [77]. Content requests for different content follows Zipf-distribution [77] with $\sum_{r=1}^{R} C/r^{\alpha} = 1$, where the probability of a request for the *r-th* popular content is C/r^{α} with α being the popularity factor, where R is the total number of content units. The caching insertion and eviction policy can be selected according to the popularity of contents.

When a cache store is full, the selected content will be evicted according to the caching eviction policy in the event of an arrival of a new uncached content, and the new content is inserted into the cache according to the content insertion policy.

In order to evaluate the efficiency of finding the contents in the caching system, cache hit is used to record for a request finding a matching content along the content delivery path. Otherwise, a cache miss is recorded. In the event of a cache miss, the content request traverses the full content delivery path to the content nodes with the content or to the server.

The analysis of modeling behavior and performance of caching system relates to the architecture of caching system. The performance is also different depending on whether the caches are cooperative. Some works study

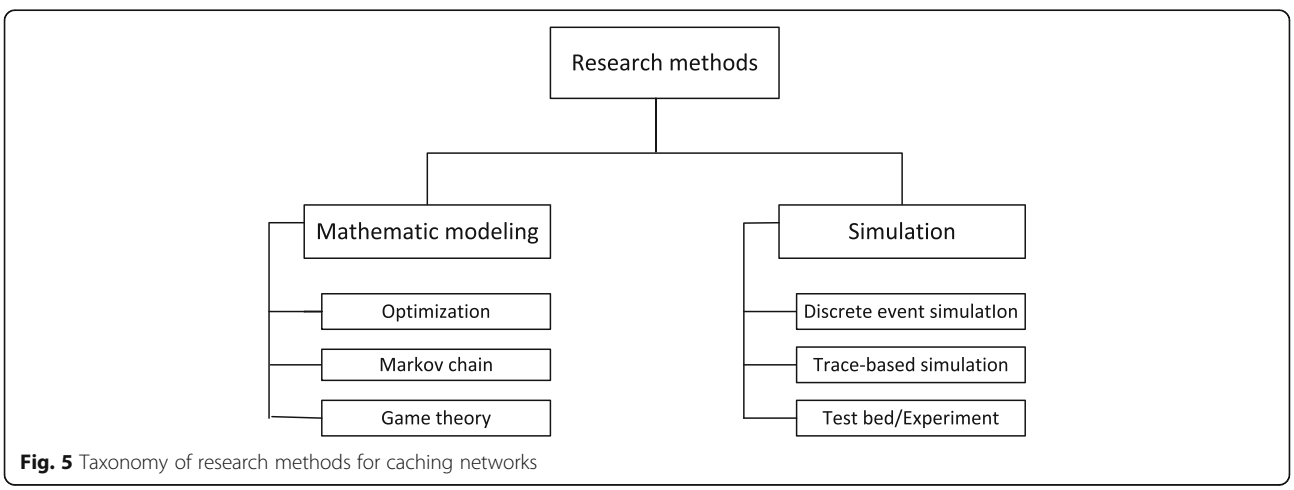

Fig. 5 Taxonomy of research methods for caching networks

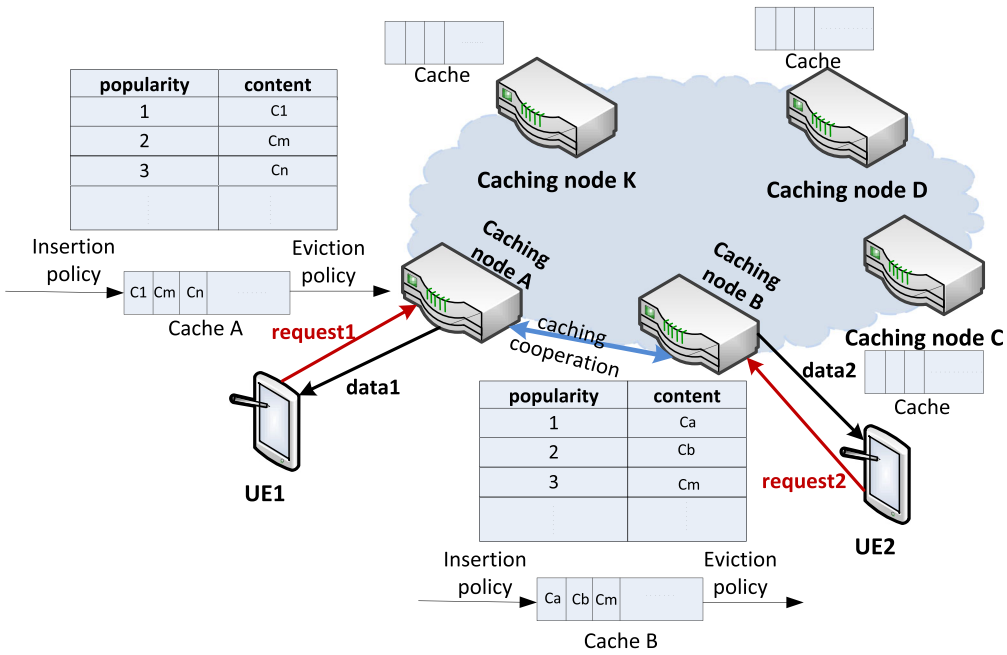

Fig. 6 A typical caching system

the caching behaviors by taking mobility into account, and some works only involve the caching schemes. In this section, research issues are addressed on caching behavior modeling and performance analysis.

Che et al. analyze the performance of two-level web hierarchy caches which a single cache is deployed at the root level and a set of caches are set at the leaf level. LRU algorithm is used at each cache locally [72, 73]. A stochastic model is proposed to characterize the uncooperative hierarchical caching system. An approximation technique is employed and the miss ratio is obtained to match the exact results accurately, which is called Che's approximation. A characteristic time for a given cache is identified as a function of the request processes, the cache size and the request pattern. Additionally, a cache can be viewed roughly as a lowpass filter with its cutoff frequency which is the inverse of the characteristic time, so contents with access frequencies lower than the cutoff frequency will have good chances to pass through the cache with the result of cache misses. According to the above idea, the cache tree can be considered as a tandem of lowpass filters at different cutoff frequencies. Based on the concept of filtering defined by the characteristic time, the hierarchical caching principles are then presented to guide the design of a cooperative hierarchical caching architecture.

Since caches are usually organized in a hierarchical manner, the cache of each level needs to serve the miss sequence from the preceding level cache, it is important to study the performance properties of tandem caches. [74] analyzes the miss stream from one single cache since it represents the input to the next level cache, and miss stream from one single cache is approximated by the superposition of a number of asymptotically independent renewal processes. Approximation of a renewal sequence is provided for predicting the fault probability in the second cache, which proves to be more accurate than IRM by experiments. The analysis can be used to rigorously characterize the performance of tandem caches. However, from a practical perspective, experiments indicate that the difference in fault probabilities in the second cache obtained by comparing the independent reference model approximation and the actual miss stream is not large, although measurable. Hence, using the IRM approximation for dimensioning networks of caches may be acceptable.

In [78], the behavior of a Time-To-Live (TTL) based caches is analyzed with a Pending Interest Table (PIT) in NDN, in which each content is associated with a timer assuming that the request arrivals can be described as renewal processes which generalizes the classical IRM. The PIT keeps track of all the Interests that a router has forwarded but are not satisfied yet, a content is evicted from the cache when its timer expires. Replacement-based caching policies are mapped to TTL-based

caching policies, and then expressions for cache hit probability, response time perceived by the users, and the size of the Pending Interest Table are derived. Applying the model proposed in [78], the traditional caching policies such as LRU, FIFO, and RANDOM can also be analyzed.

In [79], the performance of two families of replacement algorithms FIFO(m) and RANDOM(m) under the IRM model is studied by decomposing caches into lists. Explicit expressions for miss probability are presented by analyzing the steady state probabilities. Furthermore, the authors develop an algorithm with a time complexity that is polynomial in the cache size and linear in the number of items to compute the overall and per content item miss probabilities of the FIFO (m) and RANDOM (m) algorithm. Lower and upper bounds on the overall miss probability of RANDOM (m) are presented.

An analytical model of ICN storage and bandwidth sharing is proposed under fairly general assumptions on total demand, topology, content popularity and limited network resources in [80–82]. Given the assumptions that the caching network topology is a linear and binary tree, each user implements a receiver-driven flow control protocol yielding fair and efficient bandwidth utilization along the path to the content repository, and caches are managed by LRU per-packet replacement policy. A closed-form expression for expected stationary delivery time is derived as a function of hit/miss probabilities on network caches, content popularity and cache sizes. The research results can be applied to a class of content oriented receiver-driven packet-based communication networks with in-network storage. The analytical results supported by packet level simulations can be used to analyze fundamental trade-offs of ICN architectures. An essential building block for the design and evaluation of ICN protocols is also provided.

In ICN, since caching becomes a ubiquitous functionality available at each router, it is of paramount importance to develop efficient tools for the performance analysis of large-scale interconnected caches for content distribution. However, the complexity of evaluating the performance of cache networks grows exponentially with increase of the cache size and the number of contents, especially in constructing exact models of a specific multi-cache system.

The authors of [63] propose a unified methodology to analyze the hit probability of caches (both isolated and interconnected) for caching eviction policies (e.g. LRU, q-LRU, K-LRU, RANDOM, FIFO) and caching insertion policies (e.g. LCE, LCD). In [63], a simple closed-form expression of cache hit probability is obtained by extending Che's approximation under the small cache regime, which reveals superiority of the k-LRU policy by considering traffic model as renewal traffic and the effects of temporal locality (e.g. if a content is requested at a given point in time, then the content will be requested again with great probability in the near future).

In [75], an algorithm called a-NET is proposed to approximate the behavior of multi-cache networks by leveraging existing approximation algorithms for isolated LRU caches. Graph is used to denote the network of caches with arbitrary topology. The average rate of requests for a content at a cache consists of a Poisson stream of exogenous requests and miss streams of the cache's neighbors. The miss probability of each cache and the average number of response hops are used to evaluate the accuracy of a-NET.

The behavior of caches can be modeled by Markov chains or Markov Process. Several works characterize the content dynamics in a single cache or caching network by Markov chain [65, 69, 75, 83, 84].

The authors in [75] also use a discrete Markov chain to model the behavior of a tagged content k at a generic cache with size N. Markov chain consists of states (i, j), i represents the number of cache misses for contents except tagged content k, since k entered the cache, j represents the state of k, when i is not greater than N, k is in the i-th location of the cache. Otherwise k might experience eviction state, miss state and absorbing state. The Markov Model is used to measure distribution of the number of requests between two requests for content k.

In [83], a cache network is modeled as a discrete-time Markov chain with state space Ω_0, in which the content inserted or evicted from one of the cache nodes would cause the state transition of the system. Supposing that the network has m cache nodes, each cache space is N and the total number of contents is n. The system state is s, $s = (s[1], s[2],\ldots, s[j], \ldots, s[m])$, where $s[j]$ corresponds to states space of the j-th cache node. The state space Ω_0 has cardinality $\left(\binom{n}{N}N!\right)^m$.

In [65], the work process of LRU replacement policy is modeled as a homogeneous Markov chain. The state of the chain represents the corresponding content position in the cache. Supposing that each content has the same size and the cache space is N. The Markov chain states of content f_i is $\Omega = \{0, 1, \ldots, N\}$, where state 0 denotes the content f_i is not in the cache, and state j denotes content f_i is at the j-th position of the cache. According to LRU policy, the transfer rate matrix can be obtained. Then the steady-state probability of each state is derived by the balance equation. The closed-form expression of content sojourn time at a cache employing LRU policy is got, and content expectation sojourn time is provided to get the expression of hit probability.

In [69], an analytical model is presented for estimating the instantaneous hit ratio of a single cache for both LRU and FIFO. Markov chain is used to model each object in the cache. The state representation is obtained, and the instantaneous hit ratio of object is presented.

In [84], CCN is modeled using Markov chains. For a single router, the state of the chain represents the exact slot that the packet currently occupies in the cache, and the Packet of Interest (PoI) arrives as a Poisson process with rate λ. State i ($1 \le i \le N$) denotes the PoI is in the i^{th} position of router, and state $N+1$ denotes the packet is not in the cache. The requests for packets in cache, but further down than the PoI or not in cache are assumed to arrive as a Poisson process with rate μ. All states (except state 1) have a transition to state 1 with rate λ. All states i ($1 \le i \le N$) have a transition to state $i+1$ with rate μ. The chain can be trivially shown to be ergodic, and the equilibrium probability of PoI in every state is obtained, and the modeling can be extended to the multiple router system. The system of two routers is analyzed as a use case.

Some works analyze the caching behavior considering user mobility [62, 85–87].

In [62], user mobility is characterized by a stationary Markov model. The transition matrix is calculated based on the real trace data of mobile users. The statistics from the trace-based data indicates that it is a finite, irreducible and ergodic Markov chain. For a mobile user, staying on each site represents a state. Let $T_{i,j}$ denotes the probability of transition from site i to site j. The request rate $R_{k,j}$ for content k at site j is derived, and a threshold ε is set to stop the iteration process if the variance of $R_{k,j}$ is small enough.

[85] provides an analytical modeling on Information-Centric MANETs supporting mobility. The model of content retrieval for ICMANETs in 1-dimensional case is presented, which extracts locations, nodes, and contents as three basic objects. The nodes' location distribution on the generalized mobility model is analyzed and the close-form expression for the location distribution model is derived under a stationary moving process. Then, the process of content retrieval in ICMANETs is analyzed and the expression of the miss probability at a location is obtained. The closed-form expressions for accurately estimating the throughput and delay of content retrieval are derived which are used to evaluate the performance of content retrieval in ICMANETs. However, the analytical model only covers 1-dimensional case which is meaningful in the actual scene in the paper.

In [86], the model is extended from 1-dimensional case in [85] to 2-dimensional case by constructing an analytical model of content retrieval for ICMANETs based on the content hit/miss probability called PRCRM in [85]. The model is divided into node space, content space and session space. By investigating the distribution of content popularity, receiver-driven mechanism, content insertion and eviction mechanism and generalized mobility model in 2-dimensional space, PRCRM can be used to estimate the content retrieval-related performance.

In [87], a publish-subscribe system exploiting CCN system in MANET is investigated including CCN functionality, data dissemination, in-network caching, multicasting and delay tolerant delivery. Different design approaches are provided. The topic based publish-subscribe CCN system is evaluated in an emulated environment based on Linux virtual machines. The results show that Pull and Push are the two dissemination models on which publish-subscribe CCN systems may be based. Pull approach can be developed by using the CCN architecture without modifying, but requires polling, which in general leads to an undesirable overhead. However, in case of MANETs, polling is also used to refresh the topology of the dispatching multicast tree, which otherwise should be done with other means. The evaluation also reveals that the effectiveness of the CCN functionality increases with the area side and the number of subscribers.

4.2 Optimization based on information-centric mobile caching network

Optimization based on caching networks can be categorized into two classes depending on the optimization objectives, namely network centric optimization and user centric optimization. In the view point of network optimization, reducing capital expenses (CAPEX) and operation expenses (OPEX) of content delivery, wireless access networks, as well as core networks are the goals of operators, therefore the metrics are usually traffic load, link traffic, energy consumption (or energy saving rate, energy cost, energy efficiency) and throughput, some of the issues addresses the joint optimization by using total cost or utility function. In the aspect of user centric optimization, in order to improve the QoS of mobile users, user access delay and content hit rate are the evaluation metrics which are most often used.

The taxonomy of the performance metrics is provided in Fig. 7 on the optimization in information-centric mobile caching networks. In the following section, network centric optimization based on caching networks is addressed in 4.2.1, and user centric optimization based on caching networks is analyzed in 4.2.2, respectively.

4.2.1 Network centric optimization based on caching networks

In this section, research issues on network centric optimization in mobile caching networks are investigated. Mathematic methods mainly include optimization theory

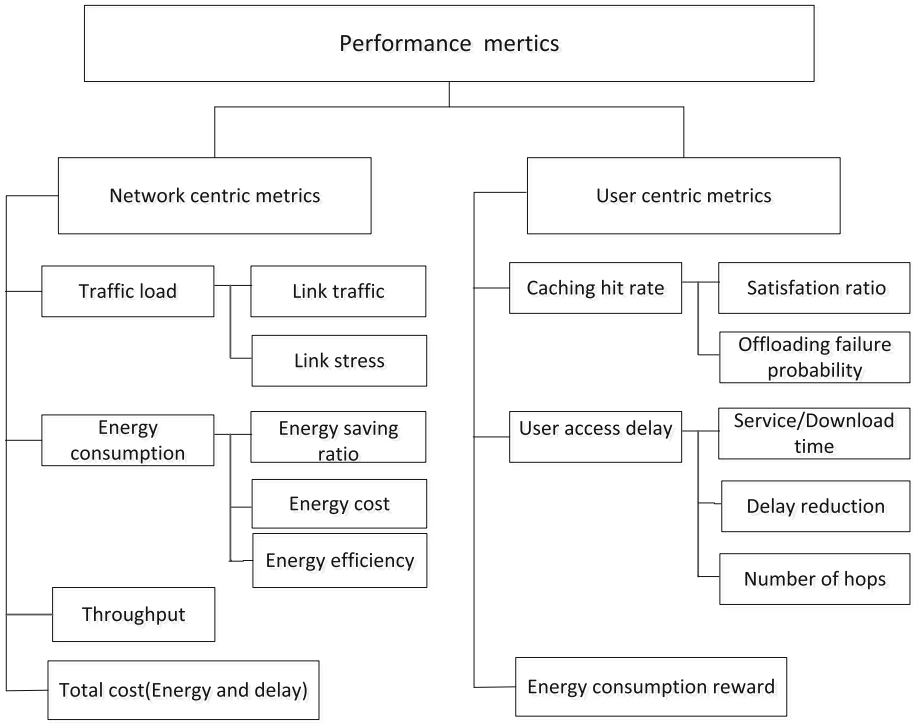

Fig. 7 Performance metrics used in information centric mobile caching networks

and game theory. Some optimization on caching schemes and caching architectures are evaluated by simulation.

Optimization theory is applied in [31, 70, 88–94] to obtain network performance improvement, while game theory is considered in [4, 95, 96].

In [31], the three-tier heterogeneous wireless caching network is modeled in order to analyze the energy saving of the system. Caching storages are available in both the relays and some of the users. The impacts of network characteristics such as the heterogeneity of multimedia contents, node density and the limited caching capacity are elaborated. The locations of the base stations, relays and device-to-device pairs are modeled as mutually independent Poisson point processes (PPPs), and four content access cases in the three-tier network model are provided. The average energy consumption of the network based on network architecture and the content access protocol is formulated. Numerical results show that the proposed mobile caching system has significant performance gain compared with the system in which neither users nor relays have caching ability considering the influence of the content size and popularity. By analyzing the influence of node density, the results reveal that there exists optimal node density arrangement to make full use of the caching resource through switching off or waking up some relays.

In [93], a cache-enabled wireless heterogeneous network framework based on software defined features is presented, in which content delivery strategy is combined with SDN to facilitate the management of the variety of user behaviors. Macro cells and small cells are modeled as independent homogeneous PPPs, and they are cooperative to cache contents based on the popularity ranking. The hit probability, the coverage probability, throughput, and energy efficiency are derived as closed-form expressions or the functions of the SINR ratio threshold, path loss exponent, transmission power and density of macro and small cells, cache ability, content popularity, and backhaul capacity. Numerical results show that the proposed cache-enabled framework has much higher throughput and improved energy efficiency than that of current LTE networks.

The energy efficient caching for CCN is studied in [70]. The energy consumption model is defined including energy consumption for transmission, caching and devices. The objective of energy optimization is transformed into the problem of minimizing average response hops. An energy-efficient aging popularity-based caching scheme is designed to improve energy efficiency and reduce network traffic.

An energy consumption model for CCN is constructed to minimize the total energy consumption of CCN [88]. The energy consumption model consists of two major parts, namely the caching energy and the transport energy. An energy-proportional model is used to define the caching energy. The caching energy is proportional to

power efficiency of caching which depends on the caching hardware technology, and the transport energy includes the energy consumption in core and edge networks. Minimization of total energy consumption of CCN is then formulated as linear and nonlinear programming problems. A genetic algorithm approach is proposed to find energy-efficient cache locations. Numerical results indicate that energy-proportional caching with sufficient caching capacity is required for realizing green CCN.

In [89], the problem of energy-conscious cache placement in wireless ad hoc networks is addressed in order to get the trade-off performance between energy consumption and access latency. Considering limited amount of available energy of mobile devices and scarcity of communication bandwidth in wireless networks, caching strategy is designed to minimize the total cost considering both the access latency and the energy expenditure for the nodes to access the cache. A connected graph G = (V, E) is used to represent a multi-hop ad hoc network with V nodes connected by E links, a caching strategy is defined by a vector with length of E, where the value of element equals to '1' if a copy of content is transmitted on corresponding edge during the dissemination phase, and the value equals to '0' otherwise. The optimization problem is then formulated as an integer linear program, and the problem is the same as a special case of the connected facility location problem, which is known to be NP-hard. A polynomial time algorithm which provides a sub-optimal solution is derived. The performance of the algorithm outperforms comparing to that of three other caching schemes.

The optimization problem is investigated in [90] to minimize overall network overhead by finding the optimal assignment of the objects in the available caches and satisfying the required constraint. By considering a network where the nodes are considered as Autonomous Systems in ICN, the Genetic Algorithm is used to support both local and global searching for finding best locations for caching contents.

In [94], an optimization model is formulated which focuses on reducing the total amount of traffic exchanged through the network in CCN and CDN. Based on a time varying content popularity evolution model that accurately represents the dynamic behavior of users and the most notable cache replacement policies, numerical results show that CDN reduces the overall traffic exchanged between network nodes in most of the cases, whereas CCN performs well in those scenarios where CDN can not quickly react to popularity evolution. Different cache replacement algorithms have very limited effects on the performance evaluation results.

The optimization problem is investigated for minimizing traffic load in large wireless networks in [91], in which the popularity of a single file varies with time. In addition, traffic in Information-Centric based multi-hop wireless network is studied in [92]. Some works present their own optimization objectives, such as the utility function illustrating the amount of bits delivered to users by a certain SBS in [4], the utility function specifying the utility related to the prices, caching costs, demands and fraction of content stored in [96].

In [92], the sustainability of multi-hop wireless communications in the context of Information-Centric Networks is addressed when content is replicated in caches over the network. In a flat wireless network for a given content popularity distribution with the network size N, the content volume M and the cache capacity K per node, the optimization objective is to select a joint replication and delivery scheme that minimizes the link traffic. By letting the three size parameters jointly scale to infinity, the scaling laws is found about the link rates ranging from O ($N^{1/2}$) down to O (1), and the regime that the network becomes sustainable subject to the scaling of the three network size parameters and the Zipf rank exponent is determined. The merit of network resources and the induced trade-offs about network expansion are analyzed.

The SBSs are faced with some constrained backhaul bandwidth and limited storage capacity. [4] exploits caching capabilities in SBSs to overcome the backhaul capacity limitations and enhance the QoS of users. A network deployed with M mobile UEs and N SBSs is considered based on Manhattan mobility model. Each SBS experiences unique network conditions due to the number of UEs currently serviced and the existing traffic load on the backhaul, therefore the SBSs individually decide which UEs to serve, which is then transformed into the problem of UE-SBS association. A utility function is defined as the amount of bits that SBS i ($i \in N$) deliver to UE m ($m \in M$) during the service time in order to formulate the UE-SBS association problem, and the problem is modeled as a one-to-many matching game for a SBS and many UEs. To solve the problem in a decentralized approach, the aim of each SBS is to maximize its own utility. An algorithm based on the deferred acceptance scheme is proposed to identify a UE-SBS matching. The results show that the proposed cell association scheme yields significant performance improvement comparing to a cell association technique without caching considerations.

In [95], an energy-efficient distributed in-network caching scheme for CCN is proposed. In the proposed scheme, each CR only needs make caching decisions locally considering both caching energy consumption and transport energy consumption. The energy-efficient distributed in-network caching problem is then formulated as a non-cooperative game, and the optimization

objective is energy cost and energy saving rate. Pure strategy Nash equilibria is proved to exist in the distributed solution, and it always has a strategy profile that implements the socially optimal configuration, even if the routers are self-interested in nature. Simulation results indicate that the proposed scheme is competitive to the centralized scheme, and has superior performance compared to other schemes widely used in CCN. Besides, it exhibits a fast convergence speed when the capacity of CRs varies.

In [96], joint caching and pricing strategies based on content popularity are investigated among access ICNs, a transit ICN and a CP. Each of the ICNs (players) has its own pricing strategy for charging others by providing service and the caching cost of access and transit ICNs is inversely proportional to the content popularity, while CP has constant cost for caching all types of content. The utility function for each player is defined as the utility obtained by providing the services for others, which is related to the prices, caching costs with respect to content popularity, demands and fraction of content stored. A non-cooperative joint caching and pricing strategies game among players is formulated, where each player tries to maximize its own utility. A unique Nash equilibrium exists for the problem of joint caching and pricing strategies for the case of symmetric Access ICNs, and the 0-1 caching strategy is determined by a threshold value which can be obtained by decomposing the joint problem into two independent caching only and pricing only problems, then contents rank based on popularity before the threshold are cached. Numerical results show that as the Zipf's factor increases, only those requests for less popular contents are forwarded by the transit ICN to the CP, since requests for the more popular contents are locally served by access ICN.

In addition to mathematic methods, simulation can also be used to evaluate the optimization of the mobile caching system, which include discrete event-driven simulators, test beds and trace-based simulation.

In [25], an architecture called SAHA is proposed to address the control plane scalability problem in SD-ICN for intra-domain communication. The SAHA supports scalable awareness of network resources and content resources as well as guarantees efficient interest matching and resource adaptation. The effects on scalability performance of resource awareness on the control plane are analyzed. Relations of number of routers, number of areas and number of interests, number of resources and average cache hit rate at area controllers are investigated respectively. Simulation experiments under OMNET++ show that the proposed SAHA can achieve outstanding scalability in resource awareness and content-based communication.

In [30], techniques for caching in current mobile networks and potential techniques for caching in 5G mobile networks are surveyed including evolved packet core (EPC) network caching, RAN caching and CCN-based caching. Using trace-driven simulations, the performance of different caching techniques is compared with respect to traffic load. Besides, the Internet inter-ISP cost, eternal cost and total network OPEX are compared with the No cache, EPC cache and EPC + RAN cache policy.

In [47], Breadcrumbs is evaluated in wireless multihop networks by using QualNet, and simulation results show that Breadcrumbs improves not only popular content throughput but also unpopular content throughput.

In [52], a new ICN use case is presented which ICN-enabled IEEE 802.11 APs are designed as nano data centers at the edge of the network, especially at multiple standalone IEEE802.11 APs. LRU is used as the cache replacement scheme. The prototype over the virtualization platform of JGN-X (Japan Gigabit Network eXtreme) and youtube trace files collected from the university of Massachusetts Amherst campus are used to evaluate the performance of network resource saving. The results show that network resource saving is a function of cache size at the edge of the network. Mobile operators can reduce operational cost in the mobile backhaul areas. Even a small cache size, e.g., 1Gbyte can achieve around 0.15 cache hit rate, which implies that 15% of total requests from mobile users for youtube video clips could be served from wireless access points directly.

In [58], a popularity-based and collaborative caching scheme is presented in content-oriented networks called WAVE, in which an upstream node suggests its downstream node to cache the number of chunks adapting to request count. A discrete event-driven simulator GT-ITM is used. The results show WAVE outperforms other on-demand caching schemes in terms of network performance such as link stress and Inter-ISP traffic.

In [62], a Mobility-aware caching policy is presented which takes user mobility into account. The proposed policy is evaluated compared with other caching strategies using trace-driven simulation. The results show that the proposed approach can achieve a 33.84% reduction on the network traffic compared with caching strategies only considering the content popularity.

In [55], a mechanism is presented to extend caching on persistent storage enabling the completion of disrupted content transfers in the context of short opportunistic contacts based on CCNx networking architecture in wireless mesh network. Evaluations are performed on PCEngines ALIX 3D2 system boards running with ADAM, and the results show that the proposed mechanism can support content transfers in opportunistic environments without significant processing and storing overhead.

In [97], existing content dissemination architectures and energy efficiency of various networking devices used for content delivery are surveyed and compared, including networking devices such as DSL, OLT, GPON, DSLAM and core network devices etc., and the energy consumption performance of different deployment scenarios of CCN nodes are compared with nano data center and the case without using CCN. The results show that even with 20 percent deployment of CCN routers in the core, CCN can effectively reduce the hop length, thereby significantly reducing energy consumption. Also, it is more effective to deploy CCN nodes at the edge because requests travel across a shorter distance. The energy efficiency comparison using simple trace-based simulations reveals that a change from a host-oriented to a content-centric networking model can substantially improve energy efficiency of content dissemination.

The design, implementation, and evaluation of Ditto is presented in [59], which is a system that opportunistically caches overheard data to improve subsequent transfer throughput in wireless mesh networks. Ditto exploits on-path as well as opportunistic caching based on overhearing to improve the throughput of data transfers and to reduce load on the gateways. The authors use the MAP testbed at Purdue University and Emulab's indoor wireless testbed to evaluate the performance of the system, and the results show that Ditto outperforms simpler caching schemes which only cache along the actual data transfer path in terms of throughput.

In [98], AMVS-NDN (Adaptive Mobile Video Streaming and Sharing in Wireless Named Data Networking) is designed and implemented as a framework of adaptive mobile video streaming and sharing in the NDN architecture to address the video traffic problem. In AMVS-NDN, once a mobile station (MS) obtained video segments from the video server via the BS, other MSs nearby can opportunistically receive the video segment from the MS who holds the segments. In general, a MS always requests a segment over local Wi-Fi connectivity. If no MS that holds the segment is available, it will request the segment via 3G/4G. The framework is evaluated by a testbed consisting of a WiMAX base station and some Android phones. It achieves higher video quality and less cellular traffic than that of other solutions.

An intra-domain traffic engineering approach for ICNs is investigated in [99, 100] to reduce the link stress of network. To minimize the average link stress of the network, or to minimize the link stress (min-max) of the most congested link, two variations of a distributed, on-line gradient descent type cache management algorithm are presented. By using realistic network topologies and synthetic workload generators, the evaluation results show that the proposed algorithms can extremely decrease link utilization.

In [101], a content delivery architecture called ISP-centric content delivery (iCODE) is developed, and it provides efficient content delivery services since an ISP can cache contents in content router. GT-ITM is used to generate topologies for evaluating the performance of iCODE, the results indicate that iCODE can offer incentives to ISPs by reducing inter-ISP traffic and link stress comparing to CDN and P2P systems.

In [102], a content-centric MANET architecture called CHANET is proposed which is built on a connectionless layer designed on top of legacy IEEE 802.11 to provide content-based routing and transport functionality. In CHANET, nodes only cache the content matching with a pending Interest and apply LFU policy. The CHANET has been simulated in Network Simulator 2. Simulation results show the great benefits offered by CHANET in terms of high throughput and network multiplication factor compared to traditional TCP/IP-based MANETs.

4.2.2 User Centric Optimization based on caching networks

In user centric optimization based on mobile caching networks, access delay, caching hit rate and caching failure probability are often chosen as significant metrics for optimization. The metric of delay is discussed using optimization theory in [28, 29, 32, 35, 36] and game theory in [103]. Minimizing caching failure probability or maximizing caching hit rate are formulated in [33, 34, 53, 62, 104]. From the perspective of simulation, trace-based simulation is used to evaluate user experience delay in [30, 43, 107], and discrete event-driven simulators are adopted to evaluate caching hit rate in [45, 46, 58].

In [29], minimization of the expected sum user delay is addressed in mobile cellular network with cooperative cell caching strategies. The round-trip time is defined as user delay for obtaining a required content by a mobile user via the MNO networks. To solve the optimization problem, an equivalent transformation is provided which converts the problem into a linear programming optimization problem. A distributed suboptimal method is also presented which has polynomial-time and linear-space complexity to solve the problem based on the equivalent transformation.

Jiang et al. propose a caching architecture where several local caches of distributed cloudlets work in a cooperative way to realize content sharing based on Distributed Cloud Service Network architecture [36]. The cooperative cloudlets which could be co-located with BSs form a cooperative caching domain in a region. The authors use content state which is determined by content requests, editorial updates and new arrivals to formulate a State based Content Distribution (SCD) framework. The optimization objective is to minimize the average total content delivery latency for all users in

a cooperative caching domain to obtain content services. Software tools are used to solve the nonlinear integer 0-1 programing problem, and then the overall SCD algorithm for content distribution is proposed. Numerical results show that the proposed framework significantly improves content cache hit rate and reduces content delivery latency.

A joint user clustering and caching scheme for wireless heterogeneous networks is presented which small-cell users have different preferences over different content types in [32]. Service delay experienced by UE to retrieve the requested content from BS is defined as the ratio between content size and downlink transmission rate from base station (MBS or SBSs) to user, and the caching optimization problem is formulated to minimize the total delay by finding the optimal caching strategy. To solve the problem in a decentralized manner, the problem is decoupled into two subproblems. Firstly, a clustering algorithm is proposed by grouping users into clusters based on their content request similarities. Secondly, a reinforcement learning algorithm based on the distributed regret learning approach is presented to minimize the delay within a specific SBS. The proposed algorithm outperforms random caching and learning without clustering schemes in terms of lower service delay due to the edge caching strategy which more popular contents are provided close to small cell UEs.

In [35], a distributed caching optimization algorithm based on belief propagation (BP) is designed for minimizing the downloading latency. The definition of delay of downloading a file in [35] is similar to service delay in [32]. A file placement matrix is used to minimize the average delay of downloading a file. In the BP algorithm, the factor graph is established based on the underlying HCN topology. The vertex set consists of factor nodes and variable nodes, and each factor node is related to a mobile user and each variable node is related to an SBS. The expressions of the average factor graph degree distribution and an upper bound of the outage probability for random caching schemes are derived. A heuristic BP algorithm is also proposed in order to reduce the complexity of the BP algorithm.

In [28], hash-routing and domain clustering techniques are applied to ICN environment of arbitrary topology, which aims to reduce the retrieval latency of ICN. In the network, when a request arrives, an edge router calculates the hash of the content identifier and forwards it to the responsible cache in the domain. The latency for the retrieval of content is formulated, which is pertinent to the caching capability of the domain, the efficiency of the used hash function and the size of the domain. To handle the retrieval delay, nodal/domain clustering techniques are investigated, including k-split clustering and k-medoids clustering. By partitioning the domain into clusters/sub-domains based on a hybrid similarity metric which captures the topological distance and the pairwise Euclidean distance of the content popularity at each node of the network, and then using hash-routing in the subset of nodes of each cluster, the average retrieval latency of a content is reduced while keeping a large cache hit ratio.

In [53], a joint caching assignment is addressed in a hybrid network based on cellular network and D2D. Seeds and relays are helpers to contribute their caches as content caches and allow other nodes to download content from them through D2D links. Seeds are those stable content caches which serve for all requests to a given content. The system allocates a certain number of seeds for each piece of content to ensure the availability of content. Relays are temporary caches allocated to a specific request, which are selected according to the information of the given request to enhance storage utility. The content cached in relays may dynamically change due to arrival or completion of relaying requests. The problem is then formulated to jointly optimize the cache assigned to seeds and relays so that the overall offloading failure probability is minimized for contents with different request rates. The results show that the static system can only offload a constant amount of contents while in the relaying case, the amount of contents increases with the number of helpers. In the case where subscribers have less than 10 friends, the relay system can support 10 times more contents than the static caching system.

In [105], a heuristic cache management scheme for energy efficiency of ICN in wireless content dissemination is addressed. In the system, A receiver-driven chunk-based ICN enabled with the edge caching of infrastructural mobile networks is considered, where the contents are identified and stored as uniquely identifiable chunks (segments). These chunks are transported at chunk level with built-in network storage for caching. An energy consumption model and reward structure is proposed, and the caching replacement problem is formulated as the maximization of the reward, which the expected reward of caching an item in a single of AP (i.e. AP_0) is the expected avoidance of energy consumption for fetching the item from any of the network nodes other than AP_0 minus the cost of having and fetching that chunk locally in AP_0. In order to reduce the complexity of solving the problem, a greedy heuristic algorithm called energy aware caching for wireless ICN (ENACI) is presented for cache management which incorporates energy reward, popularity, TTL and delay (namely chunk loss due to delay sensitivity). The performance and low-complexity of ENACI is evaluated compared to the LRU algorithm.

In [103], the optimization of caching networks is investigated by considering a real network and a virtual

network based on an Online Social Network, in which the real network consists of a set of UEs, SBSs, Service Provider Servers (SPSs), radio links which connect UEs to SBSs and backhaul links which connect SBSs to SPSs. Through the Online Social Network, UEs can communicate and share information with their friends. A SPS decides to cache a video in a number of SBSs originating from different SPSs, therefore the matching of SPSs to SBSs is needed. The caching problem is formulated as a many-to-many matching game between small base stations and service providers' servers in order to offer the smallest download time for end-users. A new matching algorithm is proposed and it can reach a pairwise stable outcome. The satisfaction of requests by the small base stations can reach up to three times of that comparing to a random caching policy, and the experienced delay by the end-users is reduced significantly since a set of videos are cached at the SBSs and user requests are satisfied mostly from SBSs instead of SPSs.

In [33], a proactive caching algorithm called PropCaching (proactive popularity caching) is presented to maximize the satisfaction ratio under resource constraints. In the caching system, a central scheduler is designed to collect statistics on the popularity of requested files in small cell networks and a set of small cells are deployed with high capacity storage units and limited capacity backhaul links. The satisfaction ratio is defined as the ratio of the satisfied requests over the total requests. Since not all requested files can be cached due to storage constraints, the PropCaching algorithm chooses to store the files with the highest popularities until the storage capacity of small cells are achieved. The results show that PropCaching outperforms random caching in most cases.

In [34], the optimal caching policy is introduced in the cellular networks enhanced with backhaul-constrained SBSs. Supposing that user requests are unsplittable which implies that each request is entirely satisfied by a base station, the joint routing and caching policy for unsplittable requests is then formulated as an NP-hard problem which minimizes the requests routed to the MBS (maximize the fraction of content requests served locally by the deployed SBSs). A novel reduction to a variant of the facility location problem is presented which enables the derivation of a set of polynomial time algorithms with approximation guarantees. The proposed scheme outperforms typical greedy algorithms up to 38%.

In [62], a mobility and popularity-based caching strategy named MPCS is proposed to increase cache hit rate through Wi-Fi when users are moving from one Wi-Fi hotspot to another. The system model is an ICN composed of multiple sites/clusters which are able to cache popular contents, and centralized servers provide a full source of requested contents from the outside network. Mobility-aware caching policy is presented which takes user mobility into account. Content request rate considering user mobility on the content popularity at different ICN sites is derived, and integer linear programming problems are formed to maximize total cache hit rate in any site for mobility-aware caching scheme.

In [104], an optimization problem is addressed which aims to determine the optimum caching probability for each content in order to minimize the average caching failure probability in a mobile content delivery network supporting D2D. Mobile devices are designed as caching severs (caching-sever device: CSD), which can provide user devices in the proximity with some contents that they request via D2D communication in order to reduce the traffic load of backbone network. A low-complexity solution approach called optimum dual-solution searching algorithm is proposed to solve the optimization problem with limited storage of mobile CSDs. The results show that the optimal caching policy is caching those less popular contents with some given probabilities while caching more popular contents with a higher probability.

Besides analytical modeling, simulation are used to verify the performance optimization for the proposed caching policies or frameworks. Discrete event-driven simulator is used in [45, 46, 54, 58], and trace-based simulation is performed in [30, 43, 106, 107].

In [45], a holistic caching framework called Split-Cache is presented and evaluated in multi-hop wireless network, in which cache C is splitted into two parts C1 and C2 for popular and less popular items. Cache-eviction decisions are made for C1 and C2 respectively. Split-Cache is evaluated by NS-2 simulator compared with other two replacement policies. The results show that Split-Cache provides higher request satisfaction ratio and less query-solving time both in static and mobile scenarios.

In [46], an interest-based caching insertion policy is proposed based on the concept of social-distance. The main idea is to store only those contents which appear to be of interest for local users. A simulator is developed in Omnet++ to evaluate the performance of the proposed interest-based caching policy, which has superior performance on cache hit probability and access delay (hops to retrieve the contents) compared with no-cache, probability-cache and all-cache policies.

In [58], a popularity-based and collaborative caching scheme called WAVE is proposed in content-oriented networks. A discrete event-driven simulator GT-ITM is used to evaluate the cache hit ratio of the caching scheme.

In [30], several caching techniques are evaluated including no caching, EPC caching, RAN caching, EPC +

RAN caching and CCN-based caching. CCN-based caching exhibits best on the performance of user content access delay.

In [43], a hierarchical cooperative caching scheme is presented in a mobile opportunistic social network. The cache of the mobile users is divided into three hierarchical components: self, friends, and strangers. Access delay and successful ratio are evaluated. The results show that the proposed caching scheme performs better compared to the performance of random cache, selfish cache and unselfish cache policies.

In [44], a cluster-based cooperative caching strategy called ACCC is proposed for MANETs. A new administrative module is developed in ACCC to control the caching process. The network is divided into a set of overlapping clusters, and each cluster is managed by a Cluster Manager as well as a ClusterBackup. By using Java Caching System JCS2 with platform Java Enterprise EditionEE8 and client/server model, ACCC is simulated and experimental results show that it outperforms recent cluster-based caching strategies as it introduces higher cache hit ratio as well as better data availability.

Based on the highway VANET scenario by using ad hoc Wi-Fi 802.11a, two social cooperation schemes are presented in [54] to maximize video playback quality without excessive startup delay and minimize the playback freezing ratio, one is the partner-assisted cooperation caching scheme which is based on cooperative video segment caching among neighbor vehicles, the other is the courier-assisted cooperation scheme which aims to enlarge the searching range of each Interest among vehicles in different lanes and enhance distributing the Interest packets along the road so as to promote content discovery. Extensive simulations have been executed by NS3, and a considerable improvement in terms of start-up delay and playback freezing is evaluated compared with probabilistic caching.

In [106], a mobile video-centric proxy cache named iProxy is designed in cellular network for mobile video providers, and an Info-aware Cache Replacement policy called LFU-based IBR-score is provided and evaluated. Information-bound references (IBRs) is used to map multiple URLs to one IBR value (associates with a single video file) in order to collapse multiple related cache entries into a single one. The cache replacement policy depends on the IBR scores. The optimization metrics are chosen as buffering, bit rate, and start up delay. The evaluation of iProxy design relies on two sets of real traffic traces, one is the Web video data set, and the other is collected over the University of Wisconsin's Wireless Network. The results show that iProxy with information-centric replacement policies can improve video start time (by up to 13 s), buffering rates and average bit rates (by 16%).

In [107], an in-network cache assisted eNodeB caching mechanism called InCan is presented for 4G LTE Networks, which aims to cache content objects at eNodeBs, save bandwidth for mobile operators and reduce delay for end users. InCan adopts a loosely collaborative approach between eNodeBs and routers rather than the full collaboration in traditional hierarchical caching systems. The cost structure considering both bandwidth and delay is modeled, and a framework is proposed in which eNodeBs can use the information from in-network caches to improve caching performance. The problem is then formulated to maximize delay reduction and it is a NP-complete problem. A 2-competitive offline algorithm and a practical online algorithm are provided to solve the problem. By trace-driven simulations and experiments under both synthetic and real network topologies, the mechanism is evaluated and the results show that the proposed approach can significantly save bandwidth for mobile operators and reduce delay for end users.

From all the issues surveyed above, optimization of information-centric mobile caching networks can be classified by optimization goal, network architecture, optimization objective and research method. Table 4 gives a summary of the optimization on information-centric mobile caching networks.

5 Applications based on mobile caching

Content distribution based on content based naming and routing are realized thanks to Information Centric Networking. Information centric mobile caching enables typical mobile content based applications, such as mobile video streaming and content based applications of connected vehicles. Since information-centric mobile caching networking has the advantages of mobility support, in-network data caching as well as multicast data delivery regardless of intermittent and short-lived connectivity, it also becomes a promising solution for typical applications such as situational awareness applications in tactical scenarios, sensing service of IoT, control of smart grid and management of traffic-heavy applications of mobile devices.

Mobile video streaming is one of the major content based applications of explosive mobile traffic, which causes a high burden on current networks. ICN is resorted to alleviate link traffic of mobile video streaming. In [98], an adaptive video streaming and sharing application in NDN called AMVS-NDN is designed and implemented, considering that most of MSs have multiple wireless interfaces (e.g., 3G/4G, WiFi, NFC). AMVS-NDN supports adaptive streaming strategy so that a MS dynamically decides which bit rate (segment) is suitable for the current link condition and sends the corresponding interest. In-network caching is exploited,

Table 4 Optimization of information-centric mobile caching networks

Optimization goals	Network architecture	Optimization objective	Works	Research method
Network centric optimization	Mobile cellular networks	Energy consumption/Energy efficiency	[31, 93]	Optimization
		Traffic load	[30]	Trace-based simulation
		Utility function	[4]	Game theory
	Information Centric Networking	Energy consumption	[70, 88]	Optimization
			[97]	Trace-based simulation
		Energy saving rate/Energy cost	[95]	Game theory
		Utility function	[96]	Game theory
		Link stress	[58]	GT-ITM(simulation)
			[99, 100]	Test bed and synthetic workload generators
		Traffic load	[62]	Trace-based simulation
			[98]	Test bed
			[101]	GT-ITM(simulation)
			[94]	Optimization
		Network overhead	[55]	Test bed
			[90]	Optimization
		Number of resources	[25]	OMNET++ (simulation)
	Wireless ad-hoc networks	Total cost	[89]	Optimization
		Throughput	[47]	QualNet(simulation)
			[59]	Test bed
	ICN WLAN/ICMANET	Traffic load	[52]	Test bed
		link traffic	[92]	Optimization
		Throughput	[102]	NS-2(simulation)
User centric optimization	Mobile cellular networks	User access delay	[29, 32, 35, 36]	Optimization
			[103]	Game theory
			[30]	Trace-based simulation
		Delay reduction	[107]	Optimization and Trace-based simulation
		Caching hit rate	[33, 34]	Optimization
			[106]	Trace-based simulation
	Information Centric Networking	Caching hit rate	[62, 104]	Optimization
			[58]	GT-ITM(simulation)
			[25]	OMNET++ (simulation)
		Delivery latency	[28]	Optimization
		Energy consumption reward	[105]	Optimization
	Wireless ad hoc networks	User access delay/ Distance in hops	[45]	NS-2(simulation)
			[46]	Omnet++ (simulation)
		Cache hit rate	[45]	NS-2(simulation)
			[46]	Omnet++ (simulation)
			[44]	Experiment
	Hybrid work	Offloading failure probability	[53]	Optimization
		video playback quality	[54]	NS-3(simulation)
	Social network based on DTN	User access delay	[43]	Trace-based simulation

which enables MSs to share content downloaded from BS to other MSs nearby who request the same content through local WiFi, thus reducing 3G/4G link traffic. In [108], a peer-to-peer application for live streaming of video content encoded at multiple bit rates is presented, which exploits an ICN API using a Java implementation running on plain laptops to simplify the application development. In the application, peers are a small set of neighboring mobile cellular devices and they have access to a remote cellular network through the cellular interface and a local full mesh one hop network through a proximity channel, which enables cooperative downloading of a live video stream. The quality of video playback is increased by the design of cooperation, during which the peer can redistribute the chunks downloaded through cellular interface on the proximity interface to request peers in a multicast fashion, and the part is also cached in the CCN content store in order to be shared with other peers requesting it on the proximity interface. Multicasting, in-network caching and multi-rate encoding approach with dependent substreams are utilized to improve the peer to peer sharing. As an extension of [108], by deploying the proof-of-concept application on commercial Android devices, the limitations of Android devices is dealed with, and the quality of video playback streaming is improved for collaborating devices that exploit all the available radio access technologies [109].

In the scenarios of connected vehicles, it is a challenging problem to improve driving and traveling experience due to poor-quality wireless links and mobility of vehicles of VANET. ICN-based VANETs promise enhancements in the areas of application, mobility, and security. Since ICN matches the described vehicular applications' pattern better than the current Internet through named data and routing by name, content discovery is easy because name-to-IP-address resolution and continuous connection to producer are not needed. Besides, content can be retrieved from the most convenient provider for vehicles on account of anycasting and in-network caching properties in ICN, which reduces data latency and network traffic. ICN-based VANETs simplifies mobility support with named data. In addition, ICN provides content-based security with protection and trust implemented at the packet level rather than at the communication channel level [110].

For situational awareness applications in tactical and emergency response scenarios, efficient, robust, and secure network communication is required. ICMANET supports situational awareness applications at the tactical edge. Declarative Attribute-based Naming, network coding, utility-based content caching and symmetric encryption are used in ICMANET to provide efficient communication for multiple simultaneous classes of situational awareness traffic in the presence of severe disruptions to network connectivity [111].

As one of the most common applications of IoT, sensing service uses massive number of sensors to monitor environment states. However, sensors are usually energy-constrained since they have not fixed power supply. Caching is employed at the access point which acts as a gateway with a fixed power supply to store and cache sensing data temporarily and sends cached sensing data to the user, thus avoiding frequently activating the sensor and reducing energy consumption of sensor devices, and at the same time offloading data management functionalities whenever possible [112, 113].

ICN is also introduced to provide higher degree of flexibility in supporting data sharing and smart grid control. In [114], an overlay ICN-based communication framework for smart grid applications is presented, which is based on publish/subscribe operations and decouples information from location and time, yielding simplicity and efficiency of management of communication flows. The framework supports in-network management of smart grid data including caching and processing such as rate adaptation, aggregation and filtering, and it also enables real-time state estimation in the medium voltage power grid. The case of smart electric vehicle charging in smart grids with ICN is studied in [115], which integrates in-network caching and lightweight in-network processing of named data. The information of state of charge can be reported by electric vehicles while moving without the connection to each anticipated recipient, thus the complexity of communications is reduced, scalability and the growth of the smart charging ecosystem is enhanced.

Information-Centric Networking and mobile cloud computing paradigms are integrated to solve the increasing usage of mobile devices for traffic-heavy applications and reduce operational costs in [116]. A Mobile Follow-Me Cloud (M-FMC) model is proposed, which consists of two parts, one is the ICN as a Service (ICNaaS) running on the cloud, and the other is FMC components enhancing the migration of content caches located at the edge of cloudified mobile networks. The model is evaluated in a testbed, and the results show that it enables content migration in Mobile Cloud Computing networks, and also improves ICN content distribution by considering users mobility, content popularity and overall network resource optimization.

From the applications surveyed above, it is perspective that the mobile Internet would be constructed based on information centric mobile caching networking in future because of its properties and advantages. However, some challenges and open issues still lie ahead before its widespread commercial deployment.

6 Challenges and Open Issues

In this section, we present some of the challenges and open issues on information centric mobile caching in the respect of context-aware mobile caching integrating with ICN, joint optimal deployment strategy of mobile in-network caching, chunk-level content popularity, intelligent cache policies with low complexity, analytical modeling of caching networks considering user mobility, mobile caching networking integrated with SDN and NFV, information-centric mobile caching and cloud computing, as well as security and privacy of information-centric mobile caching network.

6.1 Context-aware mobile caching integrating with ICN

The application based on ad hoc and hybrid framework concerning sensor network has become indispensable in future networking. Therefore, facilitating a context-aware mobile cache-enabled cooperative system is a promising research direction, which involves opportunistic sensing, spectrum sharing and performance optimization based on content awareness. Opportunistic sensing is a new paradigm for signal and information processing in which a network of sensing systems can automatically discover and select sensor platforms based on an operational scenario, determine the appropriate set of features and optimal means for data collection based on these features, obtain missing information by querying available resources, and use appropriate methods to fuse the data, resulting in an adaptive network that automatically finds scenario-dependent, objective-driven opportunities with optimized performance. Theory and algorithms of opportunistic sensing are needed for advancing autonomous sensing that not only ensures effective utilization of sensing assets but also provides robust optimal performance [117]. With the capability of promoting spectral utilization by accessing licensed primary bands opportunistically, cognitive radios based dynamic spectrum sharing is considered as a key feature of future mobile communication. To implement cognitive radios, spectrum sensing is one principal constituent for the intelligent mitigation of harmful inferences to primary user. Spectrum sensing in cognitive communication enables user devices to acquire the mobile location of primary/incumbent user when deep sensing the spectrum availability [118], and realize the localization of primary user even in time-varying fading channels [119], thus make users access different radio access networks for context-aware content based application communication. Performance optimization based on content awareness refers to optimizing the performance of the mobile cache-enabled cooperative system according to different requirements, metrics and various resources obtained from opportunistic sensing and spectrum sharing.

From the investigation surveyed, study on context-aware mobile caching integrated with ICN is still a promising aspect in the optimization of information centric mobile caching.

6.2 Joint optimal deployment strategy of mobile in-network caching

In-network caching can enhance content delivery by storing content in every node, but such a universal caching strategy is unnecessarily costly and sub-optimal [19]. Alternative in-network caching strategies have been investigated to enhance the content delivery performance in the respect of reducing network bandwidth consumption, server load and delay experienced by end users, respectively. However, joint optimization including joint caching and routing, joint network/user performance metrics mentioned above under resource constrained conditions is still an open issue.

Liu et al. [7] provides detailed descriptions on content routing based on ICMANET, and presents a concept model for content routing, and categorizes content routing into proactive, reactive and opportunistic types, then analyzes representative schemes, which can be referred to for the study of joint optimization between content routing and caching in ICMANET.

In addition, some studies focus on the energy consumption of ICN. The energy efficiency optimization in ICN is widely investigated in [120], and several research works on green ICN are analyzed in [6], which indicates that designing a caching protocol and leveraging the cooperative caching to achieve the trade-off between energy consumption and quality of service still need to be addressed.

Furthermore, although the cost of caching contents is declining, the content replacement strategy considering tradeoff between the redundancy of content caching and caching energy consumption in caching network still remains a research direction, especially in large-scale content networks.

6.3 Chunk-level content popularity

Receiver-driven chunk-based transport is supported in CDN and NDN, that is to say, content items stored in the repository are splitted into self-identified chunks, and in turn request process is divided into content and chunk level [80, 82]. As popularity is one of the most important features which affect cache efficiency [57], the popularity of contents follows Zipf or Zipf-like distribution, e.g. a content can be divided into small size chunks [58, 80, 81], the probability distribution of requests of contents follows Zipf distribution [58], and Markov Modulated Rate Process of requests with Zipf-distributed content popularity [80, 81], but the popularity of chunk-level contents is not mentioned yet. From the analytical perspective, establishing the chunk-level content popularity model from prior knowledge is of

great significance. From the experimental point of view, since there is no large-scale operational ICN network infrastructure and applications at present, it is difficult to measure the chunk-level object popularity directly [121]. To the best of our knowledge, whether the popularity law is suitable for chunk-level content still lacks analytical and experimental study.

6.4 Intelligent cache policies with low complexity

Intelligent cache insertion policies can increase cache efficiency and the diversity of cached contents, and reduce the requests flow traveling to content servers. However, high dynamic cache environment and some constrains are needed to be considered in designing low complexity intelligent cache policies, such as heterogeneity of traffic, node mobility, dynamic changes of content popularity, limited cache space of CRs and user devices as well as energy constrains for user devices. Different types of content traffic possess different content sizes, for example, files have smaller size while videos have larger size; and different types of traffic have their own caching objective. Hence, caching schemes are expected to meet different demands for caching requirements. Additionally, in information-centric mobile caching networks, numerous new contents may be published in the systems, so it is also a promising research topic to design intelligent caching schemes with low complexity considering dynamics of contents, user-centric demands and dynamic cache environment.

6.5 Analytical modeling of caching networks considering user mobility

Analytical modeling of mobile caching networks is indispensable for the fundamental understanding of the behavior of information-centric mobile caching networks. Some works lay the foundation for the study of caching networks and some also take user mobility into consideration, e.g. an approximate model for general cache networks in [75], a unified approach that can be used as the basis of a general performance evaluation tool for caching systems in [63], an analytical model of content retrieval for ICMANETs in 1-dimensional case and 2-dimensional case in [85, 86], a random-walk mobility model in mobile ad-hoc network under a content-centric traffic scenario in [122] and a heterogeneous cellular networks deployed with caches based on Manhattan mobility model in [4]. Human mobility behavior research issues indicate more and more factors in modeling mobility [123], besides, with the dramatic development of IoT, mobile cloud computing and other new emerging communication paradigms would bring challenges on flexibility, efficiency, and scalability for future access network architecture. Mobile caching network will involve immense amount of users and objects in IoT [124]. Mobility management and related optimal caching strategies considering user and content mobility would become interesting research directions for future mobile wireless network.

6.6 Mobile caching networking integrated with SDN and NFV

The software-defined networking (SDN) and network function virtualization (NFV) promote the new trend of virtualizing mobile network functions into software-based cloud servers with the aim to resource optimal utilization [41, 125, 126]. The use of virtualization in mobile caching network can stimulate MNOs to adjust their functions in an online manner dynamically. Since SDN is mostly focusing on the adaptability and controllability of network functions using virtualized resource management systems, this would lead to the controllable networking of nodes, scalability and elasticity for resource utilization adapting to user and content dynamics [30].

Several research issues focus on the SDN and NFV of mobile caching networks. The information-centric wireless network virtualization architecture is proposed in [127] by integrating wireless network virtualization with ICN. In the architecture, radio spectrum resource, wireless network infrastructure, virtual resources and information centric wireless virtualization controller are regarded as key components. Virtual resource is divided by content-level slicing, network-level slicing, and flow-level slicing. The virtual resource allocation is formulated as an optimization problem compared with in-network caching strategy. However, the detailed virtualization method is still an open issue, especially at the interfaces between in-network caching and virtualization controllers. In [128], an ICN based edge-cloud service framework is presented leveraging NFV and SDN technologies. In [129], the application of SDN in managing resources of different types of networks is analyzed in Wireless Sensor Networks (WSN) and mobile networks, the utilization of SDN for information-centric networking, and how SDN can leverage Sensing-as-a-Service (SaaS) as a key cloud application in the IoT are addressed. In [130], a SDN based framework is presented for C-RAN in HetNet to achieve more efficient communication and information management.

From the literatures summarized above, it reveals that the research works of SDN and NFV in mobile caching network are primary studies conducted from theoretical point of view. Due to the rapid growth of traffic load and user demand for information-centric mobile content services, it is believed that mobile caching strategy integrated with SDN and NFV is a new potential research direction for further study.

6.7 Information-centric mobile caching and cloud computing

The infrastructures of cloud computing provider are usually hosted by large data centers due to management reasons. ICN makes it possible to distribute cloud service objects across a mobile or a home networking environment. Thus optimization of cloud computing management and supporting cloud computing applications by using Information-centric mobile caching framework and ICN paradigm is a potential research direction [6].

From the perspective of caching cloud based on edge cloud computing, information centric mobile caching integrated with cloud computing brings about different radio access network architectures. Caches can be placed in different network elements such as SBS, BBU pool and user devices based on various deployment strategies of C-RAN [37, 130]. From the point view of management, by constructing a cloud computing platform above ICN and providing an interface for the interaction between ICN and cloud computing platform [131], malicious attacks to the data center network can be easily avoided due to security mechanism in Information-centric mobile caching by ICN, and data recovery becomes extremely easy in the cloud over ICN. From the respect of service providers, data availability is significantly improved because of universal unique information centric data name, which allows the data flow across distinct cloud computing service providers [6]. In [113], a general framework is constructed where global cloud and ICN platforms are complemented in a totally synergic way by local clouds formed at the edge of the network by mobile devices so that service provisioning and data management functionalities are offloaded whenever possible.

All of the studies show that integration of Information-centric mobile caching and cloud computing could be a long-term goal in the optimization of cloud computing management and supporting cloud computing applications in future.

6.8 Security and privacy of information-centric mobile caching network

In-network caching seems not to be attractive for some content providers since it may cause copyright problems or legal issues [6]. In open mobile system, security and privacy has been a long-run challenge of much concern, especially in the cooperative caching system where mobile nodes cache contents due to their own limited cache space and energy. In other words, a mobile node can act as a content subscriber which requests contents from cooperative nodes and also play a role of content publisher or a relay node which provides contents to cooperative nodes. The process of cooperative caching brings about severe privacy issues, which is the reason why some mobile users would not like to take part in the caching cooperation. In ICN, content name is required to be included in requests, which would lead to privacy issues. In MANETs, connectivity provided over access points would offer an acceptable level of privacy for users who trust their access points [52]. This is particularly challenging in densely connected networks that easily permit packet sniffing. For ubiquitous mobile cache deployment, it is apparent that security and privacy problem needs to be addressed, and it is also worth discussion on combination of technical mechanism and new laws on content propagation during the evolution of ICN and information centric caching.

7 Conclusion

In this paper we provide a survey on information centric mobile caching. Firstly, the motivation on information centric mobile caching is analyzed, and key issues focusing on information centric mobile caching are categorized, namely, cache placement selection, cache policy design and cache content selection. Secondly, current development of information centric mobile caching networks is surveyed; and information centric mobile caching frameworks based on information centric networking, mobile cellular network, wireless ad hoc network and hybrid network are illustrated. Thirdly, state-of-the-art content cache policies are introduced, including cache insertion policies and cache eviction policies. Fourthly, research issues based on information centric mobile caching networks are investigated, which consists of modeling behavior and performance analysis of caching systems and optimization on information centric mobile caching network. Network centric optimization and user centric optimization issues are analyzed respectively. In addition, typical applications based on mobile caching are given. Finally, challenges and promising research opportunities on information centric mobile caching network are addressed.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.61471062, 61431008).

Competing interests

The authors declare that they have no competing interests.

About the authors

Hao Jin received the PhD degree from Beijing University of Posts and Telecommunications in 1996. She is currently an associate professor in the Key Laboratory of Universal Wireless Communications for Ministry of Education, Beijing University of Posts and Telecommunications, China. Her research interests include optimization of mobile wireless communication and mobile cloud computing.

Dan Xu received the BEng degree from Shandong University in 2015. She is a graduate student for master degree in the Key Laboratory of Universal Wireless Communications for Ministry of Education, Beijing University of

Posts and Telecommunications, China. Her current research interests include mobile caching, in wireless network and mobile cloud computing systems.
Chenglin Zhao received the PhD degree from Beijing University of Posts and Telecommunications in 1997. He is currently a professor in the Key Laboratory of Universal Wireless Communications for Ministry of Education, Beijing University of Posts and Telecommunications, China. His research interests include wireless resource management, cognitive radio network and wireless sensor network.
Dong Liang received the PhD degree from Beijing University of Posts and Telecommunications in 2005. He is currently a lecturer in the Key Laboratory of Universal Wireless Communications for Ministry of Education, Beijing University of Posts and Telecommunications, China. His research interests include wireless resource management, cognitive radio network and location techniques in mobile wireless network.

Received: 5 August 2016 Accepted: 2 January 2017
Published online: 17 February 2017

References

1. G Xylomenos et al., A Survey of Information-Centric Networking Research. IEEE Commun Surv Tutorials **16**(2), 1024–1049 (2014). Second Quarter 2014
2. G Tyson, N Sastry, I Rimac, R Cuevas, A Mauthe, *A Survey of Mobility in Information-Centric Networks: Challenges and Research Directions, NoM '12 , South Carolina, USA*, 2012, pp. 1–6
3. GC Polyzos, VA Siris, G Xylomenos, GF Marias, S Toumpis, *I-CAN: Information-centric future mobile and wireless access networks* (Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference, Rhodes, 2014), pp. 139–141
4. F Pantisano, M Bennis, W Saad, M Debbah, *Cache-aware user association in backhaul-constrained small cell networks* (Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014 12th International Symposium, Hammamet, 2014), pp. 37–42
5. VA Siris, N Fotiou, D Dimopoulos, GC Polyzos, *I-CAN: Information-Centric Access Networking* (2015 European Conference on Networks and Communications (EuCNC), Paris, 2015), pp. 418–422
6. AV Vasilakos, Z Li, G Simon et al., Information centric network: Research challenges and opportunities. J Netw Comput Appl **52**, 1–10 (2015)
7. X Liu, Z Li, P Yang, Y Dong, *"Information-centric mobile ad hoc networks and content routing: A survey" Ad Hoc Networks*, 2016
8. VA Siris, D Dimopoulos, "Multi-source mobile video streaming with proactive caching and D2D communication," 2015 IEEE 16th International Symposium on A World of Wireless, Mobile and Multimedia Networks (WoWMoM), Boston, MA, 2015, pp. 1-6.
9. The COMIT project, available at: https://www.ee.ucl.ac.uk/comit-project
10. The POINT project. Available at: https://www.point-h2020.eu
11. The UMOBILE project, available at: http://umobile-project.eu
12. The RIFE project, available at: https://rife-project.eu/
13. The BONVOYAGE project, available at: http://bonvoyage2020.eu
14. European commission horizon 2020 programme. Available at: http://ec.europa.eu/programmes/horizon2020/
15. T Koponen, M Chawla, B Chun, A Ermolinskiy, KH Kim, S Shenker, I Stoica, A data-oriented (and beyond) network architecture, in *ACM SIGCOMM*, 2007, pp. 181–192
16. A Ghodsi, S Shenker, T Koponen, A Singla, B Raghavan, J Wilcox, Information-centric networking: seeing the forest for the trees, in *ACM Workshop on Hot Topics in Networks (HotNets)*, 2011
17. NSF Named Data Networking project. [Online]. Available: http://www.named-data.net/
18. I Psaras, WK Chai, G Pavlou, Probabilistic in-network caching for information-centric networks, in *ACM Workshop on Information-Centric Networking (ICN)*, 2012, pp. 55–60
19. WK Chai, D He, I Psaras, G Pavlou, Cache "less for more" in information-centric networks. Comput Commun **36**, 758–770 (2013)
20. G Xylomenos, X Vasilakos, C Tsilopoulos, VA Siris, GC Polyzos, Caching and mobility support in a publish-subscribe Internet architecture. IEEE Commun Mag **50**(7), 52–58 (2012)
21. V Sourlas, P Flegkas, GS Paschos, D Katsaros, L Tassiulas, Storage planning and replica assignment in content-centric publish/subscribe networks. Comput Netw **55**(18), 4021–4032 (2011)
22. SAIL deliverable B.1 (3.1): The network of information: Architecture and applications. [Online]. Available: http://www.sail-project.eu/deliverables/
23. The Convergence project. Available at: http://www.ict-convergence.eu. Accessed 19 Dec 2016.
24. S Nelson, G Bhanage, D Raychaudhuri, GSTAR: Generalized storage-aware routing for MobilityFirst in the future mobile Internet, in *ACM MobiArch*, 2011
25. S Gao, Y Zeng, H Luo, H Zhang, Scalable control plane for intra-domain communication in software defined information centric networking. Futur Gener Comput Syst **56**, 110–120 (2016)
26. M Diallo, V Sourlas, P Flegkas, S Fdida, L Tassiulas, A content-based publish/subscribe framework for large-scale content delivery. Comput Netw **57**, 924–943 (2013)
27. P Mendes, Combining data naming and context awareness for pervasive networks. J Netw Comput Appl **50**, 114–125 (2015)
28. V Sourlas, I Psaras, L Saino et al., Efficient Hash-routing and Domain Clustering Techniques for Information-Centric Networks. Comput Netw **103**, 67–83 (2016)
29. X Li, X Wang, S Xiao, VCM Leung, *Delay performance analysis of cooperative cell caching in future mobile networks* (2015 IEEE International Conference on Communications (ICC), London, 2015), pp. 5652–5657
30. X Wang, M Chen, T Taleb, A Ksentini, VCM Leung, Cache in the air: exploiting content caching and delivery techniques for 5G systems. IEEE Commun Mag **52**(2), 131–139 (2014)
31. C Yang, Z Chen, Y Yao, B Xia, *Energy efficiency analysis for wireless heterogeneous networks with pushing and caching* (2015 IEEE Wireless Communications and Networking Conference (WCNC), New Orleans, 2015), pp. 123–128
32. MS ElBamby, M Bennis, W Saad, M Latva-aho, *Content-aware user clustering and caching in wireless small cell networks* (2014 11th International Symposium on Wireless Communications Systems (ISWCS), Barcelona, 2014), pp. 945–949
33. E Baştuğ, JL Guénégo, M Debbah, *Proactive small cell networks* (Telecommunications (ICT), 2013 20th International Conference on, Casablanca, 2013), pp. 1–5
34. K Poularakis, G Iosifidis, L Tassiulas, Approximation Algorithms for Mobile Data Caching in Small Cell Networks. IEEE Trans Commun **62**(10), 3665–3677 (2014)
35. J Li, Y Chen, Z Lin, W Chen, B Vucetic, L Hanzo, Distributed Caching for Data Dissemination in the Downlink of Heterogeneous Networks. IEEE Trans Commun **63**(10), 3553–3568 (2015)
36. L Jiang, G Feng, S Qin, *Cooperative content distribution for 5G systems based on distributed cloud service network* (2015 IEEE International Conference on Communication Workshop (ICCW), London, 2015), pp. 1125–1130
37. Z Zhao, M Peng, Z Ding, W Wang, HV Poor, Cluster Content Caching: An Energy-Efficient Approach to Improve Quality of Service in Cloud Radio Access Networks. IEEE J Sel Areas Commun **34**(5), 1207–1221 (2016)
38. Z Zhao, S Jia, Y Li, M Peng, C Wang, *Performance Analysis of Cluster Content Caching in Cloud-Radio Access Networks* (2015 IEEE Globecom Workshops (GC Wkshps), San Diego, 2015), pp. 1–6
39. H Xiang, M Peng, Y Cheng, HH Chen, *Joint mode selection and resource allocation for downlink fog radio access networks supported D2D* (11th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness(QSHINE 2015), Taipei, 2015), pp. 177–182
40. SC Hung, H Hsu, SY Lien, KC Chen, Architecture Harmonization Between Cloud Radio Access Networks and Fog Networks. IEEE Access **3**, 3019–3034 (2015)
41. X Li, X Wang, C Zhu, W Cai, VCM Leung, *Caching-as-a-Service: Virtual caching framework in the cloud-based mobile networks* (2015 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Hong Kong, 2015), pp. 372–377
42. R Friedman, *Caching web services in mobile ad-hoc networks: opportunities and challenges[C], Proceedings of the second ACM international workshop on Principles of mobile computing, ACM*, 2002, pp. 90–96
43. Y Wang, J Wu, M Xiao, *Hierarchical cooperative caching in mobile opportunistic social networks* (2014 IEEE Global Communications Conference, Austin, 2014), pp. 411–416
44. SE El Khawaga, AI Saleh, HA Ali, An Administrative Cluster-based Cooperative Caching (ACCC) strategy for Mobile Ad Hoc Networks. J Netw Comput Appl **69**, 54–76 (2016)
45. NE Majd, S Misra, R Tourani, *Split-Cache: A holistic caching framework for improved network performance in wireless ad hoc networks* (2014 IEEE Global Communications Conference, Austin, 2014), pp. 137–142
46. J Iqbal, P Giaccone, *Interest-based cooperative caching in multi-hop wireless networks* (2013 IEEE Globecom Workshops (GC Wkshps), Atlanta, 2013), pp. 617–622

47. K Ikkaku, Y Sakaguchi, M Yamamoto, *In-network guide performance in wireless multi-hop cache networks* (Network Operations and Management Symposium (APNOMS), 16th Asia-Pacific, Hsinchu, 2014), pp. 1–6
48. A Rao, P Kumar, N Chauhan, *Energy efficient dynamic group caching in mobile Ad hoc networks for improving data accessibility* (Recent Trends In Information Technology (ICRTIT), 2012 International Conference on, Chennai, 2012), pp. 372–376
49. B Liu, Z Liu, D Towsley, *"On the capacity of hybrid wireless networks," INFOCOM 2003. Twenty-Second Annual Joint Conference of the IEEE Computer and Communications. IEEE Societies, San Francisco, CA, vol.2*, 2003, pp. 1543–1552
50. The GreenICN project. Available at: http://www.greenicn.org
51. M Dehghan, A Seetharamz, T He, T Salonidis, J Kurose, D Towsley, *Optimal Caching and Routing in Hybrid Networks* (2014 IEEE Military Communications Conference, Baltimore, 2014), pp. 1072–1078
52. S Eum, Y Shoji, M Murata, N Nishinaga, *Design of ICN-enabled IEEE 802.11 wireless access points* (Networks and Communications (EuCNC), 2014 European Conference on, Bologna, 2014), pp. 1–5
53. W Wang, X Wu, L Xie, S Lu, Joint storage assignment for D2D offloading systems. Comput Commun **83**, 45–55 (2016)
54. W Quan, C Xu, J Guan, H Zhang, LA Grieco, Social cooperation for information-centric multimedia streaming in highway VANETs, in *IEEE 15th International Symposium on a World of Wireless, Mobile and Multimedia Networks (WoWMoM), IEEE*, 2014, pp. 1–6
55. C Anastasiades, T Schmid, J Weber, T Braun, Opportunistic content-centric data transmission during short network contacts, in *IEEE Wireless Communications and Networking Conference (WCNC), IEEE*, 2014, pp. 2516–2521
56. M Zhang, H Luo, H Zhang, A Survey of Caching Mechanisms in Information-Centric Networking. IEEE Commun Surv Tutorials **17**(3), 1473–1499 (2015). third quarter
57. G Zhang, L Yang, T Lin, Caching in information centric networking: A survey. Comput Netw **57**(16), 3128–3141 (2013)
58. K Cho, M Lee, K Park, TT Kwon, Y Choi, S Pack, *WAVE: Popularity-based and collaborative in-network caching for content-oriented networks* (Computer Communications Workshops (INFOCOM WKSHPS), 2012 IEEE Conference on, Orlando, 2012), pp. 316–321
59. FR Dogar, A Phanishayee, H Pucha et al., *Ditto: a system for opportunistic caching in multi-hop wireless networks[C], Proceedings of the 14th ACM international conference on Mobile computing and networking, ACM*, 2008, pp. 279–290
60. JM Hsu, HY Chiu, YS Ye, *A Partial Cache for Multimedia Content in Named Data Networking* (Platform Technology and Service (PlatCon), 2015 International Conference on, Jeju, 2015), pp. 37–38
61. H Ko, Y Kim, D Suh, S Pack, *A proactive content pushing scheme for provider mobility support in information centric networks* (2014 IEEE 11th Consumer Communications and Networking Conference (CCNC), Las Vegas, 2014), pp. 523–524
62. T Wei, L Chang, B Yu, J Pan, *MPCS: A mobility/popularity-based caching strategy for information-centric networks* (2014 IEEE Global Communications Conference, Austin, 2014), pp. 4629–4634
63. V Martina, M Garetto, E Leonardi, *A unified approach to the performance analysis of caching systems* (IEEE INFOCOM 2014 - IEEE Conference on Computer Communications, Toronto, 2014), pp. 2040–2048
64. K Psounis, B Prabhakar, "A randomized Web-cache replacement scheme," INFOCOM 2001. Twentieth Annual Joint Conference of the IEEE Computer and Communications Societies. Proceedings. IEEE, Anchorage, AK **3**, 1407–1415 (2001)
65. J Liu, G Wang, T Huang, J Chen, Y Liu, Modeling the sojourn time of items for in-network cache based on LRU policy. China Commun **11**(10), 88–95 (2014)
66. D Lee, J Choi, JH Kim, SH Noh, SL Min, Y Cho, CS Kim, LRFU: A spectrum of policies that subsumes the least recently used and least frequently used policies. IEEE Trans Comput **50**(12), 1352–1361 (2001)
67. EJ O'Neil, PE O'Neil, G Weikum, The LRU-K page replacement algorithm for database disk buffering, in *Proc. ACM SIGMOD Conf*, 1993, pp. 297–306
68. M Bilal, SG Kang, *Time Aware Least Recent Used (TLRU) cache management policy in ICN* (16th International Conference on Advanced Communication Technology, Pyeongchang, 2014), pp. 528–532
69. H Gomaa, GG Messier, C Williamson, R Davies, Estimating Instantaneous Cache Hit Ratio Using Markov Chain Analysis. IEEE/ACM Trans Networking **21**(5), 1472–1483 (2013)
70. J Li, B Liu, H Wu, Energy-Efficient In-Network Caching for Content-Centric Networking. IEEE Commun Lett **17**(4), 797–800 (2013)
71. N Megiddo, D Modha, Outperforming LRU with an adaptive replacement cache algorithm. Computer **37**(4), 58–65 (2004)
72. H Che, Y Tung, Z Wang, Hierarchical Web caching systems: modeling, design and experimental results. IEEE J Sel Areas Commun **20**(7), 1305–1314 (2002)
73. P Rodriguez, C Spanner, EW Biersack, Analysis of Web caching architectures: hierarchical and distributed caching. IEEE/ACM Trans Networking **9**(4), 404–418 (2001)
74. PR Jelenković, X Kang, Characterizing the miss sequence of the LRU cache [J]. ACM SIGMETRICS Perform Eval Rev **36**(2), 119–121 (2008)
75. EJ Rosensweig, J Kurose, D Towsley, *Approximate Models for General Cache Networks* (INFOCOM, 2010 Proceedings IEEE, San Diego, 2010), pp. 1–9
76. AV Aho, PJ Denning, JD Ullman, Principles of Optimal Page Replacement. J ACM **18**(1), 80–93 (1971)
77. L Breslau, P Cao, L Fan, G Phillips, S Shenker, Web caching and Zipf-like distributions: evidence and implications. Proceedings of IEEE INFOCOM, NewYork, USA **1**, 126–134 (1999)
78. M Dehghan, B Jiang, A Dabirmoghaddam et al., *On the Analysis of Caches with Pending Interest Tables[C]//Proceedings of the 2nd International Conference on Information-Centric Networking. ACM*, 2015, pp. 69–78
79. N Gast, B Van Houdt, *Transient and steady-state regime of a family of list-based cache replacement algorithms[C]//Proceedings of the 2015 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems. ACM*, 2015, pp. 123–136
80. L Muscariello, G Carofiglio, M Gallo, *Bandwidth and storage sharing performance in information centric networking[C]//Proceedings of the ACM SIGCOMM workshop on Information-centric networking. ACM*, 2011, pp. 26–31
81. G Carofiglio, M Gallo, L Muscariello, D Perino, *Modeling data transfer in content-centric networking* (Teletraffic Congress (ITC), 2011 23rd International, San Francisco, 2011), pp. 111–118
82. G Carofiglio, M Gallo, L Muscariello, On the performance of bandwidth and storage sharing in information-centric networks [J]. Comput Netw **57**(17), 3743–3758 (2013)
83. EJ Rosensweig, DS Menasche, J Kurose, *On the steady-state of cache networks* (INFOCOM, 2013 Proceedings IEEE, Turin, 2013), pp. 863–871
84. I Psaras et al., Modelling and Evaluation of CCN-Caching Trees. Lect Notes Comput Sci **6640**, 78–91 (2011)
85. W Quan, J Guan, S Jia, J Zhu, C Xu, H Zhang, *A content retrieval model for Information Centric MANETs: 1-dimensional case* (2012 IEEE Globecom Workshops, Anaheim, 2012), pp. 1010–1015
86. W Quan, J Guan, C Xu, S Jia, J Zhu, H Zhang, *Content retrieval model for information-center MANETs: 2-dimensional case* (2013 IEEE Wireless Communications and Networking Conference (WCNC), Shanghai, 2013), pp. 4422–4427
87. A Detti, D Tassetto, N Blefari Melazzi, F Fedi, Exploiting content centric networking to develop topic-based, publish–subscribe MANET systems. Ad Hoc Netw **24**, 115–133 (2015)
88. N Choi, K Guan, DC Kilper, G Atkinson, *In-network caching effect on optimal energy consumption in content-centric networking* (2012 IEEE International Conference on Communications (ICC), Ottawa, 2012), pp. 2889–2894
89. W Li, E Chan, D Chen, *Energy-Efficient Cache Replacement Policies for Cooperative Caching in Mobile Ad Hoc Network* (2007 IEEE Wireless Communications and Networking Conference, Kowloon, 2007), pp. 3347–3352
90. A Mohammed, K Okamura, *Distributed GA for popularity based partial cache management in ICN[C]//Proceedings of The Ninth International Conference on Future Internet Technologies. ACM*, 2014, p. 18
91. S Gitzenis, S Toumpis, L Tassiulas, *Efficient file replication in large wireless networks with dynamic popularity* (Heterogeneous Networking for Quality, Reliability, Security and Robustness (QShine), 2014 10th International Conference on, Rhodes, 2014), pp. 164–168
92. S Gitzenis, GS Paschos, L Tassiulas, Enhancing wireless networks with caching: Asymptotic laws, sustainability & trade-offs. Comput Netw **64**, 353–368 (2014)
93. J Zhang, X Zhang, W Wang, Cache-Enabled Software Defined Heterogeneous Networks for Green and Flexible 5G Networks. IEEE Access **4**, 3591–3604 (2016)
94. M Mangili, F Martignon, A Capone, Performance analysis of Content-Centric and Content-Delivery networks with evolving object popularity. Comput Netw **94**, 80–98 (2016)
95. F Chao, F Richard Yu, H Tao, L Jiang, L Yunjie, An energy-efficient distributed in-network caching scheme for green content-centric networks. Comput Netw **78**, 119–129 (2015)

96. M Hajimirsadeghi, NB Mandayam, A Reznik, *Joint Caching and Pricing Strategies for Popular Content in Information Centric Networks[J]*, 2016. arXiv preprint arXiv:1609.00852
97. U Lee, I Rimac, D Kilper, V Hilt, Toward energy-efficient content dissemination. IEEE Netw **25**(2), 14–19 (2011)
98. B Han, X Wang, N Choi, T Kwon, Y Choi, *AMVS-NDN: Adaptive mobile video streaming and sharing in wireless named data networking* (Computer Communications Workshops (INFOCOM WKSHPS), 2013 IEEE Conference on, Turin, 2013), pp. 375–380
99. V Sourlas, P Flegkas, P Georgatsos, L Tassiulas, *Cache-aware traffic engineering in Information-Centric Networks* (2014 IEEE 19th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks (CAMAD), Athens, 2014), pp. 295–299
100. V Sourlas, L Tassiulas, *Effective cache management and performance limits in information-centric networks* (Computing, Networking and Communications (ICNC), 2013 International Conference on, San Diego, 2013), pp. 955–960
101. K Cho, H Jung, M Lee, D Ko, T Kwon, Y Choi, How can an ISP merge with a CDN? IEEE Commun Mag **49**(10), 156–162 (2011)
102. M Amadeo, A Molinaro, G Ruggeri, *An energy-efficient content-centric approach in mesh networking* (2012 IEEE International Conference on Communications (ICC), Ottawa, 2012), pp. 5736–5740
103. K Hamidouche, W Saad, M Debbah, *Many-to-many matching games for proactive social-caching in wireless small cell networks* (Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt), 2014 12th International Symposium on, Hammamet, 2014), pp. 569–574
104. HJ Kang, KY Park, K Cho, CG Kang, *Mobile caching policies for device-to-device (D2D) content delivery networking* (2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS), Toronto, 2014), pp. 299–304
105. G Gür, Energy-aware cache management at the wireless network edge for information-centric operation. J Netw Comput Appl **57**, 33–42 (2015)
106. SH Shen, A Akella, *An information-aware QoE-centric mobile video cache[C]// Proceedings of the 19th annual international conference on Mobile computing & networking. ACM*, 2013, pp. 401–412
107. Z Ming, M Xu, D Wang, InCan: In-network cache assisted eNodeB caching mechanism in 4G LTE networks. Comput Netw **75**, 367–380 (2014)
108. A Detti, B Ricci, N Blefari-Melazzi, Mobile peer-to-peer video streaming over information-centric networks. Comput Netw Int J Comput Telecommun Netw **81**, 272–288 (2015)
109. F Malabocchia, R Corgiolu, M Martina, A Detti, B Ricci, N Blefari-Melazzi, *Using Information Centric Networking for Mobile Devices Cooperation at the Network Edge* (2015 IEEE 81st Vehicular Technology Conference (VTC Spring), Glasgow, 2015), pp. 1–6
110. M Amadeo, C Campolo, A Molinaro, Information-centric networking for connected vehicles: a survey and future perspectives. IEEE Commun Mag **54**(2), 98–104 (2016)
111. S Wood et al., *ICEMAN: A System for Efficient, Robust and Secure Situational Awareness at the Network Edge* (MILCOM 2013 - 2013 IEEE Military Communications Conference, San Diego, 2013), pp. 1512–1517
112. D Niyato, DI Kim, P Wang, L Song, *A novel caching mechanism for Internet of Things (IoT) sensing service with energy harvesting* (2016 IEEE International Conference on Communications (ICC), Kuala Lumpur, 2016), pp. 1–6
113. E Borgia, R Bruno, M Conti, D Mascitti, A Passarella, *Mobile edge clouds for Information-Centric IoT services* (2016 IEEE Symposium on Computers and Communication (ISCC), Messina, 2016), pp. 422–428
114. KV Katsaros, WK Chai, N Wang, G Pavlou, H Bontius, M Paolone, Information-centric networking for machine-to-machine data delivery: a case study in smart grid applications. IEEE Netw **28**(3), 58–64 (2014)
115. KV Katsaros, WK Chai, B Vieira, G Pavlou, *Supporting smart electric vehicle charging with information-centric networking* (10th International Conference on Heterogeneous Networking for Quality, Reliability, Security and Robustness, Rhodes, 2014), pp. 174–179
116. AS Gomes et al., *A mobile follow-me cloud content caching model* (NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium, Istanbul, 2016), pp. 763–766
117. Q Liang, X Cheng, SC Huang, D Chen, Opportunistic Sensing in Wireless Sensor Networks: Theory and Application. IEEE Trans Comput **63**(8), 2002–2010 (2014)
118. B Li, J Hou, X Li, Y Nan, A Nallanathan, C Zhao, Deep Sensing for Space-Time Doubly Selective Channels: When a Primary User Is Mobile and the Channel Is Flat Rayleigh Fading. IEEE Trans Signal Process **64**(13), 3362–3375 (2016)
119. B Li, S Li, A Nallanathan, C Zhao, Deep Sensing for Future Spectrum and Location Awareness 5G Communications. IEEE J Sel Areas Commun **33**(7), 1331–1344 (2015)
120. C Fang, FR Yu, T Huang, J Liu, Y Liu, A Survey of Green Information-Centric Networking: Research Issues and Challenges. IEEE Commun Surv Tutorials **17**(3), 1455–1472 (2015). third quarter
121. M Amadeo, C Campolo, A Molinaro, G Ruggeri, Content-centric wireless networking: A survey. Comput Netw **72**, 1–13 (2014)
122. G Alfano, M Garetto, E Leonardi, Content-Centric Wireless Networks With Limited Buffers: When Mobility Hurts. IEEE/ACM Trans Networking **24**(1), 299–311 (2016)
123. M Papandrea, KK Jahromi, M Zignani, S Gaito, S Giordano, G Paolo Rossi, On the properties of human mobility. Comput Commun **87**, 19–36 (2016)
124. H Wang, S Chen, H Xu, M Ai, Y Shi, *SoftNet: A Software Defined Decentralized Mobile Network Architecture toward 5G, IEEE Network*, 2015, pp. 16–22
125. IF Akyildiz, SC Lin, P Wang, Wireless software-defined networks (W-SDNs) and network function virtualization(NFV) for 5G cellular systems: An overview and qualitative evaluation. Comput Netw **93**, 66–79 (2015)
126. Z Feng, C Qiu, Z Feng, Z Wei, W Li, P Zhang, An effective approach to 5G: Wireless network virtualization. IEEE Commun Mag **53**(12), 53–59 (2015)
127. C Liang, FR Yu, X Zhang, Information-centric network function virtualization over 5 g mobile wireless networks. IEEE Netw **29**(3), 68–74 (2015)
128. R Ravindran, X Liu, A Chakraborti, X Zhang, G Wang, *Towards software defined ICN based edge-cloud services* (2013 IEEE 2nd International Conference on Cloud Networking (CloudNet), San Francisco, 2013), pp. 227–235
129. A El-Mougy, M Ibnkahla, L Hegazy, *Software-defined wireless network architectures for the Internet-of-Things* (2015 IEEE 40th Local Computer Networks Conference Workshops (LCN Workshops), Clearwater Beach, 2015), pp. 804–811
130. C Yang, Z Chen, B Xia, J Wang, *When ICN Meets C-RAN for HetNets: An SDN Approach, IEEE Communications Magazine*, 2015, pp. 118–125
131. J Tong, R Pi, K Xu, *Cloud computing infrastructure based on named content* (2011 6th International Conference on Pervasive Computing and Applications, Port Elizabeth, 2011), pp. 429–434