**Chapter #:** **Enhancing Mobile Data Offloading with In-Network Caching**

**Xu Zhang, Xi’an University of Technology, China**

**Yue Cao, Northumbria University, UK**

**Linyu Peng, Waseda University, Japan**

Contents

[1. Introduction 1](#_Toc515287822)

[2. Mobile Data Offloading Open Research Issues 2](#_Toc515287823)

[2.1 Wi-Fi Offloading 3](#_Toc515287824)

[2.2 Small Cell Offloading 4](#_Toc515287825)

[2.2 BS Offloading 5](#_Toc515287826)

[2.3 D2D offloading 6](#_Toc515287827)

[2.4 Mobile Data Offloading Efficiency 8](#_Toc515287828)

[3. In-network Caching for Mobile Content Delivery 9](#_Toc515287829)

[3.1 A Brief Survey on Information-Centric Networking 9](#_Toc515287830)

[3.1.1 Overview of the ICN Architecture 11](#_Toc515287831)

[3.2 In-Network Caching Approaches 12](#_Toc515287832)

[3.2.1 Independent in-network Caching Scheme 13](#_Toc515287833)

[3.2.2 Cooperative in-network Caching Scheme 14](#_Toc515287834)

[3.2.3 Cache Replacement Policies 16](#_Toc515287835)

[3.3 In-network Caching Efficiency 16](#_Toc515287836)

[4. Incentives and Challenges 18](#_Toc515287837)

[5. Summary and Conclusions 20](#_Toc515287838)

[References 20](#_Toc515287839)

# 1. Introduction

This chapter discusses the most recent solutions for accelerating mobile data delivery, focusing on the different issues for achieving this goal. These issues include the efficiency of mobile data offloading, the impact of in-network caching on the alleviation of ever-growing mobile content demands, incentives of how these techniques could contribute to 5G research and finally the challenges in bringing the concept into reality.

Mobile data offloading enables to dynamically deliver much of the cellular traffic through complementary wireless technologies. These techniques utilize cost-effective options to provide ubiquitous connectivity for mobile Internet access. Various open research areas where mobile data offloading will show great efficiency are explained. This includes the eNodeB or base station (BS) offloading that makes use of existing cellular infrastructures (i.e., BSes), the Wi-Fi and Femtocell offloading that deploy cost-effective facilities (i.e., access points), and device-level offloading (i.e., device-to-device (D2D) offloading) that utilizes end users’ equipment (e.g., smartphones and tablets, etc.). Offloading efficiency are discussed to show the effectiveness specific operation regimes can achieve.

There are great concerns on tightly coupled dependency on original data sources, and mobile users could easily suffer from the scarcity of content object resources. Typical examples are P2P-like content delivery networks, where end users (or user devices) are both consumers and contributors. By additionally accounting for such issues, researchers further account for the possibility of decoupling content availability dependency from original content sources. And a key enabling factor to achieve such goal could be the recently emerged technology, i.e., in-network caching, whereby the emerging Information-Centric Networking (ICN) performs its critical functionalities.

Coupling ICN with mobile cellular network could be an attractive solution especially for the 5G network [1][2], in order to resolve the tension between the tremendous growth of mobile content traffic and backhaul bottleneck. The implications of in-network caching on network efficiency are discussed, highlighting various approaches focuses on the different issues for enhancing the mobile data offloading efficiency. These approaches include independent in-network caching scheme and cooperative caching approach. Some of the main challenges and incentives are elucidated. The chapter finally summarizes there research efforts and provides an overview on the mobile content delivery enhanced with offloading coupled with ICN techniques.

Service providers aim to provide *pervasive connectivity* nowadays, whereby *service velocity* could be substantially improved by enabling services at the network edge. Researchers as well as service providers have commonly agreed that offloading along with in-network caching enables great performance gains for the above two goals. And the following discussions in this chapter are based on the two aspects.

# 2. Mobile Data Offloading Open Research Issues

Cellular traffic data has been observed a dramatic increase globally. The proliferation of mobile device (e.g., smartphones, tablets) plays a key role in such massive growth, largely driven by the popularity of online social networking services. In fact, end users are more than ever in huger for *data-intensive* applications, e.g., video streaming, etc. However, it is commonly believed that the cellular capacity is being overloaded already and advanced cellular technologies (4G, LTE) cannot scale up enough to satisfy such appetite [4].

To ease the pain of carriers’ concerns on constrained mobile networks, a cost-effective approach is to automatically *offload* cellular data traffic onto small-cells (e.g., Femtocells) and WiFi networks via complementary network technologies [7]-[9]. With offload, an optimal network available is selected (e.g., examined and chosen by the mobile device), and the subscriber is automatically connected and authenticated. Mobile offload enables to free up cellular resources especially in high-traffic venues, e.g., city centers and airports. And the ubiquitous connectivity demanded by customers can be supported by offloading techniques with fast downloading speed. The focus of this section is on mobile data offloading open research issues. The most representative approaches are discussed, including pros and cons.

## 2.1 Wi-Fi Offloading

According to Cisco reports, more than half of total mobile data traffic was offloaded through WiFi or Femtocells in 2016 and will reach 63% by 2021 [4]. WiFi offers high data rates with low cost that normally works on unlicensed spectrum. And hence, there is no interference with existing cellular networks. WiFi offload has been recognized by the wireless industry as an effective as well as economical approach to satisfy growing customer demand for mobile data.

Actually, mobile network operators have deployed a number of WiFi access points (AP) [10][11], such as AT&T, Vodafone and Orange. The integration of WiFi into the mobile network is becoming essential so as to maximize the benefits of both. As illustrated in Fig. 1, by offloading cellular data onto WiFi networks, such as content (or video) data and even voice services, more network capacity is able to be added in an affordable and flexible way. Customers are enjoying good service experiences as well, such as desirable downloading speed and better voice services especially indoors. In fact, most WiFi-enabled mobile devices are designed to favor WiFi connections whenever available.



Figure 1. A simple illustration of WiFi offload paradigm

Integration of WiFi networks with the mobile core will become the mass-market business opportunity for upcoming 5G network operators. Clearly, it is a strong trend that 5G and Wi-Fi will be integrated into one service. Incentives for WiFi offload are mainly driven by the following well-known factors.

* **Cost-effectiveness**: To accommodate adequate data service, many more base stations (BSs) or access points are required. Compared to costly BSs deployment (e.g., expensive real-estate cost), carrier-class WiFi deployments allow to save massive capital cost for mobile operators, up to 40% on network CAPEX according to Ref. [12].
* **Smart-driven devices**: The widespread WiFi-capable devices favor WiFi networks when available. And these smart devices are capable of choosing the best network available intelligently and dynamically. The intelligence in the devices enables to constantly measure the connection quality, and to actively interact with mobile core to make connection shift decisions.

By enabling Wi-Fi access with mobile core integration, WiFi network can be treated as an alternative Radio Access Network (RAN), with full mobility. Offloaded data traffic is backhauled to the mobile core through GPRS Tunneling (GTP) protocol. With this approach, WiFi traffic is treated as mobile broadband and service operators can have complete control over subscribers’ connection to WiFi networks. This means that the existing cellular policy control or charging mechanisms can all be applied. Therefore, personalized services over WiFi networks are able to be offered with enhanced customer experiences. This may require to integrate policy and charging with the Wi-Fi service management system as well.

## 2.2 Small Cell Offloading

Unlike WiFi, small cell (e.g., Femtocell) operates on licensed spectrum via a low-power cellular base station with low-range coverage (radius up to 10 m). In comparison, the macro-cell is controlled by the base station with wider range (radius up to 30 km) and higher power, as illustrated in Fig. 2. Since the same licensed frequency band is shared by both small cell and macro-cell, no additional physical support is required at the side of user equipment, such as a WiFi radio unit. However, there is interference between Femtocell and macro network, and between nearby Femtocells, as well.



Figure 2. Small cell offloading illustration

As a typical example, Femtocell is designed as a solution to offer better voice and data services [7] for poor indoor performance limited by macro-cell. With backhaul, Femtocell connects to the mobile network operator usually through broadband Internet connection (e.g., cable modem or DSL). Femtocells are usually deployed by customers, and macro cell BS does not have to actively oversee Femtocell BS so this is practically suitable for large deployment. Therefore, femtocell is considered to be energy-efficient and of high security.

With small cells offloading, cellular data traffic is diverted onto small cell BS via wireless access points, which is similar to WiFi offload. However, due to different broadband frequencies adopted, they differ in the following aspects.

* **Operators guaranteed service quality**: Since small cells operates on licensed spectrum same as cellular networks, operators can have full control over the traffic and offer premium services.
* **Interference**: Since the same spectrum channels are shared between macro and small cells, coordination between them is needed to mitigate such conflict.

Often, small cells and WiFi work together especially in locations with high traffic demands, such as dense metro areas (e.g., airports and city centers), with small cells more frequently deployed outdoors and Wi-Fi indoors. The marginal cost of deploy a WiFi access point or a small cell is very low, so it is feasible to have Wi-Fi built-in modules for small cells.

## 2.2 BS Offloading

With locally cached popular contents，the potential of utilizing existing cache facilities at RAN is expected to satisfy users’ demands more efficiently. Specifically, cache-enabled BSs (or evolved NodeBs, eNBs) play key role in data offloading at RAN [13]. Many research works have exploited RAN caching at eNBs that have shown great efficiency on traffic load reduction at mobile core [2].



Figure 3. A simple example of BS offloading

BSs offloading provides desirable service experiences for customers, and enables to effectively alleviate backhaul congestion issues with intelligent content placement planning. Content objects are often split into data chunks that can be delivered via different paths to the requestor. As will be discussed in Section 3, eNBs are required to be equipped with content-aware functionality if equipped with the in-network caching capability [2][5][6]. As such, eNBs or BSs are capable of intercepting and caching passing-through data items.

Fig. 3 illustrates a simple example of eNB (or BS) offloading paradigm, wherein cache-enabled eNBs are incorporated at RAN. Without caching capability at eNB, content requests generated from user equipment (UE) have to reach far into the Internet through mobile core network (CN), where the content provider (CP) located. As compared, UEs can be satisfied locally at eNBs with great feasibility if the demanded data items are fully (or partially) cached. And hence the backhaul traffic is considerably alleviated given ever-growing UE demands on mobile data.

For example in Fig. 3, assuming that the content object is requested at UE side, which is originally delivered from CP by splitting into several data chunks. Suppose eNB *A* has cached several chunks of the content object. Mobile users subscribing to eNB *B* request for the content object. Instead of fetching all content chunks from the original source, content demand could be partially served locally by nearby eNB *A* alternatively, via coordination between eNB *A* and eNB *B*.

BS offloading can be highly effective if the local eNB has knowledge of caching states at nearby eNBs [2]. To facilitate such capability, coordination between adjacent eNBs needs to be enabled, which can be realized through the X2 interface [14]. With cooperation, service providers are able to achieve the following performance goals as compared to independent BS offloading.

* Lower access delay for customers without resorting to further remote source CP
* Elimination of potential bottlenecks on backhaul links
* Reduction on cache redundancy among cache-enabled eNBs

It is commonly believed that a transformed mobile base station (e.g., BS offload-enabled) can greatly relive the hunger for computational capacity and meet the needs for robust networking.

## 2.3 D2D offloading

Cellular offloading via device-to-device (D2D) communication can be a very enhancement technique for data distribution in 5G networks. Different from other offloading options dependent more or less on cellular infrastructure, direct communication between UEs in proximity is enabled without use of the mobile network facilities. By forming into a device-to-device (D2D) networking group, UEs operate as ad hoc relays in the following two ways: a) either to connect to non-congested BSs [15][16] or b) to cooperatively download the content. By concept, the BS unicasts original data items of the content to a selected group of UEs, which then multicast them to each other over local ad hoc networks using multi-hop cooperation.

Due to the opportunistic communication nature with D2D, non-real time content delivery is often offloaded for delay-tolerant networks. And D2D in 5G may even allow for better service experiences, in terms of decreased communication latency and higher downloading rate with reduced energy consumption. Possible applications are thus driven by user proximity, such as social networking and local exchange of information (e.g., advertisements and vehicular communication between smart cars). In particular, D2D networking could be a key enabling factor in the support for national security/public safety in case of not-spot network coverage area, where local connectivity is provided by devices at least.

D2D links can operate either on licensed bands or unlicensed bands. Unlicensed band D2D protocols operates on WiFi-direct or bluetooth are without network control, while protocols on licensed bands are with the help of cellular networks. With network-assistance, a centralized infrastructure (e.g., orchestrated by BSs) assists and controls the operation of D2D communication [15][16].

For example in Fig. 4, network control is applied when D2D shares with cellular the same licensed spectrum bands. With network assistance, device discovery is managed by the controller. The D2D link quality effect can be detected by the smart device locally, which is then reported to the serving controller (e.g., BS) when an offload mode shift need is detected. In comparison, with unlicensed spectrum adopted, D2D links can be established without centralized control. However, the discovery process can be very time and energy consuming in this way.

In fact, mobile devices are becoming smart capable of decision-making on switching between different access types (e.g., Wi-Fi or cellular). However, with network-assistance, customers could experience benefits especially on minimizing energy consumption and accelerating device discovery [17]. And operators can also gain from such functionality for concerns on traffic management, policy, charging and security, etc.



Figure 4. An example of cellular offloading via D2D communication

With network control, D2D communication could achieve potential gains as the following.

* **Energy and time consumption gain**: D2D paring and device discovery could be significantly improved with the assistance of network infrastructure, e.g., BSs. For example, the scan for other wireless technologies can be avoided, the transmission and reception of discovery signals can be synchronized to save UEs frequent information exchange.
* **Capacity gain**: By enabling spectrum resources sharing with cellular networks, the network capacity can be greatly improved with spectrum efficiently shared among multi- concurrent D2D links. By confining radio transmissions to the D2D connection, even higher spectrum reuse gain can be achieved as compared to LTE small cells.
* **Data rate and Latency gain**: High peak rates can be achieved when devices in proximity and actively engaged in data propagation. Since devices communicate over a direct link, the end-to-end latency may be reduced as compared to cellular communication, while a busy BS can be the bottleneck.

With D2D offloading, cooperation among mobile devices can be made within a single hop range or to multi-hops extent. However, one hop cooperation proves to be computationally efficient. And a tradeoff between delay and offloading extent exists, which proves to be NP-complete. To boost performance gains, D2D offloading bundled with BS offloading can significantly alleviate traffic pressure over cellular networks.

## 2.4 Mobile Data Offloading Efficiency

An efficient offloading approach is able to maximize the offloaded cellular traffic volume with improvement on performance gains, in terms of access delay, user data rate, network capacity and energy consumption, etc. The criteria is tightly coupled with each other and tradeoffs exist between one or two metrics as discussed previously. And thus a discussion on specific performance metric is meaningless without account for careful designs on the following operation process.

* **When to offload**: Such issue has been discussed previously with respect to specific offload technique. To efficiently offload traffic from cellular networks, the key issue is to manage seamless *shift control between different offload options*, especially in a heterogeneous network scenario with various options co-existing (e.g., WiFi, small cell, BS and D2D offload). Such shift control can be performed by smart devices locally, or assisted by cellular infrastructure to save more energy on devices. When the delay users can tolerant is beyond a threshold, an offload shift will be triggered. In particular, the interworking of two options, such as WiFi offload integrated into cellular networks, will offer great performance gains. Meanwhile, when there are multiple offload options available, some level of strategy is need to hierarchically shift to a favored offload mode. Such strategized operation can be assisted by cellular network, so as to achieve energy and computation efficiency.
* **Where to offload**: This could be a concern particularly with D2D offload. The rational is that a group of D2D candidates needs to be selected prior to commencing D2D communication sessions, wherein the original data is delivered to the group via cellular communication. However, how to choose such target UE group can be a challenge with concerns on tradeoffs between desirable service experiences and network performance gains. With network assistance, the BS can help to choose a set of user devices identified as target-UEs. The optimal UE-set selection proves to be an NP-hard problem [18]. Typical target-set selection algorithms are well-known as Greedy, Heuristic and Random algorithm. If mobile users actively participate in delivering content data, the performance among different set selection algorithms is not very significant. As such, a simple Random algorithm is effective enough to achieve desirable performance benefits.
* **What to offload**: While alleviating traffic load from cellular networks, BSs and D2D offload need to consider *what to offer* to customers from their local cache. To intelligently provide such functionality, BSs and target-set UEs can be equipped with *in-network caching functionality*, which will be discussed in detail in Section 3. With such in-network replica intelligence, BSs (or target-UEs) are capable of intercepting passing through popular contents and selectively making cache decisions. When content items are pushed to UE level, the BS would assist D2D systems to select content items [15]. For instance, top popular content data items could be pushed by the BS to the target-UEs, a) by off-line operation with pre-selected content catalog, or b) by on-line procedure managed in real-time the push-decision, which could rely on the demand history recorded at the BS. As such, data items cached at a BS is able to serve subscribed customers, and users would be benefited from nearby UEs with local caches for direct data delivery via D2D communication. Details regarding this part will be elaborated in Section 3.

# 3. In-network Caching for Mobile Content Delivery

The incentives of mobile edge offloading (e.g., BS offloading, D2D offloading, etc.) could be enhanced greatly by the capability of ICN-based caching technologies, whereby caching at the core or edge network in 5G-based network environments can be supported by sophisticated caching mechanisms at the device-level.

Offloading along with in-network caching technologies are among the most popular solutions for enabling efficient multimedia content delivery services. Such a feature will substantially increase the network efficiency when more and more end users become actually mobile users with smartphones, and thus will substantially improve the service in future 5G environments, given that mobile data will substantially dominate in future content distribution applications.

## 3.1 A Brief Survey on Information-Centric Networking

As the rapid growth towards the popularity of content distribution applications, end users are more caring about the content itself that can be obtained from the Internet. Consequently, tremendous overlay traffic can be incurred over the network due to users’ appetite for data-intensive applications. In fact, the cellular networks is already overloaded especially by the mobile data traffic [4]. On the other hand, substantial access burden at the content provider side can be increased considerably due to their limited uploading bandwidths capacity, compared to the massive data demanded.

In order to provide more efficient content delivery, ICN [5][6] has therefore emerged as an optimized method to relieve the above issues through caching inside network elements, e.g., at cache-enabled eNodeBs. A key feature of ICN is to deploy *in-network caching* that content objects can be cached within the network for local access by future interested clients, so as to alleviate the content access burden from the content provider’s side. Namely, caching is deployed in ICN as an inherent network capability rather than an overlay service as conventional caching promoted, e.g., the Web caching or the CDN caching.

Generally, in-network caching policy can be categorized as two main types [20][34]: *independent* and *coordinated caching*. With independent caching, individual network elements make their own caching decisions based on local information. Within this scope, a caching everything everywhere policy is generally utilized [6], whereby the concept is to cache every data item retrieved at all network devices along the delivery path.

However, given the limited capacity of network devices in comparison to the large number of content overlay traffic, the management of content distribution among the device side, or the in-network caching is crucial. Some researches thus promote content distribution among intermediate nodes in a selective way [22], by taking into consideration some local information, such as cache capacity, in the caching decisions.

Coordinated caching [23]-[25], on the other hand, requires network elements to make caching decisions in a coordinated manner by sharing their state of local caching information with each other, in order to provision the cache capacity more efficiently and thus improve the network efficiency as well as users’ perceived service quality.

In the following subsections, the main ICN architecture is presented to give an overview of how in-network caching operates. Different caching policies will be elaborated and compared in the next section, followed by a more detailed discussion on the concerns of caching efficiency.

### 3.1.1 Overview of the ICN Architecture



Figure 1. ICN architecture overview

Next, the most representative ICN architecture is introduced by describing its main components to give an overview of the operation of an ICN-deployed network before bringing out in the next section its main idea of in-network caching.

Several architectures regarding the ICN concept have been proposed [5][6][26][27], among which Ref. [6] is the most representative paradigm that introduced content-centric networking (CCN) paradigm. The main components of a CCN node include a content store entity, a pending interest table (PIT) and a forwarding information base (FIB), as shown in Fig. 1. The Request Packet Process and the Data Packet Process are key functions wherein in-network caching is involved.

* **Request Packet Processing**

A data consumer, having access to the Internet via an access point (or network element) *A* in Fig. 1, requests for a content object located next to *D* by broadcasting its interest over all available connectivity, e.g., via *A-B-D* and *A-C-D* in Fig. 1 (the interest can be decomposed as several interests for small data chunks that constitutes the content object). Any intermediate network element along the delivery path can intercept the interest and locally serve it if it has a replica cached at the Content Store entity. A longest-match lookup on the Content Name in the Interest packet will perform the cache check. Otherwise, if the interest matches one of the PIT entries, simply the Interest’s arrival face will be recorded and the interest will not be forwarded upstream. Such operation is referred to as the interest aggregation so as to avoid repeated forwarding and to reduce traffic redundancy. If no matches can be found at PIT entries, the Interest packet would be forwarded upstream towards the source data if there is a matching FIB entry, and a new PIT entry is created for this Interest and also its arrival face. Otherwise, the Interest will be simply discarded if there is no match at all, since such interest could arise from malicious behaviours.

* **Data Packet Processing**

The Data packet retrieving is relatively simple as compared to the Interest packet processing. The Data packet simply follows the symmetric reverse path of the PIT entries back to the requester as illustrated in Fig. 1, e.g., via *D-B-A* and *D-C-A*. Only a PIT match along the delivery path can enable a caching decision upon an arrival of a Data packet. The reason is straightforward, since a Content Store match only means a duplicated data and thus the Data packet is discarded. A FIB match occurs when there is no PIT match, and it can arise from a malicious behaviour, and thus the Data packet is discarded as well. In particular, based on the CCN paradigm, each content chunk is cached at every cache along the delivery path, referred to as the *caching everything everywhere policy*. Such universal caching can certainly achieve a high caching performance in terms of desirable cache hits at individual caches due to its aggressive caching for each incoming content chunk. However, huge *cache redundancy* can exist since duplicated caching exist in the whole network. Besides, cache operational cost can as well arise due to great cache evictions incurred for each new incoming chunk. These issues will be discussed in detail in the next subsection.

It should be noted that in-network caching is a key feature in the context of ICN, which differentiates it from traditional caching schemes, such as the Web caching or CDN caching that deploy the caching as an overlay service. Rather, the ICN deploys caching as an inherent network capability in order to enable more efficient content distribution and achieve network efficiency. Henceforth, the understanding of the in-network caching is important to help to have an in-depth understanding of device-level mobile data offloading. Next different in-network caching policies are elaborated and detailed comparisons are also dicussed, in the purpose of providing a hint to design an optimal caching policy.

## 3.2 In-Network Caching Approaches

The motivation of in-network replica is to effectively alleviate the problem of increasingly constrained mobile networks through device-level data chunk caching. This technology has been introduced to optimise core network resources, and also to achieve better Quality of Services (QoS) including reduced content access time as well as end-to-end transmission delay. Researchers have explored intelligent algorithms and mechanisms for content chunk caching according to specific content delivery application characteristics such as content availability/popularity, as well as end user behaviours including group join/leaving and mobility patterns [3]. Optimised chunk caching and replacing algorithms are designed at the device level in order to best utilise the content caching resources.

Next, the in-network caching polices are elaborated to provide a general understanding. Note the caching policy is orthogonal to the cache replacement policy and thus the two can operate separately. Generally, in-network caching policy can be categorized as two main types: *independent* and *coordinated caching*.

### 3.2.1 Independent in-network Caching Scheme

Within the scope of independent caching, individual network elements independently make caching decisions based on local information, without awareness of caching state of others. The main representative approaches in this scope can be categorized into two groups: the *universal caching* approach [6][28][29], and the *selective caching* approach [3][22].

* **Universal Caching** - As shown in Fig. 2, the caching decision is made solely based on local information. Network elements randomly (or aggressively) caches incoming chunks if none of the incoming chunk has been stored at the cache currently [6][28][29]. And if the cache is fully occupied, cache replacement policy will be adopted to make room for the new one, which will be discussed later. In this way, each data item of the content object can be cached everywhere within the delivery scope as shown in Fig. 2. It should be noted that such caching policy makes caching decisions locally without awareness of duplicated caching elsewhere or content diversity property, e.g., chunk popularity degrees associated with different content objects.



Figure 2. An independent caching overview

* **Selective Caching** – Based on the universal caching approach, the network cache resources are not efficiently utilized, especially given the limited cache capacity at individual network devices. As a result, an aggressive caching can easily result in cache redundancy. Towards this end, efficient caching approaches are proposed [3][22].

A simple example of selective caching is a *fixed rate caching decision* approach [30]. And its caching logic is performed uniformly over the network element with a fixed probability for every incoming content chunk. For example, if the probability is set at 0.5 for each incoming chunk to be cached, each cache will store approximately half amount of the demanded content object at some point when the network is stable. As such, a certain cache space would be released for upcoming content chunks and the cache redundancy could be reduced to a certain level. Intuitively, higher set of the fixed probability value imply higher cache hits due to more chunks to be cached. And the probability equals to one based on the universal caching, wherein severe cache redundancy will be triggered.

Some research works are more concerning the issue of *where to cache.* They suggest to deploy in-network caches at some selected localities, such as inside access networks rather than at every cacheable network device [31], in order to localize overlay traffic to reduce mobile core network traffic and also to achieve the performance improvement of end users as well.

However, *what to cache* is not a big concern with in-network caching approaches as discussed above. And a simply universal caching policy is usually adopted. However, more strategized caching approaches [3][22] emerged aiming to achieve more efficient content distribution across the network. These policies concerns more about the content object features*.* Specifically, content popularity/availability [3], and context information at local (e.g., the cache capacity and cache locality) are taken into account to make caching decisions. With the consideration of cache capacity, content chunks tend to be stored at relatively larger caches along the delivery path. In case of multiple same size caches as candidates, the one located in proximity to requesters will be preferred. As such, the content access latency will be reduced and the network resources can be more efficiently managed, especially in the heterogeneous caches network scenario. For instance, if caches in Fig. 2 all have identical cache capacity, more popular content chunks will be cached at network devices closer to the consumer side based on this strategized caching policy, e.g., more chunks would be cached towards to cache *A* rather than at cache *D*.

### 3.2.2 Cooperative in-network Caching Scheme

However, a certain level of duplicated caching can still exist even if a carefully selective caching policy is adopted. The main rational is that there is no context information exchanged between individual caches regarding their cached contents. And the lack of coordination among caches can also result in the case that a request unable to be served at a cache would be forwarded towards the original source, while the demanded data is cached somewhere near to the cache already.

In the purpose of eliminating the cache redundancy so as to achieve more efficient in-network caching, a few research works have been engaged in the context of coordinated in-network caching efforts [2][32].

A key enabling factor of coordinated caching is cooperation among neighboring caches. By concept, network elements within the same community (e.g., e.g., caches along the same delivery path or network devices nearby) are able to share with each other caching status. This includes context information, such as the popularity degree of incoming content chunks and the list of cached contents. As such, content demands could be served in a whole neighborhood range rather than at a single cache. With awareness of such context information available, caches can leverage it when making caching decisions, such that cache redundancy can be further reduced.

Generally, coordinated caching enables significant improvement on caching performance in mainly two ways. Firstly, cache hits could be greatly increased by forwarding requests to nearby caches. And thus the access burden from content providers can be relieved considerably. And secondly, the coordination operation allows caches to be aware of cached states within the overall neighbours, such that a duplicated caching can be avoided.



Figure 3. A simple coordinated in-network caching illustration

Take an example in Fig. 3 for the illustration of a simple coordination scenario with two objects traversing, content object *a* and *b*, respectively. Each content object is divided into five data chunks of a unified size. And the cache size is denoted as the number of chunks, assumed to be a five-chunk capacity. Assuming a case wherein neighbouring caches of one-hop distance coordinate with each other via caching status sharing and requests serving. For individual network elements, either a universal caching or a selective policy can be adopted. As shown in the figure, neighbouring caches are able to hold diverse content chunks due to coordination.

However, network devices beyond one-hop distance can still hold the same chunks, such as cache *A* and *D*, or cache *B* and *C*, since it is beyond the communication distance. If the cache redundancy in a neighbourhood is calculated as . Then with awareness of caching status in the neighbourhood of *A-B-C*, the cache redundancy intensity can be obtained as for content *a*, and for content *b*, and the overall cache redundancy can be deduced with all contents as . As compared with non-coordination caching approaches, wherein the cache redundancy intensity can be up to 100% based on a universal caching policy, while a reduction of 67% at least can be achieved under coordination in this case.

Theoretically, the ideal performance gains could be achieved when neighbouring caches exchange the whole list of cached contents, so as to completely avoid duplicated caching. However, the signalling overhead can be another concern. Thus a trade-off exists between the goal of cooperation (maximising the diversity of cached items) and the performance goal of minimizing the access latency.

### 3.2.3 Cache Replacement Policies

When the cache is fully occupied, a cache replacement policy needs to be adopted to evict cached data chunks. In comparison, cache replacement is orthogonal to caching policy, and thus the two can operate independently without affecting each other’s decisions. Generally, there are two main eviction policies, known as the LFU (Least Frequently Used) and LRU (Least Recently Used). Based on these two eviction policies a combination of a Least Frequently and Recently Used policy is also implemented [33]. A more simplified eviction policy known as the RND (random replacement policy) is considered as an efficient replacement policy especially for ICN in [30], due to the line-speed consideration in ICN caching operations and a RND is easy to implement in practice to make ICN scalable.

LFU is the eviction policy to remove the item in the cache with the smallest number of accesses. However, such eviction policy is hard to be implemented in practice in constant time. In Comparison, LRU is a relatively simpler policy easily implemented in practice as compared to the LFU. It replaces items that are so far have not been accessed in the longest time. LRFU is a combination of LRU and LFU, which leverages between these two eviction policies to promote the most benefits of both. A weight value is associated to each chunk to weigh its preference degree for staying in the cache, wherein a constant parameter allows for a trade-off between recency and frequency. Chunks with the lowest weight value will be evicted. A simple random eviction policy [34], however, is proved to be able to perform fairly well to achieve desirable performance gains.

## 3.3 In-network Caching Efficiency

The caching efficiency is essentially associated with the following performance aspects: *cache hits*, *cache evictions*, and *cache redundancy.*

* **Cache Hits** are important to evaluate a caching policy, which is often expressed as a ratio that refers to the fraction of requests served at a cache for a given amount of time.

Intuitively, a caching-everything-everywhere policy enables high cache hits due to its aggressive caching. However, it can suffer from low caching efficiency due to its data-insensitive caching nature. In comparison, a more strategized caching policy takes into account important context information so as to perform intelligent caching decisions’ making. Typical examples are [3][24][25] based on content popularity and routing information, such as distance-aware caching. As such, popular content items tend to be more frequently served at caches, and latency can be reduced due to demanded data stored at closer caches.

If coordination is further enabled between adjacent caches, cache hits can be further improved due to the possibilities for serving content requests are increased to larger scale. Namely, requests unable to be served locally re forwarded to nearby caches that may store the demanded data [2][32].

* **Cache Evictions** are also important to expresses the efficiency of a caching policy. The operational cost for a cache is mainly attributed to cache evictions. And thus high cache hit ratio along with low number of cache evictions usually defines the efficiency for a caching policy.

With the universal caching, high cache hits do exist based on evaluations of Ref. [3]. Nevertheless, severe cache evictions are incurred at the same time due to its content-blind caching strategy. As compared, a more selective caching can reduce huge cache evictions owing to its intelligent content-aware caching decision. It promotes more popular data items to be cached while less unpopular caches are discarded or evicted.

With coordinated caching policy, cache evictions can be further reduced as compared to the independent selective caching. The rational is the caching knowledge shared via coordination cooperation among adjacent caches. As such, unnecessary evictions can be avoided if the to-be-cached popular data items are already cached nearby.

* **Cache Redundancy** is the metric that can measure the efficiency of the management of network resources. As shown previously, the cache redundancy refers to the intensity of data replicas distributed over the network. From the network’s perspective, a low cache redundancy is expected to more efficiently utilize cache resources.

However, huge cache redundancy can be incurred under a universal caching as dicussed previously as shown in Fig. 2. Although a selective caching can reduce a certain level of the cache redundancy by taking into account some context information (e.g., the locality of the cache, or the popularity degrees of content items), cache redundancy still exists under such caching strategy among network elements due to their independency on caching operations.

In comparison, coordinated caching can greatly reduce the cache redundancy. With awareness of neighbouring caches’ caching states, duplicated caching can be avoided. However, concerning the signalling scalability, an exchange of a full list of content items among adjacent caches can be costly in terms of communication overhead. On the other hand, the coordination scope among caches can be enabled in various scales that may achieve different performance results. However, by enabling neighbouring caches with one-hop distance coordinating with each other, the performance gains are sufficient enough [2][3]. A more strategized consideration promotes selecting caches with the most connection degrees in the network, which are engaged in coordination with their nearby caches [35]. The rational behind this is that caches with high connection degree can have great potential to serve more requests from nearby nodes.

# 4. Incentives and Challenges

Largely driven by the alleviation of cellular traffic load on networks, mobile data offloading can offer improved network capacity, lower cost and desirable user data rate. These benefits provide incentives for customers to remain connected on the cellular network, most of the time. With offload, the potential incentives include:

* Minimal changes to existing network protocols and interface, by avoiding costly deployment, i.e., dependence on expensive real-estate cost
* Assured service quality in a simple way with regard to deployment

However, new challenges are implied mainly in the following aspects:

* **Interference management**: It is one of the challenges for D2D offload, when devices operates in the same band as cellular communications. For example, interference will happen when two adjacent devices share the same band but access to different networks (e.g., one with D2D mode and the other with the cellular network). Due to the coexistence of D2D and conventional cellular, harmful interference can be caused. Two types of interference can be invoked: a) cellular-D2D user interference and b) D2D-D2D user interference. By adopting some power control strategies performed by BS in a centralized way, the cellular-D2D can be relieved. And operations managed by devices in a distributed manner can alleviate the D2D-D2D interference. However, how to design an effective and efficient strategy has always been a major challenge [16]-[18].
* **Mobility management**: Since mobile data is offloaded from mobile networks that naturally supports seamless handovers for users in mobility, it is important to maintain service continuity with offload cases. However, the challenge is how to efficiently handle the mobility management in a heterogeneous access network scenario, i.e., in case of multiple offload options available (e.g., WiFi/small cell, BS, D2D offload). Some strategized offload shift policies need to be designed to make efficient decisions, such as to identify the right time to initiate the shift and to trigger optimal selections on the most appropriate interface. Meanwhile, due to D2D natures, high latency and signaling overhead are some of the major issues [16], especially with energy concerns that devices can be out of battery quickly when served as relay. With account for offload efficiency, desirable service continuity becomes challenging for users in mobility.
* **Security and privacy issues**: As in D2D communication, since the user data is transmitted through other users’ devices, the security and privacy become crucial which must be maintained. The BS can work as a trusted party that is responsible for authentication of the relaying devices and proper encryption on data, so as to maintain the privacy of the users. Without intervention of BS, devices have to establish trust based on local knowledge that can be vulnerable to malicious attacks, which is an open research problem in this case [19].

By taking advantage of network-assistance, concerns discussed above could be greatly relieved, e.g., by adopting the key generation and distribution mechanisms already available in mobile networks. However, the centralized fashion with the network-control incur concerns with issues inherently with such operation, such as communication overhead and network scalability.

Key techniques of in-network caching are discussed previously, in the hope of providing hints on the design of efficient offload paradigm. However, remaining challenges still need to be addressed in order to design an efficient in-network caching policy in a cellular network.

* **Trade-off** between users’ perceived service quality and the network efficiency.

From a consumer’s perspective, it is ideal that every content demand can be served. While from the network’s perspective, a desirable caching policy should be more selective rather than aggressive, so as to efficiently distribute diverse contents among caches with limited capacities. As such, more popular content items tend to be cached and served by intermediate nodes while less popular content items are discarded and may only be reached from original sources. With content items suffering from availabilities at original sources, access latency could be increased for end users. Hence, the trade-off between users and the underlying network is one concern worth noticing.

* **Communication overhead** is one major concern in such coordination deployment.

In order to keep caches to be updated, network elements need to periodically exchange information on their locally cached data items, in order to avoid unnecessarily duplicated caching between nearby caches. As a result, potentially huge information will be incurred for exchange. Bloom Filters (BFs) [36] are one technique effective for aggregating information on cached data items in signalling between coordinated caches in order to achieve scalability.

* **Cooperation scope** is another research issue concerning how far in the neighbourhoodcaches should search.

Most researches keep cooperation at one-hop scale. The main purpose is to ensure the access latency at a demanded level with an accepted cache redundancy. Thus a *trade-off* exists between the goal of cooperation (maximising the diversity of cached items) and the performance goal of minimizing the access latency. An optimal degree of redundancy is also seen as a challenging problem [35]. To avoid such a dilemma, the scope of cooperation is usually kept at one-hop scale, which frees researchers from routing concerns and also benefits the reduction of signalling overhead.

* **Legal issues**: The content items cached may incur copy right issues for some content providers, if the economic benefits are on the main concern of illegally wide distribution of the contents. Such issue needs to be taken into account when making caching decisions, to make sure the cached data has already acquired the certificate from the content provider. Nevertheless, such certificate needs the cooperation with the content owners, and the strategy regarding the distribution of the economic utilities between the network operator and the content provider is still unknown.

# 5. Summary and Conclusions

Offloading along with in-network caching is among the most popular solutions for enabling efficient multimedia content delivery services in upcoming 5G environments. Researchers as well as service providers have commonly agreed that the two techniques enable great performance gains by enabling services at the network edge, with regard to enhanced pervasive connectivity as well as service velocity. Given that mobile data will substantially dominate in future content distribution applications, such a feature will substantially increase the network efficiency when more end users become actually mobile users with smart user devices.

There is no doubt that mobile data offloading is a major opportunity for mobile network operators, which enables feasible data distributions with cost-effective solutions. By pushing the communication mode towards the network edge, it has the potential to achieve efficient network operation for the concerns of future cellular services. Meanwhile, offload brings forward many challenging issues that remain as open research problems, which requires further research standardization efforts.

By promoting ICN-based caching technologies at the core or edge network in 5G-based network environments, the incentives of mobile edge offloading could be enhanced greatly. In-network caching can be supported by sophisticated caching logics at the network-element or user device-level. Several standardization activities also advocate efficient in-network caching strategy. The Internet Engineering Task Force (IETF) have proposed two options to cache content chunks within the network, either implicitly or explicitly. With implicit caching, network nodes passively cache passing-through content chunks, while the explicit policy enables caches to explicitly request for content chunks to be cached locally. Most research works are based on implicit caching concept, while the explicit caching is not gaining much research interests due to its complexity for deployment in reality.

# References

[1] Cao, Yue, et al. "An electric vehicle charging management scheme based on publish/subscribe communication framework." *IEEE Systems Journal* (2015).

[2] Zhang, Xu, and Yue Cao. "A cooperation-driven ICN-based caching scheme for mobile content chunk delivery at RAN." *Wireless Communications and Mobile Computing Conference (IWCMC), 13th International. IEEE*, 2017.

[3] Xu Zhang, Ning Wang, et al., “A Distributed In-Network Caching Scheme for P2P-like Content Chunk Delivery,” *Computer Networks, Elsevier* (2015)

[4] Cisco Visual Networking Index, Global Mobile Data Traffic Forecast Update, 2016-2021 White Paper Mar. 2017.

[5] T. Koponen et al., “A Data-Oriented (and beyond) Network Architecture,” SIGCOMM, vol. 37, no. 4, pp. 181-192, 2007.

[6] Van Jacobson, etc., “Networking Named Content”, Proc. of the 5th international conference on Emerging networking experiments and technologies, ACM CoNEXT’09, 2009.

[7] J. G. Andrews, H. Claussen, M. Dohler, S. Rangan, and M. C. Reed, “Femtocells: Past, present, and future,” *IEEE Journal on Selected Areas* *in Communications*, vol. 30, no. 3, 2012.

[8] Suh, Dongeun, Haneul Ko, and Sangheon Pack. "Efficiency analysis of WiFi offloading techniques." *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, 2016.

[9] Zeydan, Engin, et al. "Integration and management of Wi-Fi offloading in service provider infrastructures." *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*, 2016

[10] Suh, Dongeun, Haneul Ko, and Sangheon Pack. "Efficiency analysis of WiFi offloading techniques." *IEEE Transactions on Vehicular Technology*, vol. 65, no. 5, 2016.

[11] Zeydan, Engin, et al. "Integration and management of Wi-Fi offloading in service provider infrastructures." *Network Operations and Management Symposium (NOMS), 2016 IEEE/IFIP*, 2016

[12] White paper from Aptilo Networks: Seamless Wi-Fi Offload: A Huge Business Opportunity for Mobile Carriers, 2013

[13] N. Golrezaei et al., “FemtoCaching: Wireless video content delivery through distributed caching helpers,” in *Proc. of IEEE INFOCOM*,2014.

[14] S. Rangan and R. Madan, “Belief propagation methods for intercell interface coordination in femtocell networks,” *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 631-640, Mar., 2012

[15] Andreev Sergey et al.,”Analyzing assisted offloading of cellular user sessions onto D2D links in unlicensed bands,” *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 1, 2015

[16] Lin Xingqin et al., “An overview of 3GPP device-to-device proximity services, ” *IEEE Communications Magazine*, vol. 52, no. 4, 2014

[17] G. Fodor et al., “Design aspects of network assisted device-to-device communications,” *IEEE Communications Magazine*, vol. 50, no. 3, 2012.

[18] Han, Bo, et al. "Mobile data offloading through opportunistic communications and social participation." *IEEE Transactions on Mobile Computing,* vol. 11, no. 5, 2012

[19] M. N. Tehrani, M. Uysal, and H. Yanikomeroglu, “Device-to-Device Communication in 5G Cellular Networks: Challenges, Solutions, and Future Directions” IEEE Communications Magazine, vol. 52, no. 5, pp. 86–92, 2014.

[20] H. Song, et al., “DECoupled Application Data Enroute (DECADE) Problem Statement,” IETF RFC 6646, 2012.

[21] R. Alimi, et al., “A Survey of In-Network Storage Systems,” IETF RFC 6392, 2011.

[22] Ioannis Psaras, Wei Koong Chai and George Pavlou, “Probabilistic In-Network Caching for Information-Centric Networks,” in ACM SIGCOMM 2012

[23] Shuo Guo, Haiyong Xie and Guanyu Shi, “Collaborative Forwarding and Caching in Content Centric Networks,” In Proc. Networking, 2012.

[24] Kideok Cho, Munyoung Lee and et al., “WAVE: Popularity-based and Collaborative In-network Caching for Content-Oriented Networks,” in Proc. of IEEE INFOCOM Workshop on Emerging Design Choices in Name-Oriented Networking, 2012.

[25] Zhongxing Ming, Mingwei Xu and Dan Wang, “Age-based Cooperative Caching in Information-Centric Networks,” in Proc. of IEEE INFOCOM Workshop on Emerging Design Choices in Name-Oriented Networking, 2012.

[26] M. Ain et al., “D2.3-Architecture Definition, Component Descriptions, and Requirements,” Deliverable, PSIRP 7th FP EU-funded project, Feb. 2009.

[27] B. Ahlgren et al., “Second NetInf Architecture Description,” 4WARD EU FP7 Project, Deliverable D-6.2 v2.0, Apr. 2010, FP7-ICT-2007-1-216041-4WARD/6.2, <http://www.4ward-project.eu/>

[28] E. J. Rosensweig and J. Kurose, “Breadcrumbs: efficient, best-effort content location in cache networks,” in INFOCOM, 2009

[29] Somaya Arianfar Pekka Nicander and Jӧrg Ott, “On Content-Centric Router Design and Implications,” in ACM ReArch 2010

[30] S. Arianfar and P. Nikander, “Packet-level Caching for Information-centric Networking,” in ACM SIGCOMM, ReArch Workshop, 2010

[31] Konstantinos Katsaros, George Xylomenos and George C. Polyzos, “MultiCache: An overlay architecture for information-centric networking,” in Proc. of Computer Networks, vol. 55, no. 4, 2011.

[32] Kideok Cho, Munyoung Lee and et al., “WAVE: Popularity-based and Collaborative In-network Caching for Content-Oriented Networks,” in Proc. of IEEE INFOCOM Workshop on Emerging Design Choices in Name-Oriented Networking, 2012

[33] D. Lee, J. Choi, J.-H. Kim and et al., “LRFU: A Spectrum of Policies that Subsumes the Least Recently Used and Least Frequently Used Policies,” *IEEE Tran. on Computers*, vol. 50, no. 12, 2001

[34] Dario Rossi and Giuseppe Rossini, “A Dive into the Caching Performance of Content Centric Networking,” IEEE International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD), 2012

[35] Wang, J. M., Zhang, J., & Bensaou, B. (2013, August). Intra-AS cooperative caching for content-centric networks. In *Proceedings of the 3rd ACM SIGCOMM workshop on Information-centric networking*. ACM.

[36] H. Liu, et al, “*A Multi-Level DHT Routing Framework with Aggregation”*, in Proc. ACM SIGCOMM ICN’12, Helsinki, Finland, August 2012