Received January 6, 2017, accepted February 15, 2017, date of publication March 30, 2017, date of current version May 17, 2017.

Digital Object Identifier 10.1109/ACCESS.2017.2689678

CaaS: Caching as a Service for 5G Networks

XIUHUA LI[1], (Student Member, IEEE), XIAOFEI WANG[2], (Member, IEEE), KEQIU LI[2], (Senior Member, IEEE), AND VICTOR C. M. LEUNG[1], (Fellow, IEEE)

[1]Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver BC V6T 1Z4, Canada
[2]Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, Tianjin 300072, China

Corresponding author: XIUHUA LI (lixiuhua@ece.ubc.ca)

This work was supported in part by the National Natural Science Foundation of China under Grant 61671088 and Grant 61271182, in part by the China Scholarship Council Four Year Doctoral Fellowship, and in part by the Canadian Natural Sciences and Engineering Research Council. This paper was presented at the IEEE International Conference on Computer Communications Workshop on Wireless Mobile Virtualization, Hong Kong, 2015.

ABSTRACT In recent years, demands for rich multimedia services over mobile networks have been soaring at a tremendous pace. Traditional dedicated networking equipment may not be able to efficiently support the phenomenal growth of the traffic load and user demand dynamics while consuming an unnecessarily large amount of energy resources. Recently, mobile content caching, whereby popular contents are cached inside the mobile front-haul and back-haul networks so that demands for these contents from users in proximity can be easily accommodated without redundant transmissions from the remote sources, has emerged as an efficient technique for multimedia content delivery. Mobile content caching is particularly suitable for fifth generation (5G) mobile systems that are being designed to incorporate advanced cloud computing technologies and network function virtualization techniques. Therefore, in this paper, we first propose the concept of "Caching-as-a-Service" (CaaS) based on cloud-based radio access networks, and virtualized evolved packet core, which provides the capability to cache anything at anytime, anywhere in the cloud-based 5G mobile systems to satisfy user demands from any service location with high elasticity and adaptivity, and to empower third-party service providers with flexible controllability and programmability. Then, we study the potential techniques related to the virtualization of caching, and discuss the technical details of virtualization and optimization of CaaS in 5G mobile networks. Some novel schemes for CaaS are proposed to target different mobile applications and services. We also explore new opportunities and challenges for further research.

INDEX TERMS 5G, caching as a service (CaaS), virtualization, content delivery network.

I. INTRODUCTION

Along with recent advances in mobile communication technologies, an ever-growing number of mobile users are attracted to enjoy a wide plethora of multimedia services using smart phones and tablets [1]. However, the capacities of wireless links, radio access networks (front-haul), and core networks (back-haul) are not keeping pace with the explosively growing mobile traffic due to the centralized nature of mobile network architectures [2]–[6]. Indeed, despite the continuous efforts of Mobile Network Operators (MNOs) and network equipment vendors to enhance the wireless link bandwidth by adopting sophisticated radio access techniques for the fourth generation (4G) mobile networks, cellular network deployments that focus on improving the efficiency of network resource utilization may not adequately handle user demand dynamics [7], [8]. Specifically, the cellular network architecture currently in use cannot efficiently handle the massive delivery of contents that are repeatedly requested by multiple users, as each request is processed as a different end-to-end (E2E) connection. The exploding traffic demands thus make it difficult to achieve good E2E service performances [9], [10].

To meet the above challenges, new approaches in mobile network architectures and advanced data transmission technologies are emerging in the development of next generation, i.e., fifth generation (5G) mobile networks. As conventional approaches of improving spectral efficiency within the allocated spectrum are fast approaching their theoretical limits, 5G network system designs are incorporating new application-aware approaches to further reduce traffic demand and improve the utilization of network resources. One key approach is to cache popular multimedia contents at the edges of mobile networks to reduce the E2E traffic due to redundant downloads and thus to optimize the E2E service delays [9], [10].

By caching popular contents inside mobile front-haul and back-haul networks,[1] due to the "Power Law" effect of content popularity, [9], service demands of mobile users in proximity for these contents can be easily accommodated without redundant transmissions from remote sources outside the mobile networks, so that additional traffic load from duplicated transmissions can be reduced significantly. However, the architectures of legacy mobile networks, which incorporate dedicated signal processing hardware in the RANs and sophisticated controlling and processing units in the EPCs, have not been designed with in-network caching in mind. Thus they are less capable of adapting to users' dynamic demands will maximizing the efficiency of resource utilization.

Recently, the new trend of Network Function Virtualization (NFV) [12] through Software-Defined Networking (SDN) [11] and cloud services is starting to be embraced by researchers and practitioners in the mobile networking community as an attractive solution to optimize network resource utilization and thus reduce MNOs' expenditures. By decoupling the software defined control plane from the underlying hardware-driven data plane, networking functions can be flexibly supported over commodity hardware platforms that emulate network devices under software control, as if they are running on their own bare-metal computing resources. Through virtualization, MNOs can dynamically adjust network functions within virtual machines (VMs) in an online manner to provide quick and elastic services on-demand to mobile users, 3rd-party service providers (SPs), and content providers (CPs).[2] One of the outcomes of this trend is the emergence of the new concepts of "Radio-Access-as-a-Service (RAaaS)" and "EPC-as-a-Service (EPCaaS)". RAaaS enables highly flexible RANs supporting multiple radio access technologies (multi-RATs) by connecting general purpose Remote Radio Units (RRUs) via high-speed fronthaul links (e.g., fiber optics) to data centers, in which VMs are employed as virtualized Baseband Processing Units (BBUs) to perform signal processing functions specific to different RATs [13]. EPCaaS virtualizes the EPC functions into VMs in the cloud to enable them to be managed elastically [12].

As another novel outcome of the trend in virtualization of mobile networks, this paper introduces "Caching-as-a-Service (CaaS)" that we originally proposed in [5]. CaaS deploys virtualized caching inside a MNO's data centers, and offers researchers a new direction with strong potential for innovations. Going beyond running traditional Content Delivery Network (CDN) services virtually, which still involves static storage of files and assignment of management units in VMs of the cloud, e.g., ActiveCDN [14], CaaS incorporates an innovative framework to enable caching functions and maintain cache resources universally inside the virtual environment, where CaaS instances can be adaptively created, immigrated, scaled (up or down), shared, and released depending on the user demands and requirements from 3rd-party SPs. Caching VMs can even be executed in the same server of RAN VMs to reduce E2E delays of serving mobile users handled by the same RAN VMs. The flexible framework of CaaS brings new challenges of designing virtualization strategies for caching resource, optimizing the resource utilization and performance for all entities, and exploiting new potential scenarios of realizing CaaS. In addition, CaaS enables MNOs to provide highly flexible and programmable virtual caching capabilities to 3rd-party SPs, allowing them to serve mobile users with highly qualified and customized services, while ensuring that the needs to maximize the utility of network facilities and resources through necessary traffic optimization, task scheduling and load balancing techniques inside the data centers are satisfied in a manner that is transparent to mobile users and 3rd-party SPs.

The main objective of this paper is to identify and discuss the key challenges of developing and deploying CaaS in the emerging 5G networks. The main contributions of this paper are summarized as follow:

- This paper introduces the concept of CaaS based on C-RANs and the virtualization of EPC, aiming at caching anything, anytime, anywhere in the cloud-based 5G mobile systems with high elasticity and adaptivity to user demands and service locations, and with flexible controllability and programmability to 3rd-party SPs.
- This paper explores potential techniques for virtualization of caching, and discusses technical details of virtualization of caching and optimization for CaaS in 5G mobile networks.
- This paper proposes some novel schemes utilizing CaaS to offload network traffic and improve user Quality of Service (QoS) in various mobile applications and services, and also explores new research challenges and opportunities.

The rest of this paper is organized as follows. We first study the techniques related to the virtualization of mobile networks in Sec. II, and then discuss technical details of virtual on-demand caching for 5G mobile networks in Sec. III. Then some novel virtualization of caching schemes are proposed with respect to some promising mobile services in Sec. IV. Performance metrics for evaluation of CaaS are presented in Sec. V. We explore some new research challenges and opportunities in Sec. VII, and conclude the paper in Sec. VIII.

II. CaaS BASED ON RAaaS AND EPCaaS IN 5G NETWORKS

A. RAaaS AND EPCaaS VIRTUALIZATION OF RAN AND EPC

An important trend in the development and deployment of 5G networks is the virtualization of traditional radio access processing functions in the cloud, so called C-RAN [13],

[1] In this paper, we use interchangeably the terms "front-haul network" and "Radio Access Network (RAN)", and "back-haul network" and "Evolved Packet Core (EPC) network".

[2] In this paper, for simplicity we use *SPs* to represent all kinds of 3rd-party service providers including content providers (CPs).

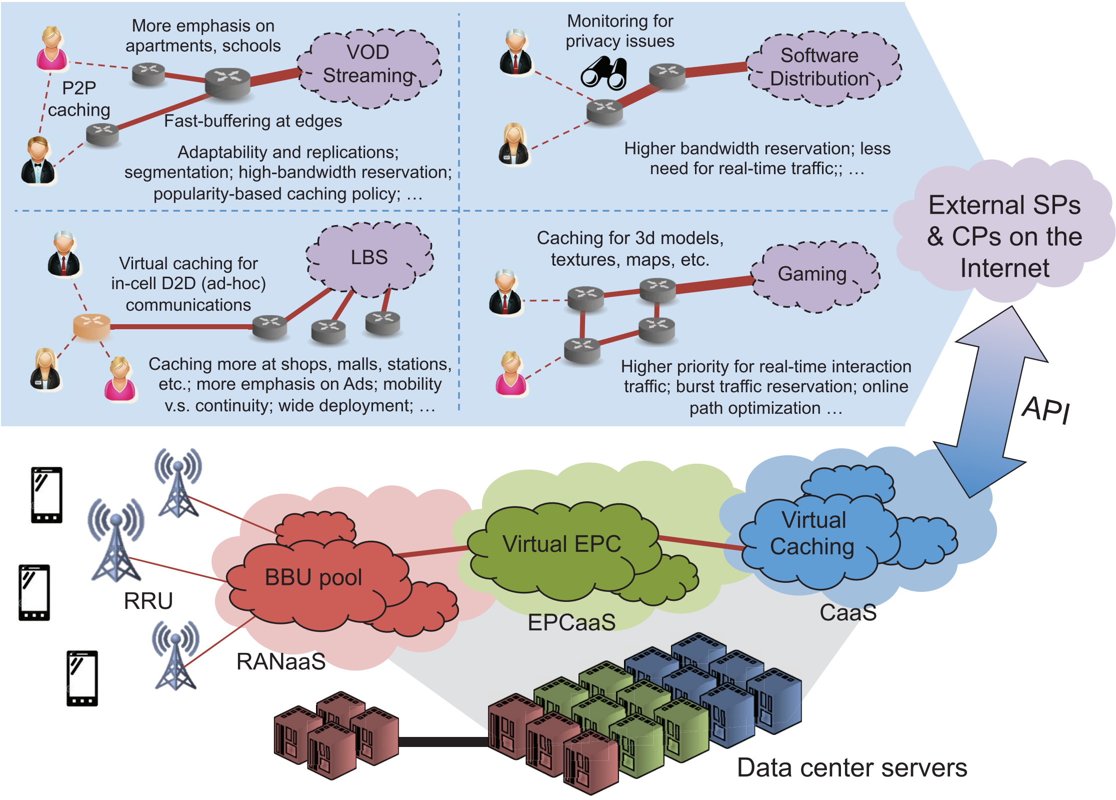

More emphasis on apartments, schools
VOD Streaming
P2P caching
Fast-buffering at edges
Adaptability and replications; segmentation; high-bandwidth reservation; popularity-based caching policy; …
Monitoring for privacy issues
Software Distribution
Higher bandwidth reservation; less need for real-time traffic;; …
Virtual caching for in-cell D2D (ad-hoc) communications
LBS
Caching more at shops, malls, stations, etc.; more emphasis on Ads; mobility v.s. continuity; wide deployment; …
Caching for 3d models, textures, maps, etc.
Gaming
Higher priority for real-time interaction traffic; burst traffic reservation; online path optimization …
External SPs & CPs on the Internet
API
RRU
BBU pool
RANaaS
Virtual EPC
EPCaaS
Virtual Caching
CaaS
Data center servers

FIGURE 1. Illustration of "Caching-as-a-Service".

which forms the basis of RAaaS. C-RANs adopt virtualized BBUs located at MNOs' data centers, which function as virtual base stations (BSs) that can flexibly support multi-RAT and are easily expandable in processing capacity. The BBUs are connected to RRUs and their remote antennas by high-speed front-haul fiber optic networks. Moreover, many MNOs are also embarking on the virtualization of the core networking functions, i.e., EPC, in computing clouds using commodity hardware but with full feature compatibility of purpose-built EPCs [12], which enables EPCaaS. MNOs can further bundle multiple functions of virtualized RAN and EPC in a single VM or a group of adjacent VMs in order to reduce the network control traffic significantly, i.e., RaaS and EPCaaS can be supported by the same data center. RAaaS and EPCaaS offer the advantages of unprecedented flexibility and scalability. For instance, massive cooperation of multiple transmit and receive points (antennas) with optimal pooling/scheduling of radio resources in different dimensions (frequency/time/space/coding/power) and effective intra- and inter-cell interference control can be realized through software control. With the virtualization of RAN and EPC into RaaS and EPCaaS using VMs in data centers, it is natural to consider the incorporation of caching functionality into VMs in the same data centers integrated with virtual RAN and EPC in order to improve the performance of multimedia content delivery services.

B. CaaS—VIRTUALIZATION OF CACHING

Integration of multimedia content caches into the VMs for elastic CaaS brings the promises of performance improvements through flexible adaptation to user demands and 3rd-party SPs' requirements, but also lead to many new challenges in the design and deployment of optimal resource allocation schemes. Fig. 1 illustrates the different dimensions of CaaS. We first present the three conceptual layers of the CaaS framework along with the RAaaS and EPCaaS in the emerging 5G networks in the following.

- ***Layer 1:*** The physical servers in an MNO's data center are running the VMs that support RAN and EPC as well as caching functions. Data centers can be centralized with servers clustered in the same location, or distributed between different locations that are interconnected with high-capacity fiber optic cable. Existing traffic optimization and task migration scheduling techniques for data centers can be deployed in this layer, where practical performance metrics are measured and reported to upper layers.
- ***Layer 2:*** With caching coexisting with RAN and EPC, where everything is virtualized in some VMs in a MNO's data centers, the caching paradigm can be considered as the universal caching discussed in [10]. For example, caching a content at a BS may involve updating the path for the cache file from the caching

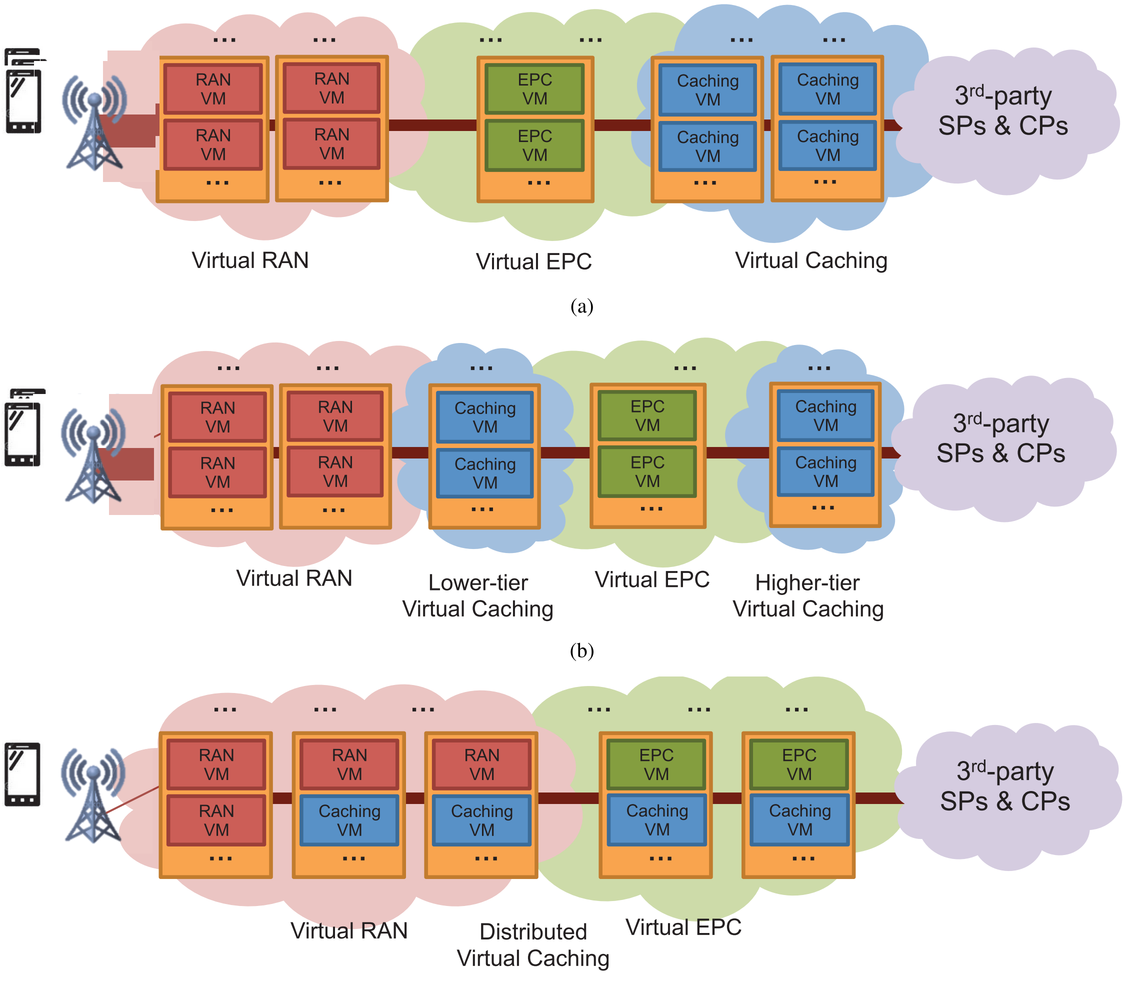


FIGURE 2. **Virtual caching schemes. (a) Traditional CDN & Virtual CDN. (b) Hierarchical virtual caching. (c) Distributed (Universal) caching.**

VMs, i.e., cVMs, to the RAN VMs by prioritized routing, or directly migration of a whole cVM into the server supporting the RAM VMs. If routing priority is not supported, the transmissions of cached contents will suffer from competition with other traffic, and the QoS for content delivery may be degraded or not guaranteed. In this layer, contents can be chunked, replicated, distributed, bundled, and redirected freely among cVMs, based on the traffic dynamics, content popularity and the diversity of user demands. Also, information and functions of other layers (e.g., D2D communication, signal quality of user devices, and mobility management) can be utilized by MNO or 3rd-party SPs for online cross-layer optimization to improve service delivery.

- ***Layer 3:*** CaaS supports application programming interfaces (APIs) that can be employed by 3rd-party SPs to programme the desired virtual caching functions, e.g., deploying static offline caching policies or invoking dynamic online caching strategies. SPs can manage their virtual caches with regard to the network topology for optimizing their own traffic and enabling QoS guarantee to their own customers. Also, MNOs can dynamically charge for the resource utilization of the SPs.

III. VIRTUAL CACHING SCHEMES FOR 5G NETWORKS

CaaS with virtual caching brings more flexible and elastic content caching and delivery services than caching only at the edge, as the former can be carried out at any location that is appropriate inside a MNO's network. In-network caching needs to be positioned and provisioned at a proper location inside the MNOs' virtualized environment, with an appropriate distance between mobile users at the edge and the involved RAN and EPC functions, while satisfying the goal of global optimality of resource utilization. Depending on the placement of the caching functionality, as illustrated in Fig. 2, we present three schemes of cache virtualization as follows: 1) virtual CDN, 2) hierarchical caching, and 3) distributed caching. For each of these schemes, we also briefly discuss the methodology to achieve the potential benefits by caching optimization and forwarding path adjustment in CaaS.

A. VIRTUAL CDN IN AN MNO's CLOUD

Fig. 2(a) shows the scenario in which caching, i.e., CDN service, is enabled on VMs in dedicated servers between the servers running virtual EPCs in the MNO's data center and the gateway to the Internet. The MNO and 3rd-party SPs can easily create or modify the virtual CDN instances corresponding to the user requirements and network conditions. Also this scheme can enjoy simple and smooth transition from current CDN services to virtual caching from the engineering perspective. Applying the principles of virtualization to CDN gives 3rd-party SPs greater control over their content distribution and traffic flows by designing optimal online caching policies for a better performance with limited budget. CDN services in VMs can also enjoy easy mobility management and automatic flow path adjustment, which reduces the time needed to install or expand the CDN for 3rd-party SPs. Virtualization also gives MNOs the agility to instantly create a new CDN cache for the purposes of minimizing any potential disruption in service in case of a hardware failure, or to elastically serve the peak traffic with ease instantly. Note that there have been a few studies on virtual CDN, e.g., ActiveCDN [14], and also Alcatel-Lucent has products that brings programmability and automation features to CDN solutions. However, in this scheme, 3rd-party SPs still cannot fully control and utilize the caching, cVMs are not very close to mobile users at the edges, and "intra-MNO traffic", caused by duplicated downloads from cVMs to mobile users via back-haul and front-haul, are not reduced.

B. HIERARCHICAL VIRTUAL CACHING

Recently, there have been some preliminary studies on caching contents at mobile edges, i.e., BSs or eNodeBs, e.g., video caching in BSs [15] and FemtoCaching [16], all of which are similar to the "RAN caching" concept presented in [10] for achieving better E2E performance. However, these approaches are challenged by the difficulty in designing large-scale cooperative caching policy as well as the high implementation complexity.

Virtualization of mobile networks and systems brings potential solutions to the aforementioned issues. As shown in Fig. 2(b), we can hierarchically deploy cVMs into servers performing the different functions as follows: servers running RAN VMs and servers running EPC VMs, possibly at different locations. This hierarchical caching architecture can potentially enhance the caching performance greatly. Caching close to the BSs can help to reduce traffic congestion and improve E2E delay, but it might suffer from a high cache-miss ratio because of limited caching resources at the edges of the mobile network. Caching near the EPC can enjoy a high cache hitting ratio due to the abundant caching resources, but the longer routing path may induce large latencies and increase traffic congestion. Thus, cVMs should be dynamically allocated to the two levels inside the MNO's data centers adapting to the dynamics of user demands and mobility, and the MNO's network configuration and traffic distribution.

The specialized caching controller module should be abstracted from traditional CDN services for caching resource management and optimization, and forwarding cached contents among cVMs. The CaaS controller as mentioned before is also an instance for virtual caching management, and thus can be scaled up and down as well as duplicated or immigrated freely depending on different working conditions.

C. DISTRIBUTED VIRTUAL CACHING IN MNO's CLOUD

The final stage of evolution of virtual caching architectures is to freely deploy cVMs in any servers of the MNO's data centers, according to the requirements from mobile users, the QoS guarantee contracted for 3rd-party SPs, the constraints from the physical facilities of the network and data centers, and their current workloads. Instances of cVMs can be attached to any servers or migrated between servers in order to achieve global optimal scheduling. The caching here is quite universal and can take place anywhere due to the possible collocation of cVMs with RAN VM and EPC VMs, since the BBU pools, EPC core networks and caching servers are within the same cloud infrastructure operated by the MNO, i.e., they are either within the same data center or distributed in difference data centers that are connected by a high-speed data network and managed by the MNO.

Furthermore, any content in cVMs can be chunked (divided) into multiple pieces and pre-packaged for efficient transmissions via the network layer, so that contents may be delivered to users expeditiously via servers running RAN VMs without flowing back to EPC VMs. If the content is cached within the cVMs running in the same servers as the RAN VMs, the latency can be kept extremely short. The centralized caching controller (possibly based on the SDN controller) plays a more important role in this scenario due to the needs to cache and share user session data, and to chunk, bundle and re-direct virtual resource and contents. Distributed caching brings more sophisticated problems for VM placement and traffic control compared to existing studies on data center network optimization in the literature. Efficient routing need to be configured automatically by MNOs according to mobile user demands and 3rd-party SPs' requirements, and effective "pre-linking" of cached contents for the routing path towards mobile users is important to reserve bandwidth and transmission opportunities, which is analogous to "pre-fetching" of the contents.

D. OPTIMIZATION OF CACHING AND FORWARDING PATH ADJUSTMENT IN CaaS

To achieve the benefits of the above three schemes, optimization of caching plays an essential role in CaaS. The efficiency of consolidating virtual resources for RAN, EPC and caching in the CaaS framework has to be balanced against the possibility of congestions and hence task deadline violations. The typical practices of cloud providers such as over-subscription of resource (e.g., processors and network bandwidth) amplify this problem even further. Limited capacity of network links

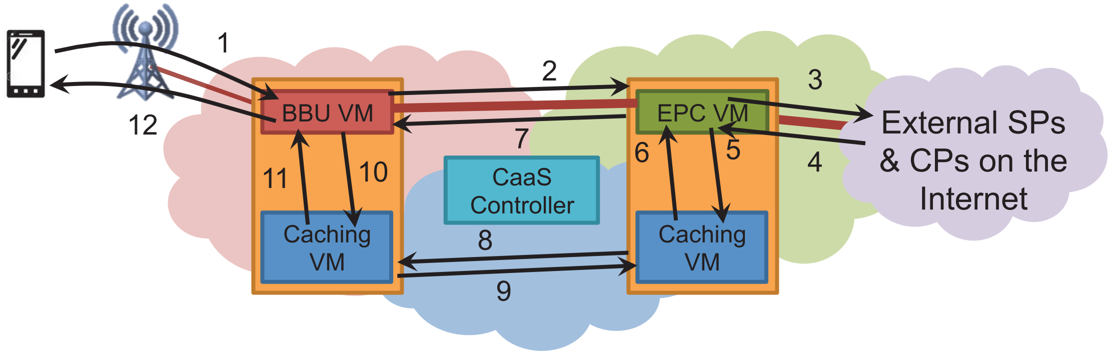


FIGURE 3. Traffic adjustment in "Caching-as-a-Service".

also leads to increased delays especially in the presence of bursty traffic. The key for MNOs to optimize virtual caching is to track the popularity of contents, optimally decide on the cache allocation in a real-time manner, and update the routing table entries interconnecting the cVMs with the other network elements. Therefore, optimal cache allocation and scheduling algorithms with dynamic updates in different time scales are needed.

A further step for optimizing virtual caching among VMs is the adjustment of cache forwarding paths in a real-time manner as illustrated in Fig. 3. Suppose a requested content is not cached. The request going to the source via path "1-2-3" results in the content being sent to the user via path "4-7-12" from the source. If virtual caching of this content occurs in a cVM near the EPC VM, the paths for the request and the returned content respectively become "1-2-5" and "6-7-12". By distributed caching with flexible cVM migration and online path optimization, based on the monitored changes of content popularity and QoS requirements from users and SPs, CaaS can directly promote the cached content or even the whole cVM to a server near the RAN VMs, e.g., via link "8", and future requests for the same content will be sent to the user via path "11-12", which can reduce latency and is more efficient.

IV. APPLICATION SCENARIOS

In the near future, the demands for more sophisticated and affordable mobile broadband services and energy-efficient service provisioning will drive the adoption of virtualization with integrated RAaaS, EPCaaS and CaaS support in 5G networks. To adopt this strategy effectively, MNOs will need to address the important technical issues of how to utilize the virtualized resources in the data centers efficiently while provisioning the required QoS reliably. Due to the dynamic nature of user traffic and network conditions, it is desirable that solutions to optimize resource allocation be amenable for online operations that effectively enable network automation. Since various applications have different requirements, we consider the integration of CaaS in 5G networks under two important application scenarios in this section.

A. PREFETCHING AND CACHING FOR MOBILE VIDEO STREAMING SERVICES

Video streaming services have been the dominating contributor to mobile traffic load for many years, but they have been confronted with highly skewed disparity of view popularity and thus suffered from redundant traffic load all the time. By utilizing CaaS for distributed caching in mobile networks, video streaming SPs can easily reserve high-bandwidth links within mobile networks in a pay-by-demand manner, where bandwidth have different costs in different time periods, while MNOs can charge for bandwidth depending on the user demands and network conditions. Also, because SPs have detailed information of users' preference on Video contents, popularity-based caching policy can be realized with ease over the CaaS framework, and SPs can even prefetch video clips prior to users' clicks.

Particularly, buffering of videos is the most critical factor for QoS and Quality of Experience (QoE) of users. A 1% increase in buffering leads to an average decrease of 3 minutes in user engagement as studied in [17] and [18], and only 54% of viewers experiencing a single start-up failure return to re-buffer the same content, as reported in [17] and [19]. With CaaS, SPs can ease the problem by splitting the first few seconds of the videos and caching them as close to the BSs as possible with the necessary amount of replicas, while maintaining active routing paths for the remaining video parts. Also based on cross-layer information such as signal quality and device battery condition, adaptive streaming of layered videos can be realized by distributing video layers with different qualities at different locations in the cloud so that QoE can be optimized based on the current system conditions. Furthermore, CaaS can optimize caching of video clips for multi-screen services to mobile users; depending on user locations and device characteristics, streaming sessions can be migrated between the screens of different devices, e.g., from a smart phone connected to a BS to a TV set connected to a WLAN, with negligible delay.

TABLE 1. Performance evaluation metrics.

Online Performance Evaluation Metrics (real-time factors)	
Response time	user perceived caching and content delivery response time
Throughput	throughput measured on the path from the caching VM server to the user
CPU Load	the average load of the caching VM server aggregated over a pre-specified time window
Cache hit ratio	the ratio of the number of cached content versus total content requested
Client requests per VM	the average number of the connected clients per VM
Average distance per VM	the average distance of the connected clients from the VM
SINR	signal to interference-plus-noise ratio for measuring the wireless channel quality of mobile users
Offline Performance Evaluation Metrics (physical factors)	
CDN deployment cost	the overall cost of deploying a traditional CDN service, based on cost models (retrieval, update, storage).
Number of VM server	the number of nodes selected by the placement algorithm to host a VM server
Link capacity	The inherent transmission capacity of links among VM servers
Link latency	fixed link delay of the connectivities among VM servers
Path length	the average number of hops between CDN end-users and the origin server in a particular CDN solution.
Shortest path betweenness centrality	for a particular virtual caching solution, the average of individual caching and transmitting nodes of the selected VMs comprising the final overall virtual caching solution.

B. RESOURCE CACHING FOR MOBILE GAMING

The mobile gaming industry is becoming one of the most profitable in the current entertainment market [20]. Gaming contents have special system requirements, e.g., real-time interactivity, security and privacy issues, and "must-download" game clients (including 3D models, textures, maps etc.), which are not of concern to the MNOs, but it would be up to the mobile gaming SPs to provision dedicated servers and high bandwidth network connections to ensure an acceptable QoE for game players. With CaaS, game SPs can pay for resources with higher priority for real-time interactive gaming traffic. CaaS can bring mobile gaming services to the next level as it enables mobile gaming SPs to have highly customizable traffic control and flexible on-demand bandwidth reservation for more a satisfying QoE to mobile game players. More notably, mobile game application stores (markets) can cooperate with MNOs by using CaaS to cache game clients at strategic locations, e.g., university dormitories, apartments and cafeterias, where there may be a high demand for gaming. All components (including 3D models, maps, textures, inter-scene videos and so on) with potentially frequent downloads, especially when groups of users are playing the same game, can be cached for efficient reuse. CaaS can even support the "click-to-play" concept, where users do not need to download entire game clients but just an initial small part and start playing the game. Game components further required as the game progresses can be preloaded into the optimized caches in CaaS to properly trade off cost and performance.

V. PERFORMANCE EVALUATION METRICS FOR CaaS

For CaaS realization, in addition to exploiting the scheme design, how to evaluate the performance is also one of the most critical issues. CaaS should offer high flexibility, programmability, and optimality for mobile users, 3rd-party SPs and MNOs, and hence performance evaluations can be complicated. We summarize how CaaS should be evaluated according to various factors as shown in Table. 1. Here, the considered factors are divided into two categories, i.e., real-time factors for online performance evaluations and physical factors for offline performance evaluations. Real-time factors include response time, delivery throughput, CPU load, cache hit ratio, client requests per VM, average distance per VM, and user SINR. Physical factors consist of CDN deployment cost, number of VM server, link capacity, link latency, path length, and shortest path betweenness centrality.

Moreover, based on the metrics in Table. 1, we should consider how to set up systematic optimization targets, tailor cache optimization algorithms and design effective API for 3rd-party SPs. Specifically, each user has particular satisfaction functions (e.g., E2E delay, enjoyed bandwidth, service convenience, and savings in access cost), and each 3rd-party SP also has particular requirements (e.g., paid QoS guarantee, reserved bandwidth, and handling of special caching requirements) on its services with a budget, while each MNO also has to optimize the caching resources in the physical environment.

VI. CASE STUDY: COOPERATIVE VM CACHING FRAMEWORK

Proving the feasibility and effectiveness of VM caching at a single caching VM (cVM) is insufficient to verify whether VM caching is promising to drive the whole mobile industry. Instead, we should consider to extend the VM caching concept with detailed optimization techniques over a large coverage with hundreds of VMs, large-scale contents, and a big number of mobile users. VMs can share caching capabilities with each other via high-speed links collaboratively, considering the content popularity, content freshness, user diversity and replica locations, and energy cost of storing and transmissions. In particular, a cVM that is connected with several BSs to serve a number of users in the corresponding local BSs is considered to be the local cVM for the corresponding users. Three main optimization issues are elaborated in this case study by using the system model in [5]. Note that this framework can be extended to more complicated ones in realistic scenarios.

- ***Minimization of inter-MNO traffic (outbound traffic)***: Due to the cost of exchanging data traffic among MNOs,

$$\boxed{\begin{array}{l} Maximize: \sum^{any\ cVM} \sum^{any\ content} (CachedInLocalcVM * Size * Populariy \\ \qquad\qquad\qquad\qquad\qquad + CachedInOthercVMs * Size * Populariy) \\ Constraints: Each\ cVM's\ cache\ storage\ limit\ cannot\ be\ exceeded \\ \qquad\qquad\quad QoS\ of\ mobile\ users\ should\ be\ satisfied\ (fully\ or\ partially). \\ \qquad\qquad\quad QoS\ requirement\ from\ 3rd - party\ Service\ Providers/ContentProviders \\ \qquad\qquad\qquad\quad should\ be\ satisfied\ (fully\ or\ partially). \end{array}} \tag{1}$$

$$\boxed{\begin{array}{l} Maximize: \sum^{any\ cVM} \sum^{any\ content} (CachedInLocalcVM * Size * Popularity) \\ Constraints: Each\ cVM's\ cache\ storage\ limit\ cannot\ be\ exceeded \\ \qquad\qquad\quad QoS\ of\ mobile\ users\ should\ be\ satisfied\ (fully\ or\ partially). \\ \qquad\qquad\quad QoS\ requirement\ from\ 3rd - party\ Service\ Providers/ContentProviders \\ \qquad\qquad\qquad\quad should\ be\ satisfied\ (fully\ or\ partially). \end{array}} \tag{2}$$

it is desired to reduce the inter-MNO traffic as much as possible.

- ***Minimization of intra-MNO traffic (traffic within the SP)***: This can be achieved by caching the most popular contents at each cVM such that most user requests can be locally satisfied with minimal data transfers are needed among cVMs.
- ***Minimization of content access delay of all users***: Contents requested be users may be fetched from local cVMs, routers in the RAN and EPC, and even remote servers, with different delays. To enhance the users' QoE, it is important to minimize the content access delays of users by caching critical contents as close to the users as possible.

Besides, in the above optimization problems, we assume that each content is entirely cached or not in each cVM in the network, in order to reduce the complexity of content management as well as the signalling overhead among cVMs. Thus, the formulated problems are linear/non-linear binary nonconvex optimization problems, which are either NP-complete or NP-hard. Moreover, considering the scale of contents and cVMs in practical implementations, it is not practicable or even impossible to obtain the optimal solutions with exact methods due to the exponential computation complexity. Instead, one needs to find low-complexity approximate methods or heuristical methods to obtain sub-optimal solutions that are suitable for practical implementation.

A. MINIMIZATION OF INTER-MNO TRAFFIC LOAD

The inter-domain routing and transmissions among MNOs usually incur costly expenses. To reduce the cost of inter-MNO data transfers, MNOs should design the caching policy properly by effectively caching contents with a high diversity so that most content requests can be satisfied by cVMs within the MNO's network. Therefore, we aim to minimize the traffic load that goes outside the MNO, which is equivalent to maximizing the satisfaction, i.e., hitting ratio, of content requests from all users within the MNO's network, either by the users' local cVMs or by other cVMs, as shown in (1) at the top this page. In (1), the first term in the objective function represents the content deliveries satisfied by users' local cVMs, and the last term denotes the content deliveries satisfied by other cVMs within the MNO's network. Some constraints are considered to make sure that the cached contents will not exceed the storage limitation, and that QoS requirements of users and 3rd-party SPs should be satisfied fully or partially.

This can be easily achieved as long as the contents cached in all cVMs have the greatest diversity so that arbitrary request for any content can be mostly satisfied by the MNO itself. One intuitive solution may be that, as the cVMs share content storage, for any content, at most one copy is stored inside the MNO's cVMs based on the popularity and the storage size of cVMs. As studied in [5], this optimization problem can be converted into a regular 0-1 multi-knapsack problem (MKP), and can be solved with a polynomial-time approximate algorithm to achieve the caching strategy.

B. MINIMIZATION OF INTRA-MNO TRAFFIC LOAD

To properly maintain the network devices for supporting MNOs' traffic load, it is also important to decrease the intra-MNO transmissions considering the costs of the deployment, operation and maintenance of cables and transport devices in the MNO's network. Thus, the second optimization objective is to minimize the traffic load within the MNO's network, which is to maximize the user demands that can be satisfied locally by each cVM. Thus, in this problem, cVMs just need to satisfy user demands locally as much as possible, as shown in (2) at the top this page. Besides, some constraints are considered on the aspects of storage limitations and QoS requirements of users and 3rd-party SPs.

As a result, the corresponding optimization problem can be formulated as a generalized assignment problem (GAP) [5]. The problem can further be split into a series of independent regular 0-1 single knapsack subproblems that can be solved in parallel with low-complexity greedy algorithms. Note that

$$\boxed{\begin{array}{ll} Minimize: \overset{any\ cVM}{\sum}\ \overset{any\ content}{\sum} & (CachedInLocalcVM * Popularity * D_{IntracVM} \\ & +CachedInOthercVMs * Popularity * (D_{IntracVM} + D_{IntercVM}) \\ & +NotCachedInMNO * Popularity * (D_{IntracVM} + D_{IntraMNO} + D_{Internet})) \\ \multicolumn{2}{l}{Constraints: Each\ cVM's\ cache\ storage\ limit\ cannot\ be\ exceeded} \\ \multicolumn{2}{l}{\qquad\qquad Inter - MNO\ and\ Intra - MNO\ traffic\ should\ be\ minized\ (fully\ or\ partially).} \\ \multicolumn{2}{l}{\qquad\qquad QoS\ requirement\ from\ 3rd - party\ Service\ Providers/ContentProviders} \\ \multicolumn{2}{l}{\qquad\qquad\qquad should\ be\ satisfied\ (fully\ or\ partially).} \end{array}} \tag{3}$$

this optimization objective may induce situations when many cVMs cache the same popular contents, which is somehow contradictory to the requirements of content diversity in the aforementioned objective.

C. MINIMIZATION OF USER DELAYS

It is important for MNOs to reduce the inter-MNO and intra-MNO traffic load. However, one essential target of the MNO is to improve mobile users' QoE, and among many factors, the most important one is the user delay, which can be defined as the round-trip time for delivering a requested content to a mobile user over the MNO's network. Thus, the third optimization objective is to minimize the overall user delays, as shown in (3) at the top this page.

In the objective function, the first part is about the delays of user requests satisfied at local cVMs, the second part is about those satisfied by neighbor cVMs within the same MNO, and the third part is about fetching content files from remote SPs' servers. Moreover, some constraints are also considered based on the storage limitations, inter-MNO and intra-MNO traffic load, and QoS requirements from 3rd-party SPs. The formulated optimization problem is non-linear and non-convex, and it is similar to the optimization problem that has been discussed in [4]. To solve the problem, in addition to providing an approximate transformation, an equivalent transformation can convert the complex problem into a linear programming problem. Rather than using exact optimal methods (e.g., branch-and-bound (BNB) method) with exponential-time and exponential-space complexity [21], [22], a distributed suboptimal algorithm, which has polynomial-time and linear-space complexity, is proposed in [4].

D. NUMERICAL RESULTS

In order to evaluate the three optimization frameworks over large-scale coverage of an MNO's cVMs, we carry out numerical analysis and simulations with the following assumptions and setting. The overall popularity of each content in the network follows Zipf distribution with the popularity of contents in each cVM is random in (0, 1], and the default exponent factor is $\beta = 0.65$. The size of each content is random in [0.001, 1] Gbits, the delay within cVMs is random in [5, 10] ms, while the delay among cVMs is random in [20, 50] ms, and the delay outside the MNO network is random in [100, 200] ms.

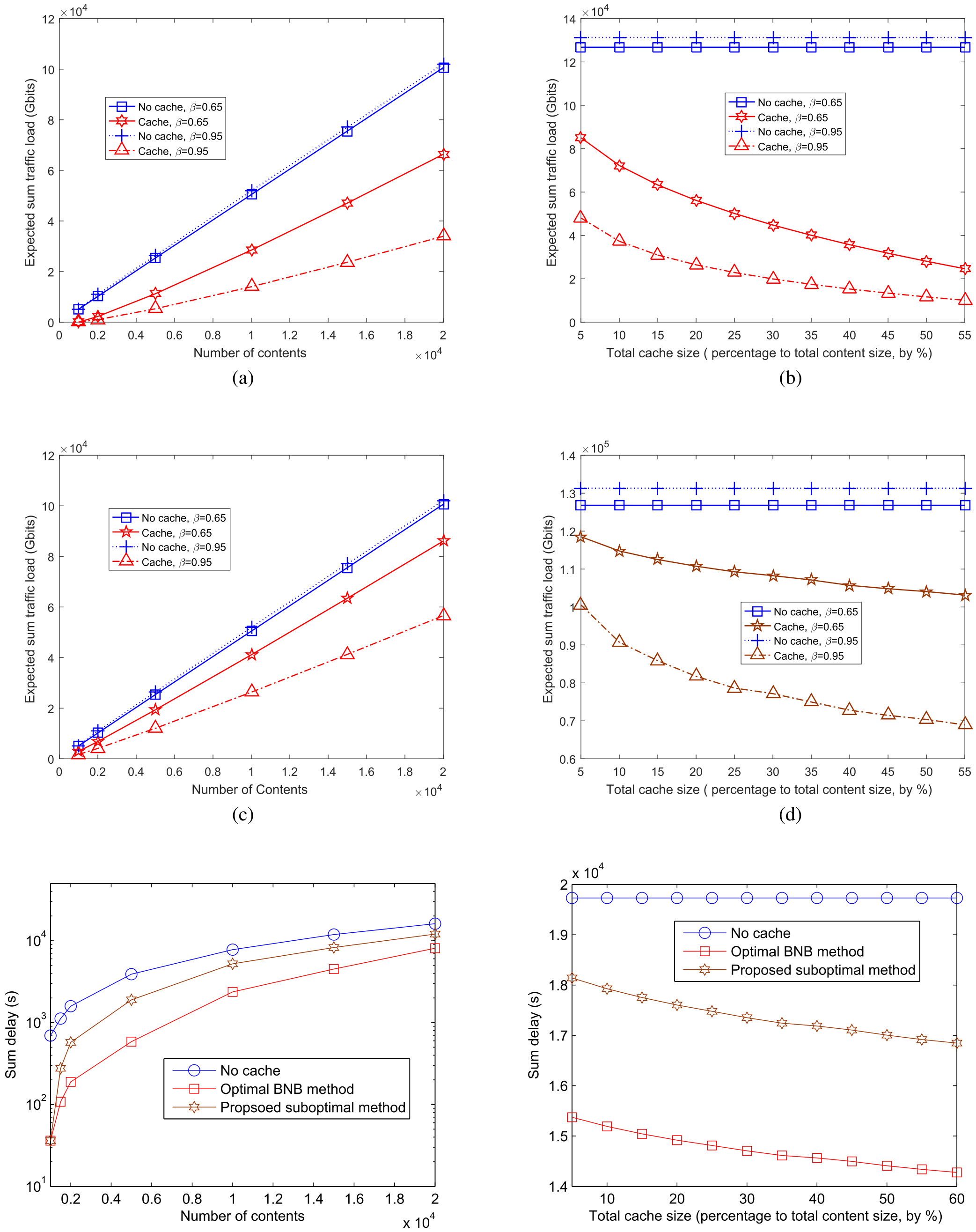
We compare the performance of non-caching strategy with our cooperative VM caching framework in terms of expected inter-MNO and intra-MNO traffic load as well as the sum delay versus number of contents and the size of total cache. The numerical results are shown in Fig. 4. Here, some specific parameters are set as: 1) In Fig. 4(a), Fig. 4(c) and Fig. 4(e), the number of cVMs is 10 and cache size of each cVM is 50 Gbits; 2) In Fig. 4(b), Fig. 4(d) and Fig. 4(f), the numbers of cVMs and contents are 50 and 5,000, respectively. From Fig. 4, it is obvious that cooperative cVM caching outperforms non-caching strategy significantly. Specifically, we can observe that all the inter-MNO traffic load, intra-MNO traffic load and the content delivery delays increase with the increase of the number of contents. Another observation is that as the total size of caches increases, the traffic load can be greatly reduced, and the content delivery delays are reduced as well.

VII. TECHNICAL CHALLENGES AND OPEN ISSUES

The emerging consensus on the emerging 5G systems is that these systems will not be based on a single technology, but rather a synergistic collection of interworking technical innovations and solutions that address the challenges of traffic growth and enhanced performance regarding various advanced mobile services. Therefore, incorporation of the CaaS framework within 5G systems is an attractive solution to address some of these challenges. However, the design, implementation and deployment of the CaaS framework bring a new set of challenges and interesting open technical issues for researchers and engineers. We discuss the core ingredients and consider the key enabling technologies for CaaS in 5G era, which require further in-depth investigations, as follows.

A. UTILIZING SDN TECHNIQUES WITH VIRTUAL CACHING

The separation of data and control planes in SDN enables the operator to flexibly steer the data traffic and assign resources to address the QoS requirements of different traffic flows inside the network. MNO can leverage SDN to provision CaaS by taking advantage of the dynamic control capability within an SDN to support the fast provisioning and content-based adaptation of caching resources within its network and extending such control to 3rd-party SPs. The latter calls for the development and standardization of protocols and APIs for the SDN control plane to enable 3rd-party SPs to effectively programming their services via CaaS. Furthermore, efficient algorithms leveraging the SDN architecture to enable fine-grained in-network caching and re-utilization of pre-, in-, and post-processing virtual



(a) (b)

(c) (d)

(e) (f)

FIGURE 4. Numerical evaluation of cooperative VM caching optimization framework. (a) Inter-MNO Traffic Optimization - Number of Contents Varies. (b) Inter-MNO Traffic Optimization - Cache Size Varies. (c) Intra-MNO Traffic Optimization - Number of Contents Varies. (d) Intra-MNO Traffic Optimization - Cache Size Varies. (e) Delay Optimization - Number of Contents Varies. (f) Delay Optimization - Cache Size Varies.

resources (VMs, temporary data, user sessions, etc.) in the virtual environment provisioned by the MNO's large scale data centers need further investigations and development.

B. VIRTUAL CACHING FOR LOCATION-BASED SERVICES

Location-based services (LBSs), which bind mobile users with locally available services based on the users' locations, can be highly profitable for MNOs and 3rd-party SPs. Through appropriate APIs for 3rd-party SPs, CaaS can enable them to flexibly deploy cached contents at locations close to where the targeted users are currently found. However, to take advantage of such flexibility, optimized decision algorithms need to be developed to enable the 3rd-party SPs to determine where contents should be cached to minimize cost while providing satisfactory QoE to users. For instance, mobile advertising images and videos can be cached at particular places for pushing to various users based on data mining. Contents can be cached in the cells or in Wi-Fi access points or VMs, and 3rd-party SPs and MNOs can dynamically adjust the cache allocation using appropriate algorithms based on the flow of human traffic. It may be interesting to develop effective auction-based business models to accelerate the adoption of virtual caching for LBSs.

C. DEVICE-TO-DEVICE CACHING AND SHARING

In reality, users are often clustered in crowds, e.g., in cafeterias, restaurants, subway trains, apartments and so on. Device-to-device (D2D) communications using the MNO's cellular spectrum can provide further opportunities for users to cache and share interesting contents with each other, while MNOs and 3rd-party SPs can effectively exploit and facilitate D2D-based caching for improved utilization of network resources. For instance, the aforementioned virtual caching LBSs can be further extended in to ad-hoc D2D networks (formed by groups of users with certain maintenance algorithms) as well. 3rd-party SPs do not need to explicitly manage the cache distribution algorithm but just utilize MNO's APIs to specify the requirements of caching certain content targeting a particular place for a group of users. One of the technical challenges of this scenario is how to maintain the persistency and consistency of virtual caching for a group of mobile users with dynamic mobility as well as a large variety of content demands.

D. SOCIALIZED VIRTUAL CACHING

Efficient deployment of CaaS needs to take into account the impact of social networking services (SNSs), especially when the transferred content is relevant to the social activities of the users. Strong social relations within a group always induce a high probability of delivering the same content to all the group members, and thus caching the content at a proper position in advance may save network resources while enhancing QoE. Also, through its API with 3rd-party SPs, CaaS can extend its social network analysis capability to the 3rd-party SPs to enable them to design and deploy suitable algorithms for ensuring that the relevant contents are kept close to the interested users [23], so that mobile users may enjoy a high hit-rate for cached contents within the serving MNO's network. Finally, there is a large potential to design better caching strategies and protocols, based on evaluations using realistic social network traces and CDN data sets.

VIII. CONCLUSIONS

In this paper, we have studied virtual caching at mobile network edges, which practical deployment is increasing attractive due to the incorporation of resource virtualization techniques in emerging 5G mobile networks. We have further discussed and extended the conceptual framework of "Caching-as-a-Service", which offers flexible and scalable caching service from mobile network operators to 3rd-party service providers. Various representative applications and scenarios utilizing CaaS have been discussed, and also for services with different requirements on bandwidth, delay, jitter tolerance, programmability, and elasticity, we have identified several performance evaluation metrics as engineering optimization targets. An interesting direction for future research is the development of practical and effective algorithms for caching resource optimization and flow control for CaaS in emerging 5G networks. Practical design and standardization of CaaS APIs for reliability, scalability, and complexity will also be needed.

REFERENCES

[1] CISCO, "Cisco visual networking index: Global mobile data traffic forecast update, 2013–2018," Tech. Rep., 2014.

[2] X. Wang, A. V. Vasilakos, M. Chen, Y. Liu, and T. T. Kwon, "A survey of green mobile networks: Opportunities and challenges," *Mobile Netw. Appl.*, vol. 17, no. 1, pp. 4–20, Feb. 2012.

[3] X. Wang, X. Li, V. C. M. Leung, and P. Nasiopoulos, "A framework of cooperative cell caching for the future mobile networks," in *Proc. HICSS*, Jan. 2015, pp. 5404–5413.

[4] X. Li, X. Wang, S. Xiao, and V. C. M. Leung, "Delay performance analysis of cooperative cell caching in future mobile networks," in *Proc. IEEE ICC*, Jun. 2015, pp. 5652–5657.

[5] X. Li, X. Wang, C. Zhu, W. Cai, and V. C. M. Leung, "Caching-as-a-service: Virtual caching framework in the cloud-based mobile networks," in *Proc. IEEE INFOCOM, WKSHPs*, Apr./May 2015, pp. 372–377.

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

[7] E. H. Ong, J. Y. Khan, and K. Mahata, "Radio resource management of composite wireless networks: Predictive and reactive approaches," *IEEE Trans. Mobile Comput.*, vol. 11, no. 5, pp. 807–820, May 2012.

[8] X. Li, X. Ge, X. Wang, J. Cheng, and V. C. M. Leung, "Energy efficiency optimization: Joint antenna-subcarrier-power allocation in OFDM-DASs," *IEEE Trans. Wireless Commun.*, vol. 15, no. 11, pp. 7470–7483, Nov. 2016.

[9] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *Proc. ACM MobiSys*, Jun. 2013, pp. 319–332.

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

[11] X. Jin, L. Li, L. Vanbever, and J. Rexford, "Softcell: Scalable and flexible cellular core network architecture," in *Proc. ACM CoNEXT*, Dec. 2013, pp. 163–174.

[12] H. Hawilo, A. Shami, M. Mirahmadi, and R. Asal, "NFV: State of the art, challenges, and implementation in next generation mobile networks (vEPC)," *IEEE Netw. Mag.*, vol. 28, no. 6, pp. 18–26, Nov./Dec. 2014.

[13] China Mobile Res. Inst., *C-RAN: The Road Towards Green RAN, Version 2.5*, White Paper, Beijing, China, Oct. 2011.

[14] S. R. Srinivasan, J. W. Lee, D. L. Batni, H. G. Schulzrinne, "ActiveCDN: Cloud computing meets content delivery networks," Columbia Univ., New York, NY, USA, Tech. Rep. CUCS-045-11, 2012.

[15] H. AhleHagh and S. Dey, "Video caching in radio access network: Impact on delay and capacity," in *Proc. IEEE WCNC*, Apr. 2012, pp. 2276–2281.

[16] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "FemtoCaching: Wireless video content delivery through distributed caching helpers," in *Proc. IEEE INFOCOM*, Mar. 2012, pp. 1107–1115.

[17] S. Singh, O. Oyman, A. Papathanassiou, D. Chatterjee, and J. G. Andrews. (2012). *Tech. Talk 4: Video Capacity and QoE Enhancedments over LTE*. [Online]. Available: http://www.profheath.org/wp-content/uploads/2011/09/sarabjot_icctalk_2012_v3.pdf

[18] F. Dobrian *et al.*, "Understanding the impact of video quality on user engagement," in *Proc. ACM SIGCOMM*, Aug. 2011, pp. 362–373.

[19] D. Whitney. (2011). *Study: Viewers Only Wait Two Seconds for Online Video to Start*. [Online]. Available: http://www.mediapost.com/publications/article/187083/study-viewers-only-wait-two-seconds-for-online-vi.html

[20] W. Cai, C. Zhou, M. Li, X. Li, and V. C. M. Leung, "MCG test-bed: An experimental test-bed for mobile cloud gaming," in *Proc. ACM Mobile Games@MobiSys*, May 2015, pp. 25–30.

[21] V. Jain and I. E. Grossmann, "Algorithms for hybrid MILP/CP models for a class of optimization problems," *INFORMS J. Comput.*, vol. 13, no. 4, pp. 258–276, Nov. 2001.

[22] S. Martello and P. Toth, *Knapsack Problems: Algorithms and Computer Implementations*. New York, NY, USA: Wiley, 1990.

[23] S. Salvatore, M. Musolesi, C. Mascolo, and J. Crowcroft, "Track globally, deliver locally: Improving content delivery networks by tracking geographic social cascades," in *Proc. WWW*, Mar./Apr. 2011, pp. 457–466.

XIUHUA LI (S'12) received the B.S. and M.S. degrees from the Honors School and the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin, China, in 2011 and 2013, respectively. He is currently pursuing the Ph.D. degree with the Department of Electrical and Computer Engineering, The University of British Columbia, Vancouver, BC, Canada. His current research interests are resource allocation, optimization, distributed antenna systems, cooperative base station caching, and traffic offloading in mobile content-centric networks.

XIAOFEI WANG (S'06–M'13) received the B.S. degree from the Department of Computer Science and Technology, Huazhong University of Science and Technology in 2005, and the M.S. and Ph.D. degrees from the School of Computer Science and Engineering, Seoul National University in 2008 and 2013, respectively.

He was a Post-Doctoral Research Fellow with the Department of Electrical and Computer Engineering, University of British Columbia. He is currently a Professor with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University. His current research interests are social-aware multimedia service in cloud computing, cooperative base station caching, and traffic offloading in mobile content-centric networks.

KEQIU LI (S'04–M'05–SM'12) received the bachelor's and master's degrees from the Department of Applied Mathematics, Dalian University of Technology, in 1994 and 1997, respectively, and the Ph.D. degree from the Graduate School of Information Science, Japan Advanced Institute of Science and Technology, in 2005. He held a post-doctoral position with the University of Tokyo, Japan, for two years.

He is currently a Professor with the Tianjin Key Laboratory of Advanced Networking, School of Computer Science and Technology, Tianjin University, China. He has authored over 100 technical papers in journals, such as the IEEE Transactions on Parallel and Distributed Systems, the ACM Transactions on Internet Technology, and the ACM Transactions on Multimedia Computing, Communications, and Applications. He is also an Associate Editor of the IEEE Transactions on Parallel and Distributed Systems and the IEEE Transactions on Computers. His research interests include data center networks, cloud computing, and wireless networks.

VICTOR C. M. LEUNG (S'75–M'89–SM'97–F'03) received the B.A.Sc. degree (Hons.) in electrical engineering from The University of British Columbia (UBC) in 1977, and the Ph.D. degree in electrical engineering in 1982. He attended graduate school at UBC on a Canadian Natural Sciences and Engineering Research Council Postgraduate Scholarship. He received the APEBC Gold Medal as the head of the graduating class from the Faculty of Applied Science.

From 1981 to 1987, he was a Senior Member of Technical Staff and a Satellite System Specialist with MPR Teltech Ltd., Canada. In 1988, he was a Lecturer with the Department of Electronics, The Chinese University of Hong Kong. He returned to UBC as a Faculty Member in 1989. He is currently a Professor and the TELUS Mobility Research Chair in advanced telecommunications engineering with the Department of Electrical and Computer Engineering. He has co-authored over 950 technical papers in international journals and conference proceedings, 37 book chapters, and co-edited 12 book titles. Several of his papers had been selected for Best Paper Awards. His research interests are in the areas wireless networks and mobile systems.

Dr. Leung is a registered Professional Engineer in the Province of British Columbia, Canada. He is a Fellow of the Royal Society of Canada, the Engineering Institute of Canada, and the Canadian Academy of Engineering. He was a Distinguished Lecturer of the IEEE Communications Society. He is serving on the Editorial Boards of the IEEE Wireless Communications Letters, the IEEE Transactions on Green Communications and Networking, the IEEE Access, the *Computer Communications*, and several other journals, and has previously served on the Editorial Boards of the IEEE Journal on Selected Areas in Communications–Wireless Communications Series and Series on Green Communications and Networking, the IEEE Transactions on Wireless Communications, the IEEE Transactions on Vehicular Technology, the IEEE Transactions on Computers, and the *Journal of Communications and Networks*. He has guest-edited many journal special issues, and provided leadership to the organizing committees and technical program committees of numerous conferences and workshops. He was a recipient of the IEEE Vancouver Section Centennial Award and 2012 UBC Killam Research Prize.