Performance and Implications of RAN Caching in LTE Mobile Networks: a Real Traffic Analysis

Tao Lin*, Hongjia Li*, Haiyong Xie†‡, Jiasi Chen§, Huajun Cui*, Guoqiang Zhang¶, Wei An*, and Yang Li*
*Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
†University of Science and Technology of China, Suzhou, China
‡Innovation Center, China Academy of Electronics and Information Technology, Beijing, China
§Department of Computer Science and Engineering, University of California, Riverside, USA
¶School of Computer Science and Technology, Nanjing Normal University, Nanjing, China

***Abstract*—Deploying caches in mobile networks, especially in the radio access network (RAN) is regarded as a promising way to improve mobile user experiences and alleviate the increasing pressure of traffic growth. However, the characteristics of mobile traffic and the performance of RAN caching still remains unclear. In this paper, we extensively analyze the traffic characteristics, the content popularity and the cache performance using a unique dataset collected from a commercial LTE network of China Mobile, from the perspective of mobile access network. The dataset spans nearly a week and consists of a collection of approximately 62.1 millions HTTP sessions, generated by more than 3200 users distributed across three base stations. Based on this realistic dataset, we observe that HTTP traffic can be reduced by 24.4% on average and the hit ratio can reach up to 42.2%, using 100GB cache size. The implications on some fundamental design issues of practical RAN caching systems, including reasonable size of RAN cache, suitable locations of cache deployment and potential benefits of collaborative RAN caching, are further presented. We believe our findings will shed light on practical RAN caching system design.**

I. Introduction

With the deployment of long term evolution (LTE) mobile networks and prosperity of mobile applications, mobile data traffic has been growing rapidly in recent years. According to the recent Cisco VNI report [1], global mobile data traffic grew 69 percent in 2014 and will increase nearly tenfold between 2014 and 2019. The growth of mobile data traffic, particularly with a large number of duplicated transmissions of multimedia content, brings increasing pressure on the mobile network in terms of network performance and operational cost.

Deploying caches in the mobile network, especially in the radio access network (RAN), is regarded by both academia [2], [3], [4] and industry [5], [6] as an effective way to reduce the redundant traffic and improve the experience of mobile users. A working group in European Telecommunications Standards Institute (ETSI), namely mobile edge computing (MEC) [7], was recently established by some leading mobile operators and industrial companies, such as Vodafone, Intel and HUAWEI. In the vision of MEC, RAN devices will have storage and computing capabilities, in addition to traditional functions of communication. Content caching in the RAN, or RAN caching for short, is regarded as one of the most important scenarios in MEC. Furthermore, content distribution and caching in the base station is also considered one of the most disruptive technologies in the future fifth generation (5G) mobile network [8], due to the trend towards computation at the edge, e.g., network function virtualization (NFV) [9], and the movement towards common-off-the-shelf (COTS) hardware with storage capabilities [10].

In order to understand the traffic characteristics and caching performance of mobile networks, recent research based on realistic traffic, e.g., [2] [4] for 3G networks, [11] for LTE network, have shown the effectiveness of mobile caching. However, the observations reported so far mainly focus on the caching performance of the *mobile core network*, e.g., evolved packet core (EPC), which aggregates a large number of mobile users. Compared with deploying cache in mobile core networks, RAN caching has more advantages in reduction of back-haul traffic and improvement of user experience, since contents can be retrieved at base stations. However, the characteristics of traffic and the performance of *RAN caching* in LTE mobile networks remains unclear.

From the point of view of the RAN, in this paper, we analyze the traffic characteristics, the content popularity distribution and the cache performance, using a unique dataset collected from a commercial LTE network of China Mobile, a leading mobile network operator in the world. The dataset spans six and a half days in April 2015 and consists of a collection of approximately 62.1 millions HTTP records which are generated by more than 3200 users distributed across three base stations, namely LTE evolved nodeBs (eNodeB).

To the best of our knowledge, this is the first study that thoroughly analyzes the performance of RAN caching in LTE mobile networks using real traffic. Our study helps answer the following questions: 1) does the content popularity still follow a Zipf distribution in the scope of the RAN, as apposed to the mobile core? 2) what performance can RAN caching achieve in terms of cache hit ratio and backhual traffic reduction? 3) how much storage does a RAN cache need in practice? 4) what is the relationship between the access user number of an eNodeB and the corresponding cache hit ratio? Our main findings and observations are summarized as follows.

- Among the 62.1 millions HTTP sessions observed over the measurement period, 21.5% HTTP sessions

978-1-5090-1732-4/16/$31.00 ©2016 IEEE

are cacheable. However, the traffic volume of these cacheable HTTP sessions account for 56.2% of the total HTTP traffic.

- Requests for images constitute 94.0% of the total cacheable HTTP requests, while the percentage of video requests is no more than 3%. However, the traffic of video requests takes up 33.9% of the total cacheable HTTP traffic.
- Over the measurement period, the popularity of cacheable contents follows a Zipf distribution with parameter α equal to 0.8, with 3200 average access users per day. This means that the requests for the top 10% popular contents account for 48.2% of total cacheable HTTP requests.
- With cache size increasing from 5 GB to 1000 GB, the cache hit ratio rises correspondingly and converges to a maximum hit ratio of 46.3%. Furthermore, we observe that the growth of hit ratio flattens out when the cache size reaches 100 GB, and the corresponding hit ratio is 42.2%, which is close to the maximum value. This indicates that 100 GB may be a reasonable RAN cache size, considering the tradeoff between the cache cost and the hit ratio, based on our measured LTE network.
- With 100 GB cache size and an average of 3200 access users per day, RAN caching can save as much as 24.4% of the total HTTP downstream traffic (including both cacheable and un-cacheable traffic) on average, with maximum savings of 27.9%.
- We investigate the effect of user number ranging from 50 to 2000, on the cache hit ratio. Not surprisingly, with increasing user number, the cache hit ratio correspondingly improves from an average value of 19.1% to 39.7%. It is worth noting that even with a smaller user number, namely 50 users, the average cache hit ratio can reach 19.1%, albeit with high volatility.
- The effect of user number on the saved traffic is also investigated. When the user number ranges from 50 to 2000, the volume of total HTTP downstream traffic is reduced by 15.9% to 23.2% on average. Similarly, it is observed that the amount of saved traffic is more volatile when the user number is small.

These observations validate the effectiveness of RAN caching. Based on above observations, we further explore the implications on some fundamental design issues of practical RAN caching systems, including reasonable size of RAN cache and suitable locations for deploying RAN cache. Moreover, the feasibility and benefits of collaborative RAN caching is investigated. We believe that our findings shed light on the design of content distribution and caching systems in future mobile networks.

II. Dataset Description

In this section, we introduce our dataset collected from a commercial LTE mobile network of China Mobile, one of the largest mobile network operator in the world, in a major provincial capital city of southern China. A packet analyzer is used to capture and analyze the packets flowing on the link between the packet data network gateway (P-GW) and the Internet, as illustrated in Figure 1. In this paper, we focus on HTTP traffic, including all HTTP requests and corresponding replies, which is considered as the dominant traffic type on current LTE mobile networks. Each detected HTTP request/reply is recorded as a session in a log file of the packet analyzer.

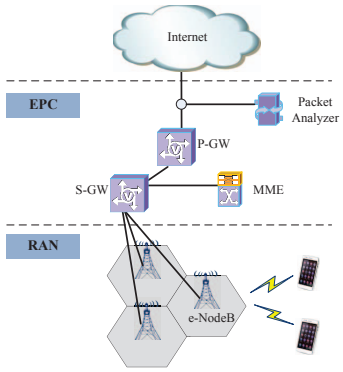

Fig. 1. Illustration of LTE network architecture and packet analyzer

Our dataset was collected from 0 : 00 AM on April 10 to 12 : 00 AM on April 16 in 2015. During the six and a half days, the packet analyzer observed the HTTP traffic of three eNodeBs. These eNodeBs are selected from three typical access scenarios in an urban region, namely, campus area, residential area and commercial area, respectively.

The average access user number of the total three eNodeBs is 3249 per day. Note that an access user is defined as a unique mobile terminal that has at least one complete HTTP request/reply session during a day. The number of access users is defined as the total number of users accessing the network in a single day. In this paper, the "number of access users" is exchangeable with the "number of users".

During the measurement period, the total number of HTTP sessions in the log file is 62.1 millions. Each session records a number of fields and the meanings of several main fields are defined in Table I.

It is worth noting that the RAN caching performance of each specific eNodeB cannot be determined, since our dataset does not contain the eNodeB association information of each access user. However, this has minimal impact on the analysis of RAN caching performance.

III. Mobile Traffic Characteristics

A. Definition of Cacheable Content

Cacheable content means that the content can be cached by the cache and reused by the subsequent requests according to some pre-defined rules. In this paper, we essentially follow the HTTP 1.1 standard [12], e.g., mainly according to the request method, request header fields, and the response status, to decide whether a content object is cacheable.

However, we make some minor modifications as follows, from a practical point of view. Firstly, only when the content-type field is video, image, application or audio, may the

TABLE I. MAIN FIELDS OF HTTP SESSION

Field	Meaning
Timestamp	The time when a HTTP request is initiated
User ID	Anonymous user identifier
Content ID	The unique resource locator (URL) of one requested content
HTTP Method	A token obtained from HTTP request message, e.g., GET, HEAD, POST
Content Type	A value obtained from the content type field in HTTP reply message, including text, image, video, audio, application and others
Content Length	The length of requested content in bytes, obtained from HTTP reply message
Downlink Length	Actual length of downlink transmission data
HTTP Status Code	Indication of the implementation state of a specific HTTP session, e.g., 200, 203, 300

content be considered as cacheable content. The content objects with the type of text is deemed un-cacheable since text consumes a small number of bytes and caching a text object makes no sense for a practical cache system, although a text object is technically cacheable. Secondly, we disregard the Cache-control field in the HTTP messages when making the cacheability decision to maximize the benefits of caching system, even though the Cache-control field is set to be "no-cache" or "no-store". The reason behind this is that some content providers, especially for on-line video websites, usually utilize the field to prevent their contents from being cached for the sake of copyright.

According to our definition of cacheable, there are $62,104,921$ requests in total, of which the number of cacheable content requests is $13,354,555$, comprising 21.5% of the total requests. The total downstream HTTP traffic volume is 1776 GB, where the cacheable traffic is 998 GB and accounts for 56.2% of the total traffic.

B. Proportion of Requests and Traffic Volume by Content Type

Figure 2 and Figure 3 show the HTTP request number and traffic volume distribution of different content types in the measured dataset, respectively.

The content types in the measured dataset consist of video, image, application, text, audio and others. Note that the content type of application represents the data of various kinds of internet applications, such as Acrobat PDF and Javascript, not just mobile applications. In Figure 2(a), we observe that

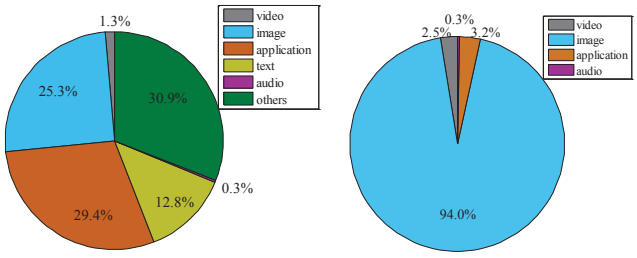

Fig. 2. The HTTP request number distributions of different content types. (a) All contents; (b) Cacheable contents.

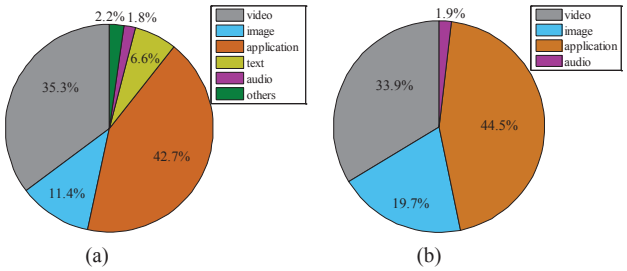

Fig. 3. Downlink traffic volume distributions of different content types. (a) All contents; (b) Cacheable contents.

application and image content account for 29.4% and 25.3% of the total number of requests. However, as seen from Figure 2(b), because most objects of application requests are unique and not cacheable, the percentage of application requests falls from 29.4% (for all-content dataset) to 3.2% (for cacheable content dataset) when considering cacheable content only. Comparing Figure 2(a) with 2(b), we find that the percentage of image requests increases from 25.3% to 94.0%, which dominates all other cacheable content types. This implies that in terms of number of requests, image objects deserve a more attention in the design of a RAN caching system.

As seen from Figure 3(a), the volume of application traffic reaches to an overwhelming 42.7% of the total traffic. Comparing Figure 3(b) with Figure 2(b), the volume of cacheable application traffic accounts for 44.5% of the total cacheable traffic, although the percentage of application requests is only 3.2% in the cacheable content dataset. The reason behind this is that the cacheable application traffic is mainly composed of downloads of various Android or iOS mobile softwares, which usually have a large file size. Similarly, although the percentage of video requests is only 2.5% in the cacheable content dataset, mobile video traffic accounts for 33.9% of the total traffic in the cacheable content dataset.

C. Traffic Volume over the Measurement Period

Figure 4 shows the hourly traffic load over the measurement period of 7 days. We observe a clear diurnal pattern where the heavier traffic load occurs during the daytime between 8 am and 12 pm. The peak traffic load is about 36 GB/hour at 12 pm (80 Mbps as per-hour average), and the lowest traffic load is 0.5 GB/hour at 4 am, 72x smaller than the peak one. Interestingly, the traffic load during the weekend, i.e., the first two days, is slightly larger than that of the weekdays. We also observe the traffic load on every weekday is bimodal. Specifically, one peak appears at lunch time (12 am - 2 pm), and the other appears at midnight, which implies that more people prefer to use smartphones for Internet access during leisure time.

D. Cacheable Content Size Distribution

Figure 5 shows the size distribution of individual cacheable contents, including total contents, image, video, application and audio. From the perspective of total content size, contents

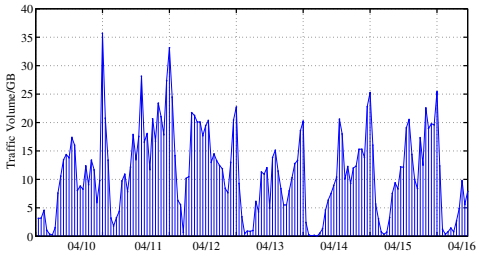

Fig. 4. Traffic volume for one week (each point is a sum of 1-hour traffic load)

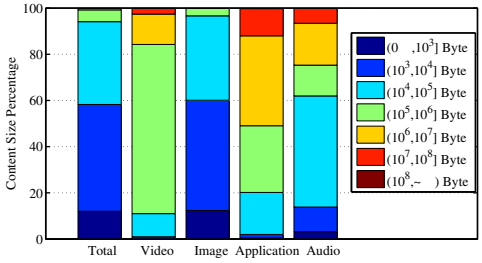

Fig. 5. Cacheable content size distribution

with size $10^3 \sim 10^5$ bytes play a dominating role, with around 82% of the total contents. Meanwhile, images achieve nearly the same distribution as total content because the most of cacheable content consists of images. It is worth noting that the size of requested video is mainly distributed between 10^5 and 10^6 bytes, which indicates mobile users in the measured LTE network are mostly watching small-size and low-grade videos. The reason behind this may be the high access cost of LTE networks in China. However, from another point of view, there is a large room for the video traffic to grow dramatically, with increasing mobile bandwidth and decreasing access cost.

IV. Cacheable Content Popularity Distribution

A. Content Popularity based on the total traffic of three eNodeBs

Figure 6 shows the cacheable content popularity distribution from April 10 to April 16, including daily distribution as well as the aggregate distribution across all seven days. Two important observations are made: (1) the curves from April 10 to April 16 each feature the same slope value; (2) the slope of each curve has the same value as that obtained from the complete 7-day dataset.

To further quantify these observations, we fit the distribution of content requests using the least squares method [13]. The results show that on the current LTE mobile network, the content popularity follows Zipf's law with respect to content request frequency and popularity rank of content, i.e. $f \cdot r^{\alpha} = C$, where r is the request rank of content i, f is the request frequency of content i, α is the Zipf slope and C is a constant.

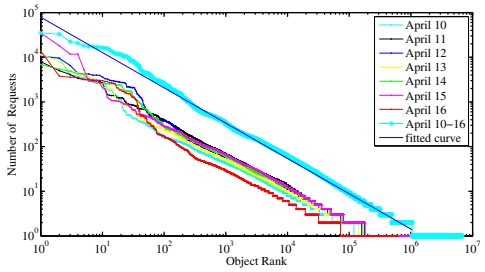

Fig. 6. Cacheable content popularity distributions

TABLE II. The obtained values of α, C and R by fitting all distributions of content requests from April 10 to April 16

Data	α	C	r
2015/04/10	0.7584	3.8551	0.9747
2015/04/11	0.7425	4.0530	0.9974
2015/04/12	0.8403	4.2719	0.9728
2015/04/13	0.7556	3.9325	0.9892
2015/04/14	0.7885	4.1235	0.9819
2015/04/15	0.7356	3.9844	0.9847
2015/04/16	0.9246	4.1816	0.9568
2015/04/10-16	0.8077	4.9569	0.9925

Table II shows the values of α, C and r. The parameter α is the most important, since it determines the concentration of the content popularity: the greater α, the more concentrated the content popularity. The mean value of α is 0.7922, which is very near the parameter $\alpha = 0.8077$ for the total contents from April 10 to April 16, and the 95% confidence interval is $[0.8554, 0.7290]$.

1) Percentage of Requests for Popular Contents

To describe the ratio of popular contents, we calculate the percentage of requests from the top 10% and 20% most popular contents. The total number of cacheable contents is $7,094,425$. The request number of top 10% popular cacheable contents is $6,434,224$, while the total number of requests for cacheable contents is $13,354,555$. Thus, the fraction of requests from the top 10% of cacheable contents is 48.2%. Similarly, the percentage of requests from the top 20% of cacheable content requests is 56.3%.

2) Characteristics of the Top 100 Popular Contents

As shown in Figure 6, the curve formed by the top 100 popular contents does not follow the Zipf distribution. Thus we investigate the characteristics of the top 100 popular contents. In the top 100 contents, there are 95 images, 4 video contents and 1 application content. From April 10 to April 16, the fraction of requests for the top 100 contents is 5.5%. Furthermore, we observe that the top 100 contents are mainly logos of website and advertising figures. Compared with normal contents that mobile users actively retrieve, these contents have much higher request frequency due to marketing strategies that passively push them to users when they are surfing certain websites. Thus, these passively requested website logos and advertising figures deviate the popularity of the top 100 contents from that of contents users really interested in.

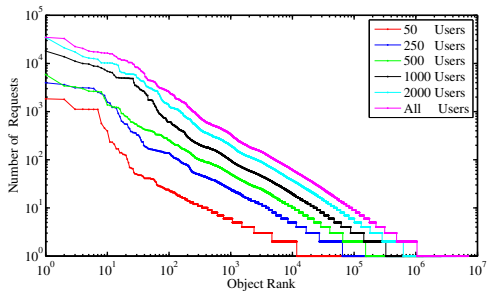

Fig. 7. The request frequency vs the request ranking under different number of users

TABLE III. THE SLOPES OF THE FITTED CURVES FOR ZIPF DISTRIBUTION

User number	50	250	500	1000	2000	All
α	0.64	0.72	0.72	0.74	0.77	0.78

B. Content Popularity verse different user numbers

We randomly select a nubmer of users each day to construct the user set, namely, 50, 250, 500, 1,000, 2,000 and all users, to run the experiments. The content popularity with varying number of users are shown in Figure 7. Firstly and surprisingly, it can be observed that the curve still follows the Zipf distribution even when the number of users is only 50.

The slope values of the fitted Zipf distribution are presented in Table III. We can observe that as the number of users increases, the value of α becomes greater and greater, which implies the requests concentrate to the top ranking contents.

Figure 8(a) and 8(b) plot two typical popularity curves and their fit lines using the least squares method. Recall from the analysis of characteristics of top 100 popular contents in the last subsection, that nearly 100 passively requested website logos deviate the popularity of the top 100 contents from that of contents interested by users in both cases of 50 and 250 users.

V. RAN CACHE PERFORMANCE ANALYSIS

A. Cache hit ratio based on the total three eNodeBs

Since we cannot establish the mapping from requests to the specific eNodeBs from the collected dataset, we simply treat the three eNodeBs as one logical node. We conduct the experiment by initializing the cache node as empty, and then replaying the requests in sequence. We define two phases of a cache: warm-up and stable. During the warm-up phase, the cache is filled from scratch with content until the hit ratio appears to be stable. Figure 9 plots the hit ratio based on the total three eNodeBs over the measured period, with 100 GB cache size and using the least frequently used (LFU) replacement algorithm. In fact, we also compare LFU with the other commonly used algorithm, namely, least recently used (LRU), and find that there is no apparent difference in terms of cache hit ratio. From figure 10, we observe that after the phase of warm-up status of about 24 hours, the hit ratio appears to be stable and the average hit ratio during the stable phase is 42.2%.

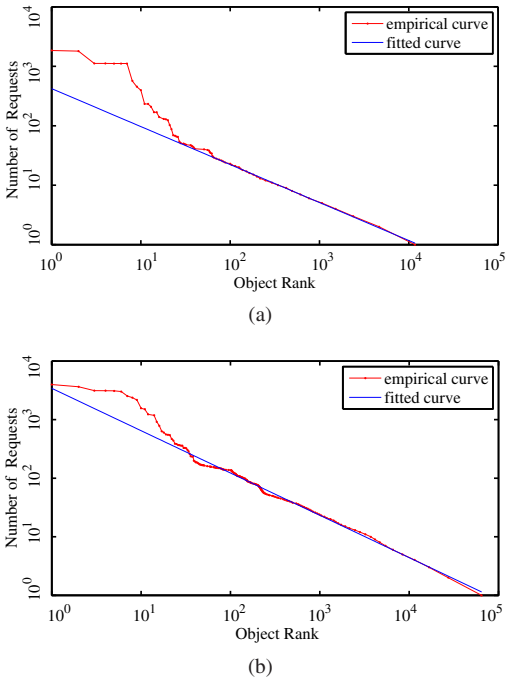

Fig. 8. Two typical popularity curves for 50 and 250 users

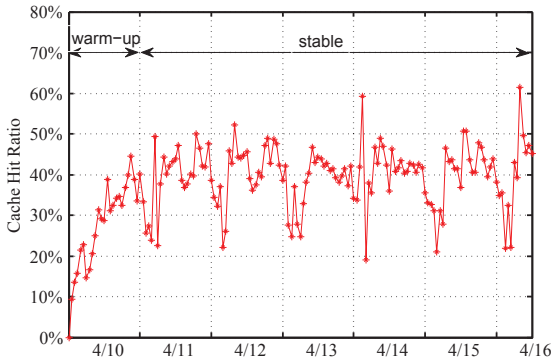

Fig. 9. Cache Hit Ratio of the total three eNodeBs

Based on the above experiment, we can also determine the cache hit ratio of different content types during the measurement period, as shown in Figure 10. This result shows that RAN caching is effective for video, image, application and audio content since the corresponding hit ratios are equal to or above 41.9%. We also observe that compared to image and video, application and audio have a higher hit ratio of 67.0% and 60.2%, respectively. This suggests that the content popularity of application and audio is more concentrated. The rational behind it may be the fact that the diversity and the amount of video and image content items available are relatively more plentiful than that of application and audio.

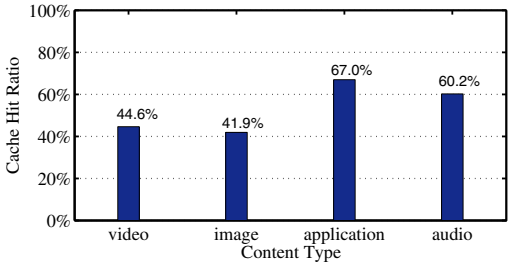

Fig. 10. Cache hit ratio of different content types

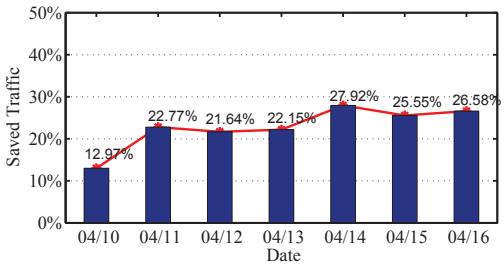

Fig. 11. Saved traffic traffic in different days

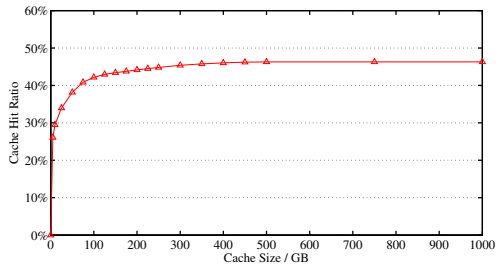

Fig. 12. Hit ratio over different cache size

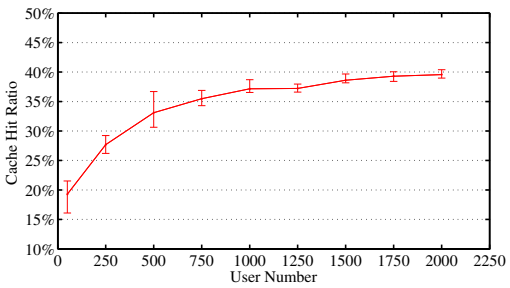

Fig. 13. Hit ratio over different user number

B. Saved Traffic based on the total three eNodeBs

Reducing backhaul traffic in the mobile traffic explosion era is one of the most important motivations of RAN caching. Figure 11 shows the daily saved traffic ratio of each day based on the aggregate of three eNodeBs, with 100 GB cache size and LFU replacement algorithm. Saved traffic ratio is defined as the ratio of the saved downstream traffic with the RAN cache, to the total HTTP downstream traffic, including both the cacheable and un-cacheable HTTP traffic, without RAN cache. As shown in Figure 11, except for a lower traffic reduction ratio on the first day during the warm-up phase, the average traffic reduction ratio on the other 6 days is 24.4% on average and has a maximum value of 27.9%.

C. Effect of Cache Size on Hit Ratio

In Figure 12, we plot the average hit ratio during the stable phase versus different cache sizes, ranging from 5 GB to 1000 GB. It is not surprising that the hit ratio increases with cache size. The maximum hit ratio reaches to 46.3% when the cache size is equal to or above 500 GB. When the cache size reaches 100 GB, the hit ratio is 42.2%, which is close to the maximum value. We also observe that when the cache size exceeds 100 GB, the increase in the hit ratio is not significant. We also observe that when the cache size is set to be 5 GB, 10 GB, 50 GB, the corresponding hit ratios are 26.1%, 29.5% and 38.2% respectively. Our observations indicate that even with a small cache size, the performance of RAN cache is obvious.

Note that the above calculation is based on the aggregate of the three eNodeBs with more than 3200 mobile users. Therefore, the cache size can be regarded as the upper bound value based on current LTE traffic. To sum up, we conclude that several hundreds of GB should be enough storage space for a single eNodeB cache in our measured LTE network.

D. Effect of User Number on Hit Ratio

A practical issue deserving consideration for industry is which eNodeBs are more suitable for deploying RAN caches. Obviously, the user number of an eNodeB is an important factor impacting the performance of the RAN cache. To this end, we investigate the relationship between the user number, ranging from 50 to 2000, and the hit ratio. Remember that we define "user number" as the number of active users in a day. For the user number of $M \in \{50, 250, 500, 750, 1000, 1250, 1500, 1750, 2000\}$, M users are randomly selected each day, and the average hit ratio and backhaul traffic reduction are computed during the stable stage over the entire measurement period. Then, the above statistical experiment are repeated 10 times, and the minimum, average and maximum values of hit ratio and reduction on backhaul traffic are plotted and shown in Figure 13 and 14.

Obviously, with increasing of user number, the cache hit ratio improves correspondingly. Even in the case of a smaller user number, 250 users, the cache hit ratio is stable, above 25%, and the average value is 27.3%. When the user number reaches to 1000, the cache hit ratios of the ten experiments is distributed around 37.2% with smaller oscillation.

E. Effect of User Number on Traffic Reduction

In Figure 14, we plot the saved traffic volume for different numbers of access users per day. Note that both the cache hit ratio and the saved traffic ratio are calculated based on the same sampled dataset as described in Section V-D.

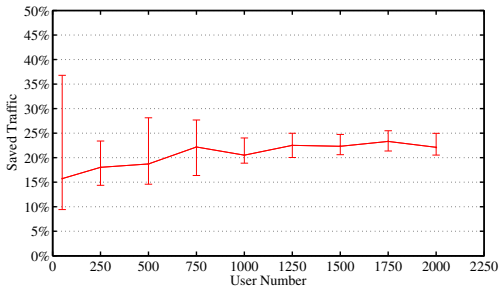

Fig. 14. Saved traffic over different user number

We observe that the average traffic reduction ratio improves much more slowly with the increase of user number, compared with an obvious increase of cache hit ratio in Figure 13. In fact, as illustrated in Figure 14, the average saved traffic ratio rises smoothly from 15.9% to 22.1% only when the user number is increased from 50 to 750. However, the average saved traffic fluctuates around 23.2% when there are more than 750 users. To some extent, the above observation is in accordance with the curve of cache hit ratio in Figure 13 where the cache hit ratio increases dramatically when the user number is no more than 750 and starts a smooth increase after the user number exceeds 750.

Another observation is that when the user number is equal or below 750, the performance of the saved traffic is more volatile. Specifically, when the user number is set to 50, the minimum saved traffic ratio is less than 10% while the maximum reaches as high as 37.1%. The reason behind this is that the randomness of mobile user behavior is amplified when the user number is very small. Once the user number is 1000 or more, the fluctuation of the saved traffic for a given user number appears to be lessened. In this case, the minimum value of the saved traffic ratio for is nearly 19% and the average values of different user numbers exceed 20%. This suggests that the performance of RAN cache in term of the saved traffic is more stable and good performance more likely when the user number of eNodeB is around 1000 or more.

VI. Implication on RAN Cache System Design

A. On Fundamental Design Issues

For a practical RAN cache system, there are several fundamental design issues for RAN cache system.

1) The first fundamental design issue is what type of content should be stored in the RAN cache.

Generally speaking, video and application content are preferred for caching in order to reduce the traffic pressure of mobile network and improve the user experience, since they make up the majority of the measured mobile traffic. However, it is worth noting that the images should not be neglected, if improving the latency perceived by mobile users becomes the primary goal of RAN cache system, rather than back-haul traffic reduction. The reason behind it is obvious when we observe that the requests for images account for 94.0% of total cacheable HTTP requests and the cache hit ratio of images is as high as 41.9% in the measured LTE network.

2) The second fundamental design issue is how large the storage size of the RAN cache should be.

As described in Section V-C, the hit ratio under a 100 GB cache size reaches 42.2% and approaches to a maximum value of 46.3% when the cache size is 500 GB. Although the traffic load of every eNodeB in the measured LTE network is light according to Section III-C, the hit ratio with a 100 GB cache size is computed based on the aggregated traffic of three eNodeBs which is generated by an average of more than 3200 users per day. Thus, we believe that a cache size of several hundreds of GB is enough for a single eNodeB of medium traffic load, and provides a good tradeoff between the performance and the cost of the RAN cache.

3) The third fundamental design issue is which eNodeBs are suitable for deploying RAN caches.

As we know, fully deploying RAN caching in the existing LTE network may be economically prohibitive due to the high cost of additional cache devices. In practice, how to choose the appropriate eNodeBs to deploy RAN caches is extremely important. According to Section V-D, the average cache hit ratio can reach 19.1% even though the average access user number per day is 50. However, in Section V-E, we observe that the corresponding saved traffic with 50 users is quite variable, and the minimum value is only 9.3%. Not until the user number reaches or exceeds 1000 does the RAN cache perform well in terms of both the hit ratio and the saved traffic. Since the access user number of eNodeB is an important factor when deciding which eNodeBs are good candidates for cache deployment, we suggest the user number of around 1000 per day can be used for reference in practical operations.

B. On the Collaborative Caching Issue

Collaborative caching in LTE mobile networks is an interesting topic for both academia and industry to improve the performance of RAN caching systems. Figure 15 illustrates a typical scenario of collaborative RAN caching where tens of cache-enabled eNodeBs in an area collaboratively cache popular contents. Here, we examine the potential gain of collaborative RAN caching based on our dataset. Specifically, we study the performance of collaborative caching among the eNodeBs with no more than 250 access users per day, due to their poor performance when caching independently. Based on the above experiments, we then investigate the role of collaboration in a cache system from a practical point of view.

In order to evaluate the performance of collaborative caching, we define an intuitive collaborative caching policy named One Copy Only (OCO). In OCO, all the individual eNodeBs are regarded as a single, large integrated cache node and each popular content is stored only once in the whole cache system, according to the overall popularity. The performance of OCO represents an upper bound of collaborative caching performance in terms of the overall hit ratio of caching system and the saved traffic of link l_0, as shown in Figure 15, especially when the transmission cost among eNodeBs are not taken into account. Obviously, the greater the saved traffic

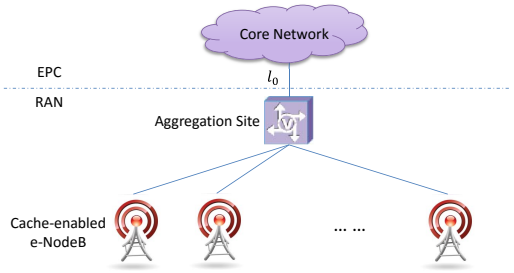

Fig. 15. Example of collaborative RAN caching

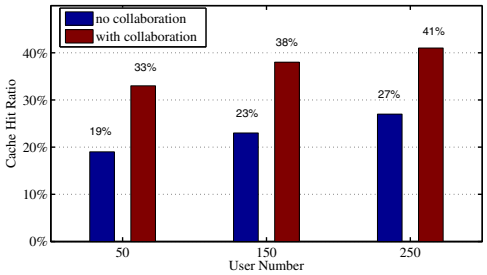

Fig. 16. Comparison of cache hit ratio over different user number

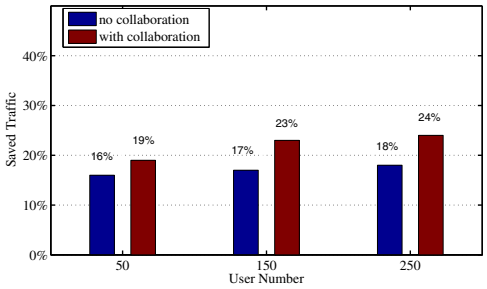

Fig. 17. Comparison of saved traffic over different user number

of link l_0, the less pressure of traffic imposed on the LTE core network. In the case of no collaboration, each eNodeB just implements the traditional LFU policy based on the local content popularity, for the purpose of comparison.

Assume that there are 10 eNodeBs or RAN cache nodes, each of which has a cache size of 100GB. Three experiments are carried out, and in each experiment, the number of users at each eNodeBs is set to be 50, 150 and 250, respectively. Each experiment is repeated 10 times. Thus, the average hit ratio and saved traffic ratio of no collaboration and collaboration in each experiment can be calculated respectively, as shown in Figure 16 and 17.

Figure 16 compares the overall cache hit ratio of caching system with and without collaboration. It is observed that the hit ratio can be significantly improved when applying the OCO collaborative caching policy. Compared with no collaboration, the cache hit ratio increases by about 73%, 65% and 57% when the user number is set to be 50, 100 and 250, respectively.

Figure 17 compares the traffic reduction on link l_0 with and without collaboration. Similarly, compared with no collaboration, we observe that the saved traffic ratio with collaboration increases by 18.7%, 35% and 33%, respectively, when the access user numbers per day per eNodeB are 50, 100 and 250. Note that the performance in terms of saved traffic can be further improved when the number of collaborative eNodeBs increases. For instance, when there are 50 eNodeBs, the saved traffic ratio through collaborative caching will reach 24%. It indicates that the saved traffic ratio increases by 50% compared with that of no collaboration.

From the results observed from Figure 16 and Figure 17, it is obvious that collaborative caching helps improve the performance of the cache system for eNodeBs or small cells when the user number is generally no more than 250.

Based on the above observation, there are two implications which may be valuable for practical LTE cache systems. Firstly, compared with deploying cache in a small cell, it is may be more effective to deploy the cache in some aggregation point of these small cells, e.g., an aggregation switch or Home eNodeB Gateway [14]. Secondly, when the number of users at cache-embedded eNodeBs are relatively small, designing an effective collaborative caching policy is critical for improving the overall performance of the cache system.

Furthermore, due to the scale of our dataset in terms of user number, the performance of collaborative caching for those eNodeBs with a large number of users remains to be investigated in the future.

VII. Related Works

Recently, there have been many prior works on the theoretical analysis and algorithm design of mobile caching in cellular backhaul links and core networks [15], [16], [17], in mobile heterogenous networks [18], [19], [20]. These works provide a solid foundation for the performance optimization of mobile caching. However, most of the results are obtained based on simulations and not on real network data.

Meanwhile, some research based on real traffic has been conducted to analyse the effectiveness of content caching at wired edge networks. A deployment of micro CDN technologies in wired ISP networks [21] is evaluated based on a large dataset collected over one and a half months of continuous online monitoring of links between the access and the backhaul network of Orange France. The gains are striking: with 100 GB memory on edge router line cards, covering 1200 users on average per day, the traffic in the access network can be reduced by 35%. Zink et al. [22] have conducted a measurement study of YouTube traffic in a large university campus network and shown that P2P-based distribution caching can reduce network traffic significantly.

As for mobile networks, Erman et al. [4] have explored HTTP caching in 3G cellular networks by using traffic traces generated by millions of users over a period of 36 hours. They report that 68.7% of the HTTP objects are cacheable, with 33% cache hit ratio in their measurements. Another work [23] finds that 36% of the 3G traffic is attributed to video streaming, and 24% of the bytes can be served from HTTP caches. By

monitoring 8.3 billion flows of 3G traffic for one week at a 10 Gbps backhaul link at one of the largest ISPs in South Korea, Shinae et al. [2] have comparatively analyzed the benefits and trade-offs of promising caching solutions. They have found that 53.9% of the total objects are cacheable, accounting for 40.7% of the downlink traffic, and Web caching can reduce download traffic up to 27.1%.

An analysis [11] based on a real LTE traffic gathered over a single day in North America reveals that 73.9% of the data volume and around 30.6% of the HTTP requests are cacheable. The above observation is similar with our results where the figures are 56.2% and 21.5%, respectively. Besides, the proportion of requests and traffic volume by content type is also basically consistent with our observations. However, it was observed that only 6.6% of the data volume can be saved with a web cache installed at Serving Gateway (S-GW) in [11]. Recall from Figure 9, the rather low ratio of saved traffic is not surprising, since the web cache may remain "warm-up" phase in the whole measurement period. Accordingly, we believe that a single day dataset is not enough to analyse the cache performance, especially for eNodeB-granularity measurement.

More importantly, compared with the above real traffic analysis mainly focusing on the performance of caching in mobile core network, we provide a thorough analysis for LTE eNodeB-granularity traffic, including traffic characteristics, content popularity and caching performance, from the perspective of RAN caching.

VIII. Conclusion

This paper provided evidence of the potential benefits of deploying RAN caches based on a real traffic collected from a commercial LTE network. Some interesting findings include that the hit ratio of RAN caching reaches 42.2% on average and that downstream traffic can be reduced by 24.4%, with 100 GB cache size and around 3200 mobile users. Moreover, we investigated the relationship between the performance of RAN caching and the number of users. Based on the above findings, we presented the implications on some fundamental design issues of practical RAN caching systems. These results may also help guide of future modeling and optimization work.

Acknowledgment

This work was supported by National Key Technology R&D Program under Grant No. 2015ZX03003004, National High Technology Research and Development Program (863 Program) under Grant No.2015AA01A705 and 2013AA013503-2, National Natural Science Foundation of China (Grant No. 61572497, 61572256, 61302031, 61302108 and 61174152).

References

[1] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update, 2014-2019," http://www.cisco.com/, 2015.

[2] S. Woo, E. Jeong, S. Park, J. Lee, S. Ihm, and K. Park, "Comparison of caching strategies in modern cellular backhaul networks," in *ACM 11th annual international conference on Mobile systems, applications, and services*, 2013, pp. 319–332.

[3] F. Malandrino, C. Casetti, and C.-F. Chiasserini, "Content discovery and caching in mobile networks with infrastructure," *IEEE Transactions on Computers*, vol. 61, no. 10, pp. 1507–1520, 2012.

[4] J. Erman, A. Gerber, M. T. Hajiaghayi, D. Pei, S. Sen, and O. Spatscheck, "To cache or not to cache: The 3G case," *IEEE Internet Computing*, vol. 15, no. 2, pp. 27–34, 2011.

[5] N. S. Networks, "Nokia siemens networks intelligent base stations," *White Paper*, 2012.

[6] Intel, "Smart cells revolutionize service delivery," *White Paper*, 2013.

[7] E. T. S. Institute, "Mobile edge computing," *White Paper*, 2014.

[8] F. Boccardi, R. W. Heath, A. Lozano, T. L. Marzetta, and P. Popovski, "Five disruptive technology directions for 5G," *IEEE Communications Magazine*, vol. 52, no. 2, pp. 74–80, 2014.

[9] M. Yang, Y. Li, D. Jin, L. Su, S. Ma, and L. Zeng, "Openran: a software-defined RAN architecture via virtualization," in *ACM SIGCOMM computer communication review*, vol. 43, no. 4. ACM, 2013, pp. 549–550.

[10] T. Wood, K. Ramakrishnan, J. Hwang, G. Liu, and W. Zhang, "Toward a software-based network: integrating software defined networking and network function virtualization," *Network, IEEE*, vol. 29, no. 3, pp. 36–41, 2015.

[11] B. A. Ramanan, L. M. Drabeck, M. Haner, N. Nithi, T. E. Klein, and C. Sawkar, "Cacheability analysis of HTTP traffic in an operational LTE network," in *IEEE Wireless Telecommunications Symposium (WTS)*, 2013, pp. 1–8.

[12] R. Fielding, J. Gettys, J. Mogul, H. Frystyk, L. Masinter, P. Leach, and T. Berners-Lee, "Hypertext transfer protocol–http/1.1," Tech. Rep., 1999.

[13] T. Strutz, "Data fitting and uncertainty," *A practical introduction to weighted least squares and beyond. Vieweg+ Teubner*, 2010.

[14] J. Robson, "Small cell backhaul requirements," *NGMN White Paper*, pp. 1–40, 2012.

[15] S. Ren, T. Lin, W. An, G. Zhang, D. Wu, L. N. Bhuyan, and Z. Xu, "Design and analysis of collaborative EPC and RAN caching for LTE mobile networks," *Computer Networks*, vol. 93, pp. 80–95, 2015.

[16] Z. Ming, M. Xu, and D. Wang, "InCan: In-network cache assisted enodeb caching mechanism in 4G LTE networks," *Computer Networks*, vol. 75, pp. 367–380, 2014.

[17] J. He, H. Zhang, B. Zhao, and S. Rangarajan, "A collaborative framework for in-network video caching in mobile networks," in *IEEE International Conference on Sensing, Communication, and Networking*, 2013.

[18] N. Golrezaei, K. Shanmugam, A. G. Dimakis, A. F. Molisch, and G. Caire, "Femtocaching: Wireless video content delivery through distributed caching helpers," in *IEEE International Conference on Computer Communications (INFOCOM)*, 2012, pp. 1107–1115.

[19] K. Poularakis, G. Iosifidis, A. Argyriou, and L. Tassiulas, "Video delivery over heterogeneous cellular networks: Optimizing cost and performance," in *INFOCOM, 2014 Proceedings IEEE*. IEEE, 2014, pp. 1078–1086.

[20] M. Dehghan, A. Seetharam, B. Jiang, T. He, T. Salonidis, J. Kurose, D. Towsley, and R. Sitaraman, "On the complexity of optimal routing and content caching in heterogeneous networks," *INFOCOM, Proceedings IEEE*, 2015.

[21] C. Imbrenda, L. Muscariello, and D. Rossi, "Analyzing cacheable traffic in isp access networks for micro cdn applications via content-centric networking," in *Proceedings of the 1st ACM international conference on Information-centric networking*, 2014, pp. 57–66.

[22] M. Zink, K. Suh, Y. Gu, and J. Kurose, "Characteristics of youtube network traffic at a campus network–measurements, models, and implications," *Elsevier Computer Networks*, vol. 53, no. 4, pp. 501–514, 2009.

[23] J. Erman, A. Gerber, K. Ramadrishnan, S. Sen, and O. Spatscheck, "Over the top video: the gorilla in cellular networks," in *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*, 2011, pp. 127–136.