Enabling Adaptive Data Prefetching in 5G Mobile Networks with Edge Caching

Chengchao Liang*, F. Richard Yu*, Ngoc Dao†, Gamini Senarath† and Hamid Farmanbar†
*Depart. of Systems & Computer Eng., Carleton University, Ottawa, ON, Canada
†Huawei Canada, Ottawa, ON, Canada

***Abstract*—The exponential growth of data traffic volume dominates the demand for the next generation mobile networks (5G). The consistent and satisfied quality of experience (QoE) is one of the leading challenges of provisioning services in 5G mobile networks. Thus, in this paper, we propose a novel adaptive prefetching scheme to compensate the undesired transmission conditions in the 5G mobile network by extending the content prefetching concept from the users to the network. Specifically, an optimization problem is proposed for a prefetching scheme that adaptively retrieves users data to access nodes and (or) user equipments (UEs) before the actual requests according to the network status, QoE status, predicted data rates, and mobility patterns of users. For the sake of tractability, the prefetching problem is transferred to a convex problem that can be solved efficiently. Accordingly, to implement the proposed schemes in the 5G network, system interactions among entities in the network are designed to realize prefetching-related functions. A signaling protocol to support the adaptive prefetching scheme is also presented. Simulation results show that an adaptive prefetching scheme can improve the network performance significantly.**

***Index Terms*—5G, adaptive prefetching, buffer management, load balance, quality of experience**

I. Introduction

The next generation mobile network, 5G [1], has been proposed to accommodate the increase in global mobile data traffic predicted to be nearly eightfold from 2015 to 2020. Significantly, 75 percent of them will be video [2]. Thus, the video streaming service is replacing voice and other applications to become the most significant service in mobile networks. To ensure that the customers remain satisfied with the video streaming services, mobile network operators have to provide a more consistent user experience and higher quality of experience (QoE) across the whole network [3], [4].

Unfortunately, even though the extremely high bandwidth can be provisioned in 5G by improving spectral efficiency and enhancing coverage, unpredictable channel conditions (e.g., fast moving velocity), uncovered areas (e.g., tunnels) and handovers may still bring drawbacks to the QoE of users. To compensate fluctuations of radio links and potential congestions of backhaul links, the prefetching technology has been introduced to mobile networks [5]–[7]. Considering that costs of the storage and transmission are less nowadays, prefetching data can be considered one of the most efficient ways to maintain the consistency of users' experience.

Based on the location of the prefetched data, prefetching in mobile networks can be classified into two main categories: *user equipment (UE) prefetching* and *access node (AN) prefetching*. The study in [5] proposes that the prefetching support should be provided by the underlying mobile system. The proposed research aims at minimizing delay with constraints of battery lifetime, data usage and network status. The authors in [7] study this topic from the view of the content. They prefetch data to the users based on a recommendation system, which can be considered as an extension of in-network caching. [8] proposes to prefetch data to users based on mobile online social networks to reduce the access delay and enhance the satisfaction of mobile users. However, these works do not consider the mobility of the user and prefetching data at ANs. To fill the lack of AN prefetching, by predicting users' next associated AN, [6] proposes to prefetching packets to possible ANs so that the handover delay can be reduced. [9] also proposes a prefetching scheme of ANs, but focuses more on the integration of video rate adaptive.

Although those studies have been used to realize and optimize the prefetching, a joint and adaptive prefetching scheme integrating UE and AN is lack of sufficient research. To fill this gap in the mobile network is not straightforward by just combining those two prefetching schemes because the network should dynamically select the location to prefetch the data (e.g., ANs or UEs), and the volume of data that it can (or needs to) prefetch. Thus in this article, we present a novel framework and scheme that can support adaptive prefetching to guarantee the QoE and improve the utility. The main features of our proposed works are summarized as follows:

- A novel prefetching scheme enabling adaptive UE and AN prefetching is proposed in this paper. The proposed mechanism takes the current and future status of the network, such as user mobility, network load, network costs, current buffer status and QoE of the user into consideration.
- A potential solution of optimizing prefetching is proposed with constraints regarding physical resource limitation and QoE requirements. The problem is formulated as a convex problem that can be solved efficiently.
- To realize a network-aware prefetching, new functionalities are required. Thus, a novel architecture, including necessary network functions (elements) and corresponding operations (call flows), is proposed.

The rest of this paper is organized as follows. Section II introduces the system model of the presented problem. Section III formulates the presented problem and describes general architecture and operations of the proposed adaptive prefetching. Simulation results are discussed in Section IV. Finally, we conclude this study in Section V.

II. System Model

To realize adaptive prefetching at UE and ANs, we assume that two types of caches are enabled at the UE and ANs separately for each application traffic flow, such as protocol data unit (PDU) session in 5G [10]. For any user served by the network, the first type of cache can be the regular buffer assigned to the application and is used to cache content at the device. When the channel is predicted to be worse than required data rate or an outage is possible, the network may request the UE to prefetch more data. The second cache is used for ANs to prefetch more data to the RAN when backhaul is underloaded or a low-cost AN is predicted to serve the user.

In this paper, we consider the downlink transmission case in a cellular network comprised of a set $\mathcal{J}$ prefetching-enabled access nodes (ANs) and a set $i \in \mathcal{I}$ of users. If user i is served by the AN j, an association indicator $a_{ij} = 1$; otherwise $a_{ij} = 0$. Those ANs connect to the core network (CN) through wired backhaul links with fixed capacities L_j (e.g., bps). The total bandwidth of spectrum used for associated users is W_j of AN j. Each user can be served by one or multiple ANs [11]–[13] with allocated radio access data rate r_i. Obviously, r_i depends on the spectrum bandwidth (number of sub-carriers if OFDMA system is considered) assigned by the corresponding AN j as well as the spectrum efficiency (SE) $\gamma_{ij}, j \in \mathcal{J}$ of the link. Without considering small-scale fading (e.g., frequency-selective fading), by using Shannon bound, the average UE i's SE γ_{ij} (bps/Hz) over the scheduling interval is calculated as $\gamma_{ij} = \log\left(1 + \frac{g_{ij}p_{ij}}{\sigma_0}\right)$, where g_{ij} is the large-scale channel gain that includes pathloss and shadowing between the transmission AN j and the receiving UE i. p_{ij} (Watt/Hz) is the normalized transmission power on the link between i and j. The fixed equal power allocation mechanism is used, which means transmission power p_{ij} is the same for all frequencies. σ_0 is the power spectrum density of additive white Gaussian noise.

According to [1], each user can have multiple data flows (PDU sessions) for different applications. Thus, we assume that total a set $\mathcal{F}_i$ of flows are running at the same time at user i. We denote the radio access and backhaul bandwidth for each flow f_i as r_{f_i} (bps) and l_{f_i} (bps), respectively. Moreover, the caches allocated for flow f_i at UE i and AN j are denoted by b_{f_i} and c_{f_i}. The maximum cache spaces of UE i and AN j are B_i (bits) and C_j (bits), respectively.

The fact of exhausted caches can lead a frozen transmission from AN to UE or a frozen streaming at UE. Thus, to avoid this annoy situation at any time, both caches for buffering flow f_i should be larger than thresholds $\bar{b}_{f_i}$ and $\bar{c}_{f_i}$. However, to reach this purpose, the network needs to predict the possible radio access and backhaul bandwidth of each flow. Let us denote the scheduled bandwidths from current time t to $t+T$ as r_{f_i} and l_{f_i}, respectively. Similary, the predicted scheduled bandwidths from current time $t+T$ to $t+2T$ as $r^T_{f_i}$ and $l^T_{f_i}$. Obviously, the remaining cached data of UE cache b_{f_i} and AN cache c_{f_i} for flow f_i after $2T$ should hold the contstraints that b_{f_i} and c_{f_i} are larger than $\bar{b}_{f_i}$ and $\bar{c}_{f_i}$.

III. Adaptive Prefetching of Mobile Data in 5G Networks

In this section, the proposed adaptive prefetching scheme is presented. Firstly, the optimization problem is formed and transformed into a convex problem. The implementation of the proposed scheme is described at last.

A. Problem Formulation

The prefetching data problem can be considered as a scheme of allocating resource in advance so that we can model the prefetching as a network resource allocation problem.

1) Variables: In the next generation mobile network, 5G, heterogeneous networks are proposed to support the high demand of coverage and bandwidth, so that it is important to select an appropriate node. Similar to existing studies [12], [14], in this paper, besides $a_{ij} \in \{0,1\}$, we use $a^T_{ij} \in \{0,1\}$ to indicate the predicted association relationship after T seconds. $a^T_{ij} = 1$ means UE i is served by AN j at the time $t+T$ while $a^t_{ij} = 0$ means the opposite. To maintain the caches in a certain level, the network should schedule the bandwidths of both radio access links and backhaul links appropriately. Thus, as mentioned before, two types of variables are defined in this paper, which are r_{f_i} and l_{f_i} as well as $r^T_{f_i}$ and $l^T_{f_i}$.

2) Objective: In this paper, we deploy a utility function based on the proportional fairness (PF) scheme [12]. The utility is the aggregated log-based utility of the achieved rate of each user shown as the following equation.

$$U = \sum_{\substack{j \in \mathcal{J} \\ i \in \mathcal{I}}} a_{ij} \log\left(\sum_{f_i \in \mathcal{F}_i} r_{f_i}\right) + \sum_{\substack{j \in \mathcal{J} \\ i \in \mathcal{I}}} a^T_{ij} \log\left(\sum_{f_i \in \mathcal{F}_i} r^T_{f_i}\right), \tag{1}$$

where the second part is based on the predicted network information.

3) Constraints: To satisfy the prefetching request, the remaining cached data of UE and AN for flow f_i after $2T$ should be larger than $\bar{b}_{f_i}$ and $\bar{c}_{f_i}$, shown as follows:

$$\begin{aligned} &\underbrace{b^0_{f_i}}_{\text{existing data}} + \overbrace{\left(\sum_{j \in \mathcal{J}} a_{ij} r_{f_i} T + a^T_{ij} r^T_{f_i} T\right)}^{\text{downloaded data}} \\ &- \underbrace{D_{f_i}}_{\text{consumed data}} \geq \bar{b}_{f_i}, \forall f_i, \end{aligned} \tag{2}$$

and

$$\begin{aligned} &\underbrace{c^0_{f_i}}_{\text{existing data}} + \overbrace{a_{ij} l_{f_i} T + a^T_{ij} l^T_{f_i} T}^{\text{downloaded data}} \\ &- \underbrace{a_{ij} r_{f_i} T + a^T_{ij} r^T_{f_i} T}_{\text{transmitted data}} \geq a^T_{ij} \bar{c}_{f_i}, \forall f_i, j, \end{aligned} \tag{3}$$

where D_{f_i} is the consumed data (e.g., played video) and T (seconds) is the scheduling interval. Moreover, as all cached data are limited by the storage space, following constraints should hold:

$$\sum_{f_i\in\mathcal{F}_i}\left(b^0_{f_i}+\sum_{j\in\mathcal{J}}a_{ij}r_{f_i}T+\sum_{j\in\mathcal{J}}a^T_{ij}r^T_{f_i}T-D_{f_i}\right)\le B_i,\forall i, \tag{4}$$

and

$$\sum_{\substack{i\in\mathcal{I}\\ f_i\in\mathcal{F}_i}}\left(c^0_{f_i}+a_{ij}l_{f_i}T+a^T_{ij}l^T_{f_i}T-a_{ij}r_{f_i}T+a^T_{ij}r^T_{f_i}T\right)\le C_j,\forall j, \tag{5}$$

where $\mathcal{I}$ means the set of users that will be associated to AN J after two scheduling interval $2T$ based on the prediction of the network status.

Besides (2)-(5), following constraints should hold due to physical resources limitations. Firstly, the scheduled radio link bandwidth should not exceed the capacity of the feasible spectrum, which can be formulated as

$$\sum_{i\in\mathcal{I}}\frac{\sum_{f_i\in\mathcal{F}_i}a_{ij}r_{f_i}}{\gamma_{ij}}\le W_j,\forall j, \tag{6}$$

and

$$\sum_{i\in\mathcal{I}}\frac{\sum_{f_i\in\mathcal{F}_i}a^T_{ij}r^T_{f_i}}{\gamma^T_{ij}}\le W_j,\forall j. \tag{7}$$

It should be noted that γ_{ij} is based on current observation of the channel status, but γ^T_{ij} only is based on the prediction of the channel status. Apparently, the predictions are not exactly accurate. However, since the prefetching scheduling is much slower than the varying of channels, we assume the prediction of average channel status is accurate enough. Similarity applied to the backhaul,

$$\sum_{i\in\mathcal{I}}\sum_{f_i\in\mathcal{F}_i}a_{ij}l_{f_i}\le L_j,\forall j, \tag{8}$$

and

$$\sum_{i\in\mathcal{I}}\sum_{f_i\in\mathcal{F}_i}a^T_{ij}l^T_{f_i}\le L_j,\forall j. \tag{9}$$

Thus, given the objective function, variables and constraints, we can define the adaptive prefetching problem **P0** as follows:

$$\begin{aligned}
\max\quad & U\\
s.t.\quad & C1: && r_{f_i},r^T_{f_i},l_{f_i},l^T_{f_i}\in\mathbb{R}^+,\forall f_i,\\
& && a_{ij},a^T_{ij}\in\{0,1\},\forall i,j\\
& C2: && \sum_{j\in(J)}a_{ij}\le 1,\forall i\\
& && \sum_{j\in(J)}a^T_{ij}\le 1,\forall i\\
& C3: && (2)-(3),\quad C4:\quad (4)-(9).
\end{aligned} \tag{10}$$

Constraints C1 and C2 are domains of variables. Constraints C2 limit the associated AN to be only one. Constraints C3 are our prefetching requirements and constraints C4 are resource limitations. Unfortunately, the mix integer variables result in the problem a mix-integer non-convex problem (MINLP) that generally is NP-hard and intractable in scheduling [15].

B. Problem Transformation

To make problem (10) tractable, some transformations are needed to approximate it to a convex problem. Following the approach in [12] and [16], we relax a_{ij} and a^T_{ij} in (10) to be real value variables that $a_{ij}\in[0,1]$ and $a^T_{ij}\in[0,1]$. Relaxed a_{ij} and a^T_{ij} can be interpreted as the time sharing factors that represent the ratios of time when UE i associates to BS j [16]. Moreover, the multiple-association scheme is realized by the packet duplication [13], which means one UE can be served by multiple ANs to enhance the reliability.

To overcome the multiplication relationship between association indicators and links bandwidths, we make following variable substitutions shown as $\tilde{r}_{f_ij}=a_{ij}r_{f_i}$, $\tilde{r}^T_{f_ij}=a_{ij}r^T_{f_i}$, $\tilde{l}_{f_ij}=a_{ij}l_{f_i}$ and $\tilde{l}^T_{f_ij}=a_{ij}l^T_{f_i}$. To avoid non-applicable results, let $a_{ij}\log\left(\sum_{f_i\in\mathcal{F}_i}\frac{\tilde{r}_{f_ij}}{a_{ij}}\right)=0$ when $a_{ij}=0$ and $a^T_{ij}\log\left(\sum_{f_i\in\mathcal{F}_i}\frac{\tilde{r}^T_{f_ij}}{a^T_{ij}}\right)=0$ when $a^T_{ij}=0$. The reformed objective function is then presented as

$$\tilde{U}=\sum_{\substack{j\in\mathcal{J}\\ i\in\mathcal{I}_j}}\left[a_{ij}\log\left(\sum_{f_i\in\mathcal{F}_i}\frac{\tilde{r}_{f_ij}}{a_{ij}}\right)+a^T_{ij}\log\left(\sum_{f_i\in\mathcal{F}_i}\frac{\tilde{r}^T_{f_ij}}{a^T_{ij}}\right)\right]. \tag{11}$$

(11) is jointly convex of $\tilde{r}_{f_ij}$, $\tilde{r}^T_{f_ij}$, a_{ij} and a^T_{ij}, because $\tilde{U}$ is the well-known perspective function [17] of $\tilde{r}_{f_ij}$, $\tilde{r}^T_{f_ij}$, a_{ij} and a^T_{ij} [1]. Due to the limited space, the detailed substitutions applied to (2) to (5) and (6) to (9) are not presented. The relaxed problem of problem (10) can then be shown as

$$\begin{aligned}
\mathbf{P0}:\max\quad & \tilde{U}\\
s.t.\quad & \tilde{C}1: && \tilde{r}_{f_ij},\tilde{r}^T_{f_ij},\tilde{l}_{f_ij},\tilde{l}^T_{f_ij}\in\mathbb{R}^+,\forall f_i,\\
& && a_{ij},a^T_{ij}\in[0,1],\forall i,j\\
& \tilde{C}2: && \sum_{j\in(J)}a_{ij}\le 1,\forall i\\
& && \sum_{j\in(J)}a^T_{ij}\le 1,\forall i\\
& \tilde{C}3: && (\tilde{2}),(\tilde{3})\\
& \tilde{C}4: && (\tilde{4}),(\tilde{5})(\tilde{6}),(\tilde{7}),(\tilde{8}),(\tilde{9})
\end{aligned} \tag{12}$$

As the objective function is a convex function and all constraints are linear, the problem (12) is a convex problem that can be solved by a lot of efficient methods (e.g., interior point method). In this paper, CVX [19] is used to solve the problem to get the adaptive prefetching solution.

C. Implementation of the Proposed Adaptive Prefetching

As we mentioned in above subsection, the adaptive prefetching based on the current and future channel conditions of users and the status of the network. Thus, in order to predict

[1] $\tilde{U}$ is a continuous function, a similar proof can be found at [18]

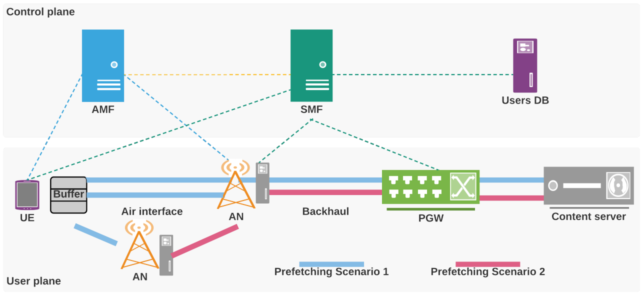


Fig. 1: Adaptive prefetching architecture.

the possible status of UE mobility, the 5G network needs to be equipped with functions that can predict mobility patterns (e.g., locations and handovers).In the control plan of 5G architecture [1], Access and Mobility Management Function (AMF) has the mobility context of users.Thus, AMF can obtain potential mobility patterns (e.g., locations or areas) of users; this information will be delivered to another entity called Session Management Function (SMF).

SMF is a core network (CN) function entity as well that manages and monitors services (protocol data unit sessions) of the network [1]. To improve the network performance, the SMF is requested to be able to handle a QoE-aware report from UEs [3]. For example, the SMF interacts with UE and content servers to coordinate the video segments transmission. In the adaptive prefetching, the SMF tries to optimize the QoE of UEs by taking consideration of the QoE report from UEs and the predicted network information from the AMF. Thus, the SMF takes responsibilities of several key functions related to adaptive prefetching, such as the QoE optimization, sending the request, and coordinating network elements in the UP of both the RAN and CN. The SMF needs to evaluate if prefetching is necessary or not by solving the problem (12). If the prefetching is active, the SMF also has to decide where to prefetch and how many data should be prefetched. Moreover, the users data base (DB) stores historical data that are used to assist the prediction and evaluation. The predictions and corresponding applications are understudying by both industrial and academic research [20], [21].

To prefetch data adaptively at UEs or edge ANs, both UE, ANs, and packet-gateway (PGW) should be prefetching-enabled. Specifically, if prefetching is active wherever to prefetching, a higher data rate is necessary for this specific traffic flow along with both the backhaul and optional (user prefetching) air interface. Thus, QoS requirements of this specific flow should be able to be modified dynamically. Moreover, since data will be downloaded at UEs and ANs, buffers at UEs and BSs should allow longer holding time. To make a better performance, prefetching requests a large buffer size, which is available currently due to the lower price of storage and the emerging wireless edge caching.

Fig. 1 shows the main logical entities involved in the adaptive prefetching. From a network view, thanks to the user-centric concept in 5G [22], [23], data can be prefetched not only at the serving AN but also a potential serving AN. SMF can set dynamically a user plane (UP) path to where the prefetching is needed. As shown in Fig. 1, Scenario 1 is a user prefetching case where the data are delivered from a content server to the buffer of the UE directly. Scenario 2 is a typical AN prefetching case where data firstly are pre-downloaded at the AN then delivered to the UE when the UE is served by the corresponding AN.

To realize adaptive prefetching, three phases of operations need to be performed by the network and UEs, which are 1) *prefetching preparation*, 2) *prefetching request*, and 3) *prefetching*. Moreover, a well-designed signaling protocol is another critical component. As shown in Fig. 2, we design a call flow that is used for UE prefetching and AN prefetching. It should be noted that this is just an instance and more signaling protocols are necessary for specific different services. Here, we use Fig. 2 to illustrate the three phases of prefetching operations. Steps 1 and 2 are the regular downloaded procedure.

- *Preparation*: Steps 3 to 5 show the necessary signaling exchanges between UE and other network elements to acquire prefetching information. At step 3, both the ANs and backhaul also need to send the network status (e.g., traffic load and radio links capacity) to the SMF for the further prediction. The QoE report is sent to SMF while AMF predicts the location of the user in the next server time periods (e.g., seconds to minutes) by using information from the UEs and ANs and report them to the SMF. The SMF calculates the future data rate of the backhaul and radio links that can be allocated to users. In this phase, SMF has to handle application-aware QoE reports including buffer status from UEs. Based on the information from QoE reports and future data rate report, the SMF makes decisions on the future achievable bandwidth for UEs, potential prefetching locations and recommends prefetching size (data size or video length).
- *Request*: Steps 6-9 and 12-16 show the request phase of prefetching. In this phase, the SMF firstly sends a request and corresponding information to UEs who may need prefetching. If the UE accepts to prefetch data, it sends Acknowledgement (ACK) back to the SMF; otherwise, negative-Acknowledgement (NACK) is sent. If the SMF receives ACK, it needs to inform the related network elements to modify the QoS of flows that will carry the prefetching data (steps 8-9). Once AN prefetching is active, the selected AN needs to allocate buffer resource to corresponding flows and holds them for the longer time if necessary. If other ANs are involved, new UP paths may need to be set up. It should be noted that to avoid exhaust radio resources, the packets stalling at the buffer used for prefetching may be set a different priority rather than regular packets (steps 12-16).
- *Prefetching*: In the prefetching phase (steps 10-11 and 17-19), the application at the UE sends requests of extra data. According to the decision made by SMF,

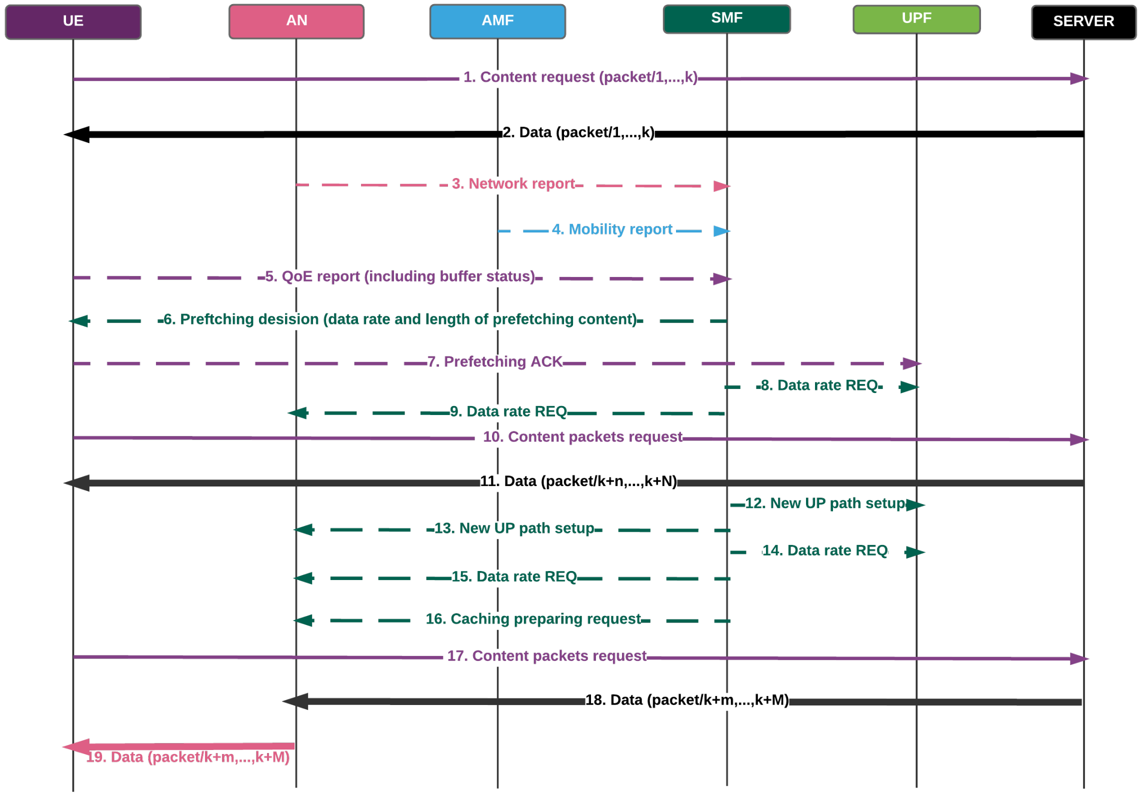

UE
AN
AMF
SMF
UPF
SERVER
1. Content request (packet/1,...,k)
2. Data (packet/1,...,k)
3. Network report
4. Mobility report
5. QoE report (including buffer status)
6. Preftching desision (data rate and length of prefetching content)
7. Prefetching ACK
8. Data rate REQ
9. Data rate REQ
10. Content packets request
11. Data (packet/k+n,...,k+N)
12. New UP path setup
13. New UP path setup
14. Data rate REQ
15. Data rate REQ
16. Caching preparing request
17. Content packets request
18. Data (packet/k+m,...,k+M)
19. Data (packet/k+m,...,k+M)

Fig. 2: Message exchange for supporting the adaptive prefetching.

the network should assist the involved ANs to prefetch the data. Taking the DASH service as an example, the content application sends video segments requests to the content server. It should be noted the prefetching phase is independent with the ANs, which means that the AN is transparent to this phase. Technically, this is just a regular data request procedure.

As the support of SDN and network virtualization in the CN and the RAN of the next generation mobile network has been proposed, those function entities may just be software applications that are located in one controller. Thus overheads between those entities can be ignored to a large extent. Moreover, the softwarization of the CN enables that signaling protocols of other services can be easily obtained based on modifications of the given example.

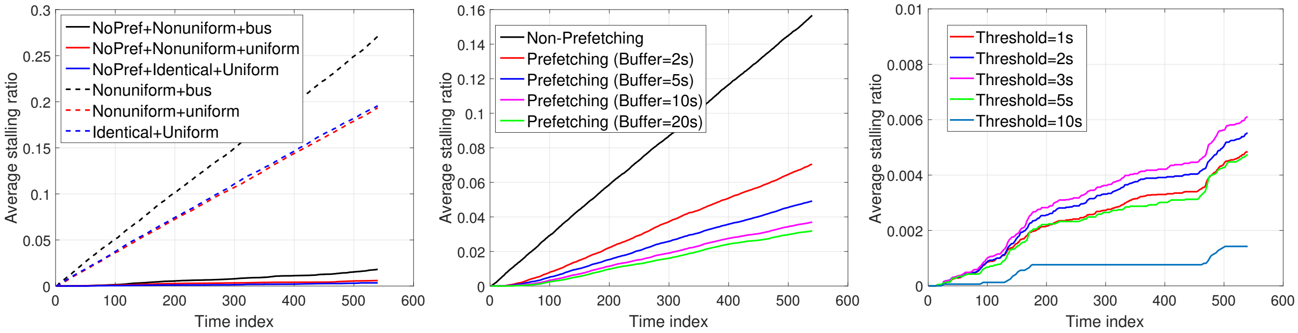
TABLE I: Simulation parameters

Number of users	120
Number of BSs	10
BS distribution	1-D Line, randomly
Minimum Inter-BS distance	250 m
Average Inter-BS distance	500m
Spectrum bandwidth	10 Mhz
Transmission power	46 dBm
Backhaul capacity	50 Mpbs
Users distribution	uniform distribution
Mobility	average speed = 15 m/s
Average video rate	2 Mpbs
Video rate distribution	Gaussian
Simulation Length	1000 s
Starting play buffer	1 s
Predicted cycle length	[2,3,5,10] s
Maximum prefeching buffer of users	[2,5,10,20] s
Maximum prefeching buffer of BSs	[2,5,10,20] s

IV. Simulation Results and Discussions

Simulations are conducted to evaluate the performance of the proposed adaptive prefetching scheme. A mobile network with 120 users and 10 ANs is considered. ANs are located along with a road at where users move with the average velocity of 55 kmph. The video consuming bandwidth ranges from $[1, 3]$ Mbps. The remaining simulation setups are listed in Table I. Three scenarios of different distributions of UEs and ANs are considered. In D1, equal AN spacing and uniform UE distribution are considered. In D2 deploys both uniform UE and AN distribution. We group UEs to several small groups (e.g., a bus) in D3 where the uniform AN distribution is assumed. Obviously, D3 is the worst case, as traffic may boost according to the clustered users.

Fig. 3a illustrates simulation results for the 3 scenarios. As the plots shown, simulation results for each of the three distributions with no pre-fetching (NP) have a higher average stalling ratio compared to the results with pre-fetching. Therefore, pre-fetching data is effective to improve the network performance. As shown in Fig. 3b, in this experience, the



(a) Different deployment scenarios

(b) Different buffer sizes of ANs.

(c) Different threshold.

Fig. 3: Average stalling ratio of different schemes

effects from buffer size of both caches of UE and AN are evaluated. With the increase of cache size, the average stalling ration of the UEs decreases significantly. It should be noted that the performance shown by buffer size of 10 seconds is similar to its peer of buffer size of 20 seconds, which means that increasing the cache size of ANs cannot always enhance the performance.

The minimum requirement of caching size is the main parameter that affects our proposed scheme. Thus, in Fig. 3c, we test the influence of the threshold $\bar{b}_{f_i}$. Shown in the figure, when the system requests the UE to maintain at least 10 seconds content, the stalling ration is approaching zero.

V. Conclusions

In this paper, an adaptive prefetching scheme has been introduced to improve the network performance. To realize this adaptive scheme, a system and methods were proposed to retrieve users data to ANs and (or) UEs before actual requests are made, where the corresponding architecture and operations were introduced. Simulation results have shown that the proposed scheme enhances the experience of UEs. Our future works are considering the enhancement brought by multiple access edge computing.

References

[1] 3GPP TS 23.501 V1.1.0, "Technical specification system architecture for the 5G system," July 2017, http://www.3gpp.org, accessed Jul. 20, 2017.

[2] Cisco, "Cisco visual networking index: Global mobile data traffic forecast update 2015–2020," Tech. Rep., Feb. 2016.

[3] 3GPP TR 26.909 V2.0.0, "Technical specification group services and system aspects; study on improved streaming quality of experience (QoE) reporting in 3GPP services and networks," Mar. 2017, http://www.3gpp.org, accessed Mar. 30, 2017.

[4] A. Jassal and C. Leung, "H. 265 video capacity over beyond-4g networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2016, pp. 1–6.

[5] B. D. Higgins, J. Flinn, T. J. Giuli, B. Noble, C. Peplin, and D. Watson, "Informed mobile prefetching," in *Proceedings of the 10th international conference on Mobile systems, applications, and services*, 2012, pp. 155–168.

[6] J. Wen and V. O. Li, "Data prefetching to reduce delay in software-defined cellular networks," in *Proc. IEEE Symp. Personal, Indoor & Mobile Radio Commun. (PIMRC)*, Aug. 2015, pp. 1845–1849.

[7] S. Wilk, D. Schreiber, D. Stohr, and W. Effelsberg, "On the effectiveness of video prefetching relying on recommender systems for mobile devices," in *Proc. IEEE Int. Consumer Commun. & Net. Conf. (CCNC)*, Jan. 2016, pp. 429–434.

[8] C. Wu, X. Chen, Y. Zhou, N. Li, X. Fu, and Y. Zhang, "Spice: Socially-driven learning-based mobile media prefetching," in *Proc. IEEE INFOCOM*, Apr. 2016, pp. 1–9.

[9] H. A. Pedersen and S. Dey, "Enhancing mobile video capacity and quality using rate adaptation, RAN caching and processing," *ACM/IEEE Trans. Networking*, vol. 24, no. 2, pp. 996–1010, May 2016.

[10] 3GPP, "Technical Specification Group Services and System Aspects; Network Sharing;Architecture and functional description," 3rd Generation Partnership Project (3GPP), TS 23.251 V11.5.0, Mar. 2013. [Online]. Available: http://www.3gpp.org/ftp/Specs/html-info/23251.htm

[11] K. Shen, Y.-F. Liu, D. Y. Ding, and W. Yu, "Flexible multiple base station association and activation for downlink heterogeneous networks," *IEEE Signal Processing Letters*, vol. 24, no. 10, pp. 1498–1502, Oct. 2017.

[12] Q. Ye, B. Rong, Y. Chen, M. Al-Shalash, C. Caramanis, and J. G. Andrews, "User association for load balancing in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 6, pp. 2706–2716, June 2013.

[13] 3GPP TS 38.300 V1.0.0 1 , "Technical specification group radio access network; nr; nr and ng-ran overall description stage 2 (release 15)," Sept. 2017, http://www.3gpp.org, accessed Oct. 13, 2017.

[14] C. Liang and F. R. Yu, "Virtual resource allocation in information-centric wireless virtual networks," in *Proc. IEEE Int. Conf. Commun. (ICC)*, June 2015, pp. 3915–3920.

[15] G. Li and H. Liu, "Downlink radio resource allocation for multi-cell OFDMA system," *IEEE Trans. Wireless Commun.*, vol. 5, no. 12, pp. 3451–3459, Dec. 2006.

[16] D. Fooladivanda and C. Rosenberg, "Joint resource allocation and user association for heterogeneous wireless cellular networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 248–257, Jan. 2013.

[17] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge university press, 2009.

[18] C. Liang, F. R. Yu, and X. Zhang, "Information-centric network function virtualization over 5G mobile wireless networks," *Network, IEEE*, vol. 29, no. 3, pp. 68–74, Mar. 2015.

[19] M. Grant and S. Boyd. (2014, Mar.) CVX: Matlab software for disciplined convex programming, version 2.1. [Online]. Available: http://cvxr.com/cvx

[20] A. Parate, M. Böhmer, D. Chu, D. Ganesan, and B. M. Marlin, "Practical prediction and prefetch for faster access to applications on mobile phones," in *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, 2013, pp. 275–284.

[21] Y. Zhang, "User mobility from the view of cellular data networks," in *Proc. IEEE INFOCOM*, May 2014, pp. 1348–1356.

[22] K. V. Katsaros, V. Sourlas, I. Psaras, S. Rene, and G. Pavlou, "Information-centric connectivity," *IEEE Comm. Mag.*, vol. 54, no. 9, pp. 50–57, Sept. 2016.

[23] C. Liang and F. R. Yu, "Bandwidth provisioning in cache-enabled software-defined mobile networks: a robust optimization approach," in *Proc. IEEE Veh. Tech. Conf. (VTC) - Fall*, Sept. 2016.