Cooperative and Weighted Proportional Cache Allocation for EPC and C-RAN

Anselme Ndikumana, Choong Seon Hong*

Department of Computer Science and Engineering, Kyung Hee University, Rep. of Korea
{anselme, cshong}@khu.ac.kr

Abstract

The demand for multimedia services in mobile networks continues to increase rapidly, and this increase is expected to continue in the coming years. However, the existing mobile network cannot practically deal with this explosive growth of mobile data traffic. To address this challenge, caching in Cloud Radio Access Network (C-RAN) has been proposed, where the contents can be cached at Base Band Unit (BBU) pools, and Remote Radio Heads (RRHs). However, there is no cooperation among these caches, which causes backhaul and fronthoul links to continue to deal with huge mobile data traffic. Furthermore, caching at RRH allows the low latency content delivery, but caching storage at RRH is usually small, and this results in lower cache hit ratio. Therefore, to address this challenge, a cooperative caching that considers hierarchical caching at Evolved Packet Core (EPC), BBU pool, RRH is needed. In this paper, we propose weighted proportional cache allocation that allows cooperation among caches, reduce delay, and improves cache hits, where Mobile Network Operator (MNO) provides shared cache allocation to multiple Content Providers (CPs) in three layers (EPC, BBU pool, RRH) hierarchical caching architecture. We show that our proposal is easy to implement in production network.

1. Introduction

In mobile networks, the demand for multimedia data and services continue to increase. This enormous extend was not caused only by the increase of mobile devices, but also the growing number of applications and services. However, the existing mobile networks cannot practically handle this explosive growth of demands. Therefore, Cloud Radio Access Network (C-RAN), which consists of centralized base band processing unit, known as Base Band Unit (BBU) pool, and distributed Remote Radio Heads (RRHs) has been introduced for improving 5G network performance. C-RAN helps Mobile Network Operators (MNOs) to address the number of challenges that they are facing, while trying to support this explosive growth mobile data traffic [1]. In additional to that, caching in C-RAN towards to the edge of network at RRHs, which is closer to mobile users contributes simultaneously in reducing mobile data traffic and improving quality of experience. However, caching at network edges are usually small and result in low cache hit ratio [2].

Currently, in the proposed solutions [3-4], caching functions mostly take place within the EPC, C-RAN, RRHs, but there is no cooperation among these caches. To address this issue, in [3], authors proposed a novel cooperative hierarchical caching framework in C-RAN, where contents are jointly cached at the BBU and at RRHs in order to minimize content access delay, and reduce backhaul traffic load.

Based on above cooperative hierarchical caching proposal, in this paper, we propose weighted proportional cache allocation that improves cooperation among caches, where MNO monetizes its cache storages, and provides shared cache allocation to multiple CPs in three layers (EPC, BBU, RRH) hierarchical caching architecture.

The rest of the paper is organized as follows, Section 2 presents our system model. Section 3 described in details weighted proportional cache allocation, while the Section 4 presents pricing framework. Section 5 presents our performance evaluation. We conclude the paper and give future directions in Section 6.

2. System Model

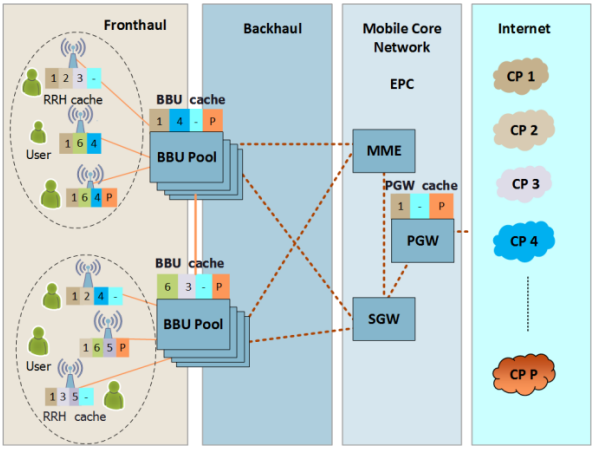

Fig. 1. System Model

Let us consider $V = \{1, \ldots, V\}$ as a set of consumers, which are served by $N = \{1, \ldots, N\}$ cache-enabled RRHs. As described in Fig. 1, we assume that each mobile user $v \in V$ is connected to its nearest RRH $n \in N$, and can request one content $\tau \in T$ at each time slot, where T is considered as a content catalogue. Each RRH $n \in N$ is connected to cache-

enabled BBU pool, through the use of capacity constrained fiber fronthoul link. We consider $B = \{1,\dots,B\}$ as a set of BBU pools, where BBU pools are connected among themselves through fiber links. Each BBU pool is connected to cache-enabled EPC (cache storage is attached to PGW) denoted E through the use of backhaul link. Each EPC is connected to one or more CPs. We denote $P = \{1,\dots,P\}$ as a set of CP servers.

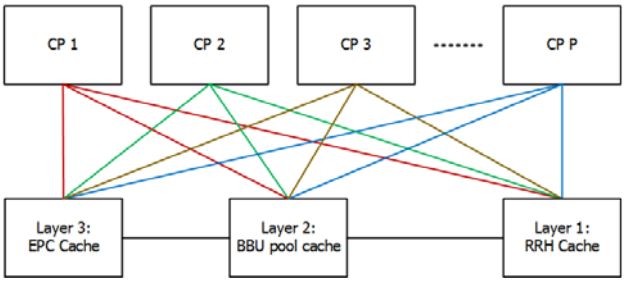

Fig. 2. Three layers hierarchical caching architecture

Cache capacity and layers: We consider that the MNO has total cache capacity $\mathrm{C} = \sum_{k=0}^{K} \mathrm{C_k}$ located in different K layers, namely (3) EPC, (2) BBU pool, (1) RRH, where $\mathrm{C_k}$ is the cache storage at each layer $k \in K$. The total cache of capacity C is used to cache the contents from multiple CPs based on their demands and payments.

Cooperative Hierarchical Caching: For retrieving content, we consider content requests arrives at RRHs according to the Poisson process, with arrival rate λ, where λ_n^{τ} is the arrival rate for the content τ at RRH $n \in N$ from consumer $v \in V$. For the content $\tau \in T$ which is not cached at RRH $n \in N$, the RRH forwards the request to BBU $\mathrm{l} \in B$, where λ_l^{τ} is the arrival rate for the content $\tau \in T$ at BBU pool $\mathrm{l} \in B$. The BBUs exchanges information for contents cached on its RRHs. For the content $\tau \in T$ which is not cached at BBU $\mathrm{l} \in B$, the BBU $\mathrm{l} \in B$ can forward request to another RRH $n' \in N$ (managed by BBU $\mathrm{l} \in B$), or to its neighbor BBU denoted BBU $\mathrm{l}' \in B$, or to the EPC E for the content $\tau \in T$ which is not cached at BBUs. The arrival rate for the content $\tau \in T$ at BBU $\mathrm{l}' \in B$ is denoted $\lambda_{l'}^{\tau}$, while λ_e^{τ} is the arrival rate for the content $\tau \in T$ at EPC. From the content which is not cached at EPC, EPC retrieve the content outside of MNO network.

3. Weighted Proportional Cache Allocation

In this section, we discuss in details our weighted proportional cache allocation [4] in three layers hierarchical caching architecture, where each CP $i \in P$ receives cache storage which is proportion to its demand. We consider that MNO implements total cache storage C at the cost of $Q(C)$. Once cache storage is implemented, it is available to P CPs, where $\mathrm{P} \geq 2$. As described in Fig. 2, each cache storage $\mathrm{C_k}$ at layer $k \in K$ is divisible, and can be used to cache the contents from multiple CPs. Moreover, we consider also that each CP $\mathrm{i} \in \mathrm{P}$ produces the content $\tau \in T$ that needs to be cached near to the consumers inside the MNO's network.

Each CP $\mathrm{i} \in \mathrm{P}$ is asked to submit bid as demand for cache storage, then MNO maps the vector of bids to the cache storage and computes the payment that each CP $\mathrm{i} \in \mathrm{P}$ has to pay for getting cache storage. Furthermore, for helping the CPs to prepare their bids, MNO announces the available cache storage $\mathrm{C_k}$ at each layer $k \in K$ to CPs, and the sum of the bids/demands placed on each network layer $k \in K$. However, the MNO does not reveal the bid(s) of one CP to another.

Let us consider CP $\mathrm{i} \in \mathrm{P}$ submits nonnegative cache demand as a bid b_{ik} $(b_{ik} \geq 0)$ for layer $k \in K$. Let us denote $b_i = (b_{i1},\dots,b_{iK})$ as the vector of bids submitted by CP $\mathrm{i} \in \mathrm{P}$. The total bids submitted for each layer $k \in K$ becomes $\sum_{i=0,k}^{P} b_{ik}$. To allocate cache storage to multiple CPs, let us denote $m = (m_1,\dots,m_P)$ as cache allocation vector, where $m_i = (m_{i1},\dots,m_{iK})$ represents the cache allocated to CP $\mathrm{i} \in \mathrm{P}$. Each CP $\mathrm{i} \in \mathrm{P}$ receives the fraction of cache storage at each layer $k \in K$ equals to:

$$m_{ik} = \mathrm{C_k} \frac{b_{ik}}{\sum_{i=0,k}^{P} b_{ik}} \qquad (1)$$

For $\sum_{i=0,k}^{P} b_{ik} = 0$, the cache storage at layer $k \in K$ is free of contents, i.e, it is not allocated to any CP $\mathrm{i} \in \mathrm{P}$.

We consider that the MNO has limited cache storage at each layer $k \in K$, where the total cache allocation at each layer $k \in K$ has to be less than or equal to cache storage $\mathrm{C_k}$.

Each CP $\mathrm{i} \in \mathrm{P}$ is associated with utility function $u_i\ (m_i)$, which describes its satisfaction based on the fraction $(m_{i1},\dots,m_{iK})$ of cache storage that is allocated to him. Each CP $\mathrm{i} \in \mathrm{P}$ gives weight to each cache storage, where we denote $w_i = (w_{i1},\dots,w_{iK})$ as a vector of the weights. The utility function $u_i\ (m_i)$ now becomes linear function:

$$u_i(m_i) = w_{i1}m_{i1} + \dots + w_{iK}m_{iK}, \qquad (2)$$

where $0 \leq w_{ik} \leq 1$ is the private preference of each CP $\mathrm{i} \in \mathrm{P}$ for cache storage located at layer $k \in K$.

4. Pricing Framework

As described in the above section 3, each CP $\mathrm{i} \in \mathrm{P}$ is asked to submit bid as demand for cache storage. From the cache storage m_i allocated to CP $\mathrm{i} \in \mathrm{P}$, each CP $\mathrm{i} \in \mathrm{P}$ has to contribute to the cost $Q(C)$ through payment, where the cache allocation function is known by CPs in advance before submitting their bids.

The cache storage located near to the consumers at RRH is more expensive that the cache storage located in long distance from the consumers (e.g. at EPC). Thus, consumers can receive the content with reduced delay. However, a robust pricing framework is needed that consider the cache storage available at each layer $k \in K$.

We consider $q = (q_1,\dots,q_P)$ and $q_i = (q_{i1},\dots,q_{iK})$ as vectors of payments, where q_{ik} represents the payment CP $\mathrm{i} \in \mathrm{P}$ has to pay MNO for cache storage allocated to him at layer $k \in K$. Based on cache locations, the cache storages have different values from both MNO and CPs. Each CP $\mathrm{i} \in \mathrm{P}$ gives w_i to each cache storage, and it is willing to get more cache

storages located near to its customers/consumers, so that its customers can get the content with minimize delay. On the other side, we consider that MNO sets also the weight ω_k for each layer $k \in K$, where the weight $\omega_k = 1/|k|$ needs to be applied to the payment q_{ik} that each CP $i \in P$ has to pay to MNO.

$$q_{ik}(b_{ik}, m_{ik}) = \omega_k b_{ik} m_{ik}. \quad (3)$$

The payoff of each CP $i \in P$, is equal to:

$$R_i(m_i) = u_i(m_i) - q_i(b_i, m_i), \quad (4)$$

while the payoff of MNO is equal to:

$$S(b_i) = \sum_{i=0,k}^{P} q_i(b_i, m_i) \quad (5)$$

Dominant Strategy: The truthful communication of demand should be a dominant strategy that guarantee a nonnegative utility for each CP $i \in P$, such that:

$u_i(m_i) - q_i(b_i, m_i) \geq u_i(m_i) - q_i(b'_i, m_i)$, for $b_i \geq b'_i$. (6)

The tuple $(u_i(m_i), q_i(b_i, m_i))$ becomes truthful bidding when $u_i(m_i) = q_i(b_i, m_i)$. The equation (5) has to guarantee that $0 \leq q_i \leq b_i$ for all CPs and bids, i.e., each CP $i \in P$ is not asked to pay more to MNO than its bid b_i. Further, we consider that each CP $i \in P$ does not have infinite budget, where B_i is the budget constraint. The total bid $\sum_{k=0}^{K} b_{ik}$ for CP $i \in P$ has to be less than or equal to its total budget B_i.

Best Response: We assume that each CP $i \in P$ is selfish, and it wants to maximize its utility based on its budget constraint B_i. From this perspective, each CP $i \in P$ chooses the best response that satisfy the following optimization problem:

$$\begin{aligned} &\underset{\{m_i \geq 0\}}{\text{Maximize}} && u_i(m_i) \quad (7) \\ &\text{Subject to:} && \sum_{k=0}^{K} m_{ik} \leq \sum_{k=0}^{K} C_k \\ & && \sum_{k=0}^{K} b_{ik} = B_i, \forall\, CP\ i \in P \end{aligned}$$

Nash Equilibrium: A Nash Equilibrium (NE) is achieved when each CP's bidding vector b_i is the best response, i.e., bidding vector b_i for any CP $i \in P$ is the best response to the system or any other bidding vectors b'_i, where:

$u_i(m_1(b_1), .., m_i(b_i), .., m_P(b_P)) \geq u_i\ (m_1(b_1), .., m_i(b'_i), .., m_P(b_P))$.

We consider NE as a desirable state that is stable, where no CP $i \in P$ has incentive to change its strategy.

Social welfare: We consider social welfare as a cache allocation that guarantee more revenue and high efficiency of of all CPs.

$$\begin{aligned} &\underset{\{m_i \geq 0\}}{\text{Maximize}} && \sum_{i=0}^{P} u_i(m_i) \quad (8) \\ &\text{Subject to:} && \sum_{i=0}^{P} m_i \leq C \\ & && \sum_{i=0}^{P} b_i = \sum_{i=0}^{P} B_i \\ & && \sum_{i=0}^{P} q_i(b_i, m_i) \geq Q(C),\ \forall\, CP\ i \in P \end{aligned}$$

5. Performance evaluation

In this section, we present the performance evaluation of our proposal. During the evaluation, we use numerical analysis, where Julia language [16] is used.

We set number of bidders/CPs equals to $P = 10$, The bids are uniformly and randomly distributed over the interval from 100 USD to 700 USD per 1 GB of cache storage. The total cache storage of MNO is set to $C = 100000$ GB, where the cost of cache storage implementation is set to $Q = 3.625$ per 1GB. The number of layers is set to $K = 3$.

The Fig.4 shows the variation of the bids of CPs, which are in range from 100 USD to 700 USD in each layer $k \in K$. Each CP $i \in P$ received cache storage which is proportion to its demands/bids and available cache storage C_k. Based on CP bids, the cache storage allocated to CPs varies from 1600GB (bidder 3) at layer 2 to 6000 GB at layer 1(bidder 1).

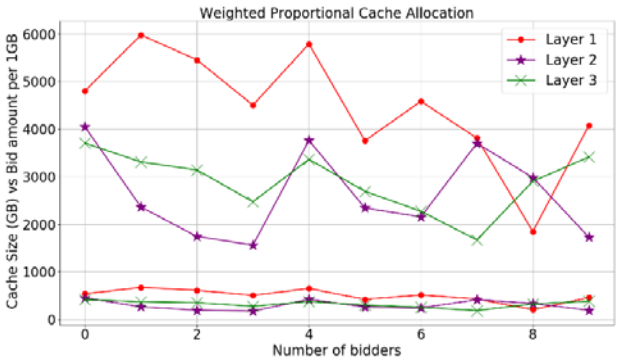

Fig. 4. Weighted Proportional Cache allocation

6. Conclusion

In this paper, we proposed a cooperative and weighted proportional cache allocation for 5G, where MNO caches the contents from multiple CPs, in three layers hierarchical caching architecture, based on their demands and payments. The numerical analysis shows that our proposal is easy to implement in production network. In the future, we aim to extend our proposal with more performance analysis.

7. Acknowledgement

This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP) (NRF-2017R1A2A2A05000995). *Dr. CS Hong is the corresponding author.

8. References

[1]. Checko, Aleksandra, et al. "Cloud RAN for mobile networks—A technology overview." IEEE Communications surveys & tutorials 17.1 (2015): 405-426.

[2]. Wang, Xiaofei, et al. "Cache in the air: exploiting content caching and delivery techniques for 5G systems." IEEE Communications Magazine 52.2 (2014): 131-139.

[3]. Tran, Tuyen X., Abolfazl Hajisami, and Dario Pompili. "Cooperative hierarchical caching in 5G cloud radio access networks (C-RANs)." arXiv preprint arXiv:1602.02178 (2016).

[4]. Ren, Shoushou, et al. "Collaborative EPC and RAN Caching Algorithms for LTE Mobile Networks." Global Communications Conference (GLOBECOM), 2015 IEEE. IEEE, 2015.

[5]. Nguyen, Thành, and Milan Vojnovic. The weighted proportional allocation mechanism. Technical Report MSR-TR-2010-145, Microsoft Research, 2010.

[6]. J. Bezanson, A. Edelman, S. Karpinski, and V. B. Shah, "Julia: A fresh approach to numerical computing," arXiv preprint arXiv:1411.1607, 2014.