MixCo: Optimal Cooperative Caching for Mobile Edge Computing in Fiber-Wireless Access Networks

Ning Wang[1], Weidong Shao[1], Sanjay K. Bose[2], Gangxiang Shen[1,*]
[1]*School of Electronic and Information Engineering, Soochow University, Suzhou, Jiangsu Province, P. R. China*
[2]*Department of Electrical and Electronic Engineering, IIT Guwahati, Guwahati, INDIA*
**Corresponding email: shengx@suda.edu.cn*

Abstract: We consider the optimal content caching problem among Mobile Edge Computing (MEC) servers in a Fiber-Wireless (FiWi) access network, to minimize the average content delivery latency subject to limited storage and computing capacity of each server. An MILP model and a Mix-Cooperative (MixCo) caching strategy are developed for efficient performance.
OCIS codes: (060.4250) Networks; (060.4256) Networks, network optimization

1. Introduction

With growing mobile data traffic, huge access capacity is required to connect nodes for real-time communications. Fiber-optics communication promises large capacity for X-hauling wireless access networks using integrated fiber-wireless (FiWi) access [1]. Cloud-Radio Access Network (C-RAN) [2] is one such integrated FiWi access network to realize future 5G access. For good Quality of Experience (QoS), latency is an important performance parameter in a 5G network with widely distributed computing/storage capacity, which users need to access quickly. This leads to the new network scenario of Mobile Edge Computing (MEC) [3]. Here, an efficient content caching strategy will decrease the content delivery latency to the users and the content load offloaded to the large data centers (DCs) by local caching.

In this paper, we consider the optimal content caching problem for MEC servers in a FiWi access network, to minimize the average content delivery latency subject to limited computing/storage capacity of the servers. Unlike most studies on content caching which only consider the limitation of storage capacity, we exploit the synergy between caching and computing capacity at the caching nodes for optimum performance. We focus on this for a FiWi access network supporting MEC and formulate the problem as a Mixed Integer Linear Programming (MILP) model. We also develop an efficient Mix-Cooperative (MixCo) caching strategy which performs close to the MILP model.

2. MixCo: Cooperative Caching among MEC Servers in a FiWi Access Network

Fig. 1 shows an example of a FiWi access network supporting MEC. This has three parts - fronthaul, mid-haul, and backhaul. The fronthaul has multiple radio nodes connected to a Distributed Unit (DU). The mid-haul connects multiple DUs to a Centre Unit (CU), and finally the backhaul connects multiple CUs to the core network via a regional switch where a large DC is attached. To support MEC, a MEC server (or micro DC) is deployed at each DU (also CU) to provide computing/storage capacity for local latency-sensitive applications.

To reduce delivery latency when the local MEC server's cache does not have the requested content, we propose that non-local MEC servers that are close to the user should handle the request cooperatively. This is called the *MixCo strategy*. Specifically, as shown in Fig. 1, we divide the FiWi access network into multiple *zones* with each CU and its connecting DUs and fronthaul networks as a zone. Each zone is subdivided into multiple *sub-zones* with each sub-zone having a DU and its fronthaul network. All the user requests in each sub-zone are served by its sub-zone DU and local MEC server. If the content is not locally cached or if the server is fully loaded, the request is sent to other MEC servers in the same zone. The CU acts as the central controller to handle forwarding of the requests. The MEC servers in the same zone cooperate to serve user requests.

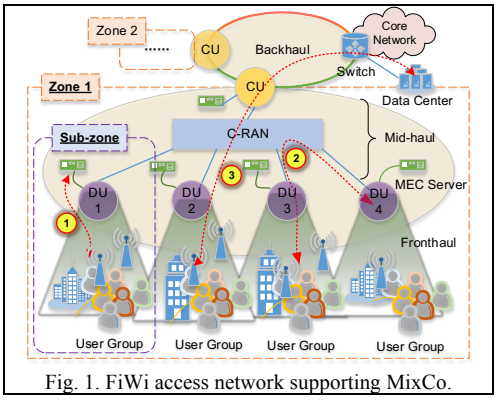

Fig. 1. FiWi access network supporting MixCo.

As shown in Fig. 1, this leads to three possible scenarios. **Scenario-1** is when the local MEC server can directly serve a request via its fronthaul network. If the local MEC server does not cache the content requested by the user or is fully loaded, the DU needs to consult with its CU to find a MEC server in the same zone that has the content and the processing capacity for that request. If there is any, the request is forwarded to this MEC server, and the content is delivered from that server to the user. This is **Scenario-2**. **Scenario-3** is when even this cannot be done and the CU forwards the request to a large DC at the border of the backhaul network and the core network, letting the DC serve the request. This forms a hierarchical serving process. With a request forwarded to higher-level servers, the content delivery latency would increase accordingly. With limited computing/storage capacity at each MEC server, *the*

caching strategy in different MEC servers in the same zone to minimize the average content delivery latency is the problem considered in the following section.

3. MILP Model and MixCo-based Heuristic Algorithm

To find the best cached content in each MEC server of a zone, we formulate the problem as an MILP model and develop a heuristic algorithm based on the MixCo strategy. The model and the heuristic algorithm are presented next.

A) MILP model: The sets, parameters, and variables for this are defined as follows. $\boldsymbol{C}$ is the set of content items. $\boldsymbol{S}$ is the set of MEC servers in a zone. For easy analysis, each content item is assumed to occupy one unit of storage. s_i and c_i are the respective maximum storage space and processing capacity of MEC server i. Here the processing capacity is derived from the computing capacity of the server and the transmission capacity of the link connected to the server. We choose the bottleneck one as the maximum processing capacity of the sever. It is measured in units of maximum content items to be processed. λ_i is the content request load at DU node i. p_k^i denotes the popularity of content item k in the sub-zone associated with DU node i, which follows the Zipf law. The three delivery scenarios shown in Fig. 1 will have different delivery latencies. We use L_1, L_2, and L_3 to represent these, respectively. In general, $L_3 \gg L_2 > L_1$. Δ is a large value. For variables, ρ_k^i is a binary variable that equals 1 if the content item k is cached at MEC server i. ψ_k is a binary variable that equals 1 if the content item k is cached in the current zone. $\xi_{i,k}^j$ is a variable (real) to indicate the fraction of the request load for content item k in sub-zone i that is served by MEC server j. This MEC server could be a local one of the current sub-zone, or the one that is responsible for the other sub-zone.

Objective: *Minimize* $\sum_{i\in\boldsymbol{S},k\in\boldsymbol{C}}(L_1\cdot\xi_{i,k}^i + L_2\cdot\sum_{j\in\boldsymbol{S},j\neq i}\xi_{i,k}^j + L_3\cdot(\lambda_i\cdot p_k^i - \sum_{j\in\boldsymbol{S}}\xi_{i,k}^j))/\sum_{i\in\boldsymbol{KS}}\lambda_i$, aiming to minimize the average content delivery latency. This is subject to the following constraints.

$$\sum_{k\in\boldsymbol{C}}\rho_k^i \le s_i \ \ \forall i\in\boldsymbol{S} \quad (1) \qquad \psi_k \le \sum_{i\in\boldsymbol{S}}\rho_k^i \le \psi_k\cdot\Delta \ \ \forall k\in\boldsymbol{C} \quad (2) \qquad \xi_{i,k}^j \le \rho_k^j\cdot\Delta \ \ \forall k\in\boldsymbol{C}; i,j\in\boldsymbol{S} \quad (3)$$

$$\sum_{j\in\boldsymbol{S}}\xi_{i,k}^j \le \lambda_i\cdot p_k^i\cdot\psi_k \ \ \forall k\in\boldsymbol{C}; i\in\boldsymbol{S} \quad (4) \qquad \sum_{k\in\boldsymbol{C},j\in\boldsymbol{S}}\xi_{j,n}^i \le c_i \ \ \forall i\in\boldsymbol{S} \quad (5)$$

In the objective, $\lambda_i\cdot p_k^i - \sum_{j\in\boldsymbol{S}}\xi_{i,k}^j$ is the load of content item k offloaded to the large DC. Constraint (1) ensures that the storage space required to cache content items at each MEC server does not exceed its storage limit. Constraint (2) ensures that a content item cached in a zone must be cached in at least one MEC server in the zone. Constraint (3) means that if a server does not cache a content item, then there is no load of the content item on this server. Constraint (4) means that for a content item, the sum of its load served by different servers should not exceed its total load. Constraint (5) means that the total content load assigned to a server should not exceed its processing capacity.

B) Heuristic algorithm (MixCo): We also develop a heuristic algorithm based on the MixCo strategy to *cooperatively* cache content and assign load to different MEC servers. Since content items have different popularities in a sub-zone and in the whole zone, we propose to split the storage space in each server into two parts. The first part is called *local sub-zone storage space*, used to cache the *most popular* content items locally for the sub-zone. The second part is called *zone-shared storage space*, used to *cooperate with* the zone-shared storage spaces in the other servers to form a large zone-wide storage space. This space is used to cache *non-locally popular* content items *as many as possible* based on their zone-wide popularities. To cache as many content items as possible, only one copy of content item is cached in the zone-shared storage space.

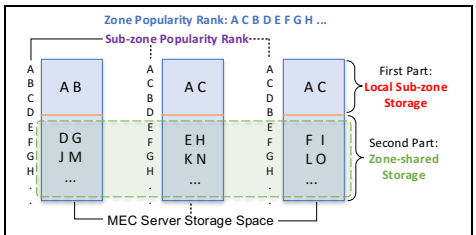

Fig. 2. An example of MixCo content caching.

Fig. 2 shows an example for the MixCo-based content caching process. There are three MEC servers that cache different content items A, B, C, etc. These content items have different sub-zone popularity ranks and zone-wide popularity ranks as shown. The key idea of the MixCo strategy is to first use a fixed local sub-zone storage space to cache the most popular content items according to their local popularity ranks. For example, for MEC server 1, content items A and B are cached in the local sub-zone storage. The remaining storage space of the server is then treated to *cooperatively* form a large zone-shared storage space with other servers, used to cache content items that are not cached in any server's local sub-zone storage space in the zone. The selection of cached content items is based on the zone-wide popularity rank. There is only one copy cached in the large zone-shared storage space for each content item so as to cache content items as many as possible. As A, B, and C have been cached in the local sub-zone storage spaces, the remaining content items to be cached in the large zone-shared storage space are D, E, F, etc. As an example, D, E, F are in turn cached in server 1, 2, 3, respectively, where each content item is only cached in one of the servers.

4. Simulations and Performance Analyses

To evaluate the performance of the proposed MixCo strategy, we performed simulation studies based on the following conditions. We assumed that there are 300 content items with each consuming one unit of storage and processing capacity. A MEC zone that contains 5 DUs (i.e., 5 MEC servers) was simulated, where each MEC server has 50 units

of storage space and can support the streaming of 100 content items simultaneously. In addition, the DC at the edge of backhaul is assumed to have a huge storage space and processing capacity, which can serve any number of user requests. The content delivery latencies of Scenarios (1)-(3) are assumed to be in the ranges of [1~3ms], [6~10ms], and [15~20ms], respectively. The content popularity distribution is subject to the Zipf law, i.e., $p_k = (1/k^{\alpha})/(\sum_{k=1}^{|C|} 1/k^{\alpha})$, where α=0.88. We employed the commercial software AMPL/Gurobi to find the optimal solution for the MILP model and Java to implement the heuristic algorithm.

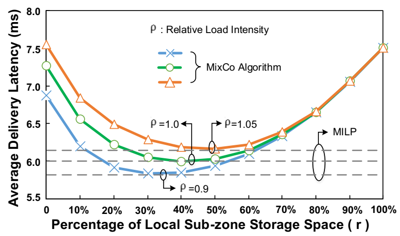

Fig. 3. Impact of sub-zone storage space.

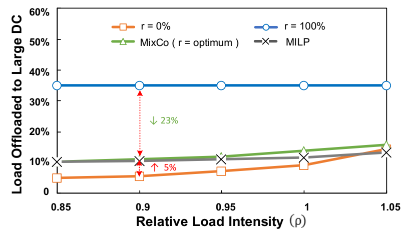

Fig. 4. Content load offloaded to large DC.

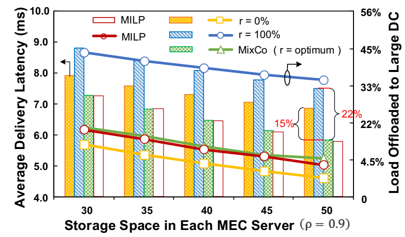

Fig. 5. Impact of MEC server storage space.

For the MixCo strategy, it is critical to decide a proper percentage of storage space for each MEC server to cache content items that are the locally most popular (i.e., the local sub-zone storage space). Simulation studies were carried out to evaluate how the average content delivery latency changes with an increasing percentage (denoted as *r*) of local sub-zone storage space at each MEC server. We show the results in Fig. 3 for different relative load intensities ρ, which is defined as *the ratio of the content request load of each sub-zone to the processing capacity of each MEC server*, i.e., λ_i/c_i. It is interesting to see that under different relative load intensities, there are different optimal values for the percentage of local sub-zone storage space. For example, for ρ=0.9, the best percentage of sub-zone storage space is ~30%, while for ρ=1.0, such a percentage is ~40%. Compared to the results obtained by the MILP model, we find that these values are very close to the optimal results, which shows the efficiency of the proposed MixCo strategy. Moreover, it is reasonable to see that an increasing relative load intensity leads to an increasing optimal percentage of sub-zone storage space since a heavier content load means more content load generated from the users in a sub-zone, That requires the local MEC server to reserve more sub-zone storage space for the locally most popular content items.

Once the optimal percentages of sub-zone storage space for different relative load intensities are found, we also compared content loads offloaded to the large DC at the edge of backhaul network for different scenarios as shown in Fig. 4. In addition to the MixCo scenario that has the optimal percentage of sub-zone storage space, we also consider two extreme scenarios with *r*=0% and *r*=100%, which correspond to the cases of zero sub-zone storage space and full sub-zone storage space at each MEC server. It is reasonable to see that more content load is offloaded to the large DC with an increasing relative load intensity since a higher load would exhaust more processing capacity at the MEC servers where one may also request more different content items. We see that the MixCo strategy is efficient to have a content load offloaded to the large DC similar to that of the MILP model. It is interesting to see that the scenario of *r*=0% shows the lowest content load offloaded to the large DC. This is because under this configuration, only one copy of a content item is cached in the storage space of all the MEC servers in a zone, which leads to the largest number of different content items cached in the zone. As a result, for any content item request, the likelihood of finding this content item cached in the zone would be high, and therefore this would minimize the load offloaded to the large DC. Similar reasoning can be made for the case of *r*=100%.

We also evaluated how the content delivery performance changes with increasing storage space at each MEC server. Fig. 5 shows the results of average delivery latency and content load offloaded to the large DC when the relative load intensity is ρ=0.9. As expected, an increasing storage space leads to a decreasing content delivery latency and less content load offloaded to the large DC. Similar performance values are found for the different schemes. The MixCo scheme performs very close to the MILP model, indicating the efficiency of the MixCo strategy.

5. Conclusion

We considered the content caching problem for a FiWi access network supporting MEC. A MixCo strategy was proposed to allocate a fraction of MEC server storage space to store the locally most popular content items and the remaining part to cooperatively cache as many less-popular content items as possible in a zone-wide fashion. An MILP model and a heuristic algorithm were developed. The results show that an optimal percentage of storage space exists for each MEC server to cache the locally most popular content items. The proposed MixCo strategy is efficient to achieve an average delivery latency and content load offloaded to the large DC which are close to the MILP model.

[1] N. Ghazisaidi and M. Maier, "Fiber-wireless (FiWi) access networks:…," *IEEE Network*, vol. 25, no. 1, 2011, pp. 36-42.
[2] A. Checko *et al.*, "Cloud RAN for mobile networks…," *IEEE Communications Surveys & Tutorials*, vol. 17, no. 1, 2015, pp. 405-426.
[3] T. Tran *et al.*, "Collaborative mobile edge computing in 5G networks: new...," *IEEE Commun. Mag.*, vol. 55, no. 4, 2017, pp. 54-61.
[4] A. Tzanakaki *et al.*, "5G infrastructures supporting end-user and operational services: the 5G-XHaul architectural...," in Proc. *ICC 2016*.