

Load-based Content Allocation Scheme for Realizing Efficient Mobile Cooperative Cache

Taiki Akiba, Celimuge Wu, and Tsutomu Yoshinaga

Graduate School of Informatics and Engineering, the University of Electro-Communications,
Chofu, Tokyo, 182-8585, Japan

takiba@comp.lab.uec.ac.jp, {celimuge, yoshinaga}@uec.ac.jp

Received: February 14, 2023

Accepted: May 23, 2023

Communicated by Shuichi Ichikawa and Takashi Yokota

Abstract

Owing to the rise in mobile users accessing high-quality video content, content delivery networks are under increased load. Therefore, a distributed cooperative caching technique, in which each mobile device functions as a cache server and shares a cache through device-to-device communication has been proposed to alleviate the network load. However, efficient content placement on mobile devices is challenging, because the cache capacity of each mobile device is limited and did not remain at a fixed location. In this study, we propose a load-based content allocation (LBCA) method that efficiently distributes cache data to mobile devices based on the load of the base station (BS). The proposed scheme uses scalable video coding (SVC) that subdivides the content data hierarchically and dynamically selects the video quality depending on the network load. Additionally, we propose a multi-stage LBCA (M-LBCA) that divides the BS distribution area into concentric clusters to efficiently manage cache capacity for SVC data while considering the distance from the BS. Simulations demonstrated a decrease in the total number of users unable to continually view content on congested networks when using LBCA and M-LBCA as compared to the existing method in four evaluation environments. Therefore, our experiments demonstrate that the proposed cache control scheme improved the user's quality of experience.

Keywords: distributed and cooperative cache, device-to-device communication, mobile network, content delivery network

1 Introduction

Currently, mobile users can watch video content anywhere anytime owing to the advancement in mobile communication technology and easily accessible video-on-demand (VoD) services. However, internet traffic increases with the number of mobile users and an increase in data size caused by higher video quality [1]. Mobile cache technology, which can address the increasing communication traffic has attracted significant attention. This enables mobile devices to act as cache servers and receive data from neighboring devices via device-to-device (D2D) communication. Sharing data among mobile devices reduces the communication load of base stations (BSs) and upper network layers considerably. A distributed cooperative cache scheme was proposed to utilize the limited cache capacity of each mobile device efficiently. This scheme shares the cache data among multiple cache servers through cooperation to increase the total cache capacity by holding different contents for each cache server [8] [6]. In addition, mobile cache servers use scalable video coding (SVC) schemes that have been proposed to improve the quality of experience (QoE) of users [10] [17]. In

SVC, the video data are divided into a base layer and an extension layer to improve the quality. A device can replay the content by receiving a minimum of one base layer. Thus, the network load can be reduced dynamically by determining the delivery of the extended layer. However, existing distributed cooperative caching methods do not use SVC or the radio attenuation characteristics of wireless networks.

This study proposes load-based content allocation (LBCA) that efficiently assigns cache data, which are then subdivided by the SVC scheme, onto mobile devices depending on the load of BS considering radio attenuation. Flexible cache control based on the BS load enables continuous viewing of content while maintaining a certain level of quality even when the mobile network is congested, which improves the QoE of users. We then extend the LBCA scheme to a multi-stage LBCA (M-LBCA) scheme that divide the BS distribution area into concentric circles, and multiple concentric clusters to efficiently manage the cache capacity of SVC data when the distance from the BS was considered.

2 Related Work

2.1 Content-access bias

The access to video content was skewed toward the most popular video. Studies analyzing the popularity trends of videos in VoD services demonstrated that the access bias of videos follows a Zipf distribution [14] [16]. Using the Zipf distribution, the access probability of rank i can be expressed as follows:

$$P(i) = \frac{1/i^s}{\sum_{r=1}^C 1/r^s}, \quad (1)$$

where C is the number of contents, and s is the strength of the access bias. The typical access bias for video content ranges from $s = 0.6$ – 0.7 . In this study, we assume an environment where the viewed content is highly skewed among users with similar preferences (i.e., a sports stadium or event venue).

2.2 Distributed cooperative cache server

The rise in content and demand for high-resolution video led to an increase in data size, necessitating an expansion of cache capacity. However, multiple independent cache servers often duplicate the same cache content, resulting in inefficient use of cache capacity. Therefore, a distributed cooperative cache in which multiple cache servers store different content to efficiently utilize the total cache capacity was proposed.

Nakajima *et al.* [8] proposed a distributed cooperative cache control method, colored cache that efficiently minimizes network traffic by supplying the requested content from the closer cache servers. Shiroma *et al.* [12] proposed templated elastic assignment (TEA) caching, which is a lightweight, suboptimal content allocation method for mobile devices. The TEA determines the suboptimal number of mobile devices that should cache each content to maximally reduce the traffic volume in an access network, based on the access probability distribution of the content and the total cache capacity. Once the allocation is calculated, it is reused as a template without recalculation costs, even in highly dynamic mobile network environments. We extended the TEA to support SVC schemes to enable detailed quality control.

2.3 Cache control of content on a chunk basis

Video content can be delivered to users in two approaches. The first approach involves downloading all the content data in advance and then playing them back. The second approach involves streaming the delivery system in which the data are downloaded in chunks for a few seconds and played back sequentially. The streaming delivery enables users to view content without downloading the entire data beforehand.

When users request content data from the chunks, Shiroma *et al.* [13] demonstrated that network traffic can be reduced by caching popular chunks, based on access frequency, which varies depending on the playback position of content [13]. Okada *et al.* [9] demonstrated that distributed cooperative caching is more efficient by assigning color tags to the segmented chunk data. We further divided the data into chunks and used the data corresponding to the SVC scheme described below.

2.4 Reducing traffic using scalable video coding schemes

When several devices are connected simultaneously to the same BS, network congestion deteriorates the QoE of users. Consequently, reducing the number of service requests and decreasing the transmission data size will help alleviate network congestion and reduce BS load. SVC [15] contributes to the efficient use of the cache capacity. SVC comprises a base layer, which is the minimum amount of data required for playback, and an extension layer to improve video quality. The BS then decides whether to deliver each extension layer based on network congestion status. Users can view the requested content of a specific quality by decoding the base layer with the received extension layer.

Zhan *et al.* [17] proposed an SVC-based cache placement method that minimized the average download time of content in mobile network environments. Ma *et al.* [7] also proposed a method to maximize D2D cache hits by optimizing the cache placement and cache search algorithms for SVC data. However, these methods do not consider significant changes in mobile network environments. Optimization calculations must be performed whenever the environment changes, including changes in access trends and device movements. Furthermore, heavy computation is necessary for optimization when content and user numbers are high. Therefore, we extended the TEA method to perform suboptimal cache allocation with lightweight computations to sustain the changes in mobile environments.

Okada *et al.* [10] proposed a mechanism to realize continuous video watching even during peak traffic, using a flexible SVC-based cache control for live streaming. However, they ignored radio network attenuation characteristics and a reduction in the communication speed caused by access concentration. In contrast, we consider the change in video quality that can be received under wireless transmission conditions, and we determine suboptimal content allocation by controlling the color tag for each data level of SVC.

2.5 Cache control in device clusters

Construction of clusters of multiple devices as a group and performing cache control based on the characteristics of each cluster can further improve the D2D cache hit ratio.

Huang *et al.* [3] proposed a method for building clusters around multiple randomly selected devices and perform cache control based on the request prediction in each cluster. Khan *et al.* [5] proposed a cache control method that constructs clusters of users in a D2D cache network based on the similarity of their social interests and maximizes the cache hit ratio for each cluster. However, these methods did not consider using SVC schemes, and they do not control the quality of the cached video based on the communication conditions.

In this study, users were clustered based on their distance from the BS. Therefore, users in the same cluster can receive the same video quality. We aimed to improve the QoE of users by controlling the cache capacity of each SVC level data on a per-cluster basis. We demonstrate the effect of SVC-based cache control using concentric device clusters around the BS.

3 Load-based Content Allocation

3.1 Color tagging assignment for SVC data

3.1.1 Color tagging for each SVC level

In a cache control method using color tags [12], popularity content ranking of the chunks was determined from the content access history. Color tags were assigned based on their popularity. In

LBCA, the chunk data were split into layers of SVC, and a color tag was assigned to each chunk of data.

Assume that SVC data consist of three quality layers: low, middle, and high. The base layer is composed of low-quality data, and the enhanced layer is comprise middle- and high-quality data. Fig. 1 shows an example of color tag assignment to the chunks of each layer. The data chunks are











Quality Level \ Chunks	Base Layer	Enhancement Layers	
	Low	Middle	High
#1			
#2			
#3			
#4			
#5			

Figure 1: Assignment of color tag per level of quality of SVC layer

sorted by access popularity from ranks #1 to #5. As shown in Fig. 1, each data chunk is assigned to at least one color independently for each visual quality level. For example, a data chunk of rank #1 has three colors (red, blue, and yellow) for low- and middle levels and two colors (red and blue) for high levels. We assigned an increased number of colors to the data chunks that ranked higher, three colors for rank #1 and two colors for rank #2. In addition, we assigned more colors to the chunks in the base layer, three colors for the low level and two colors for the high level. This is because higher-ranked data chunks are more accessible than the lower-ranked data chunks, and the data chunk of the base layer is necessary to view the content, whereas that of an enhanced layer must be viewed at a higher quality. In LBCA, each cache server is assigned a single color and only caches data chunks with the same color tag. Data chunks with more color tags can be cached on a larger number of cache servers. Therefore, individual color tags were assigned to each quality level, chunk by chunk, the quality of the video to be cached can be fully controlled.

3.1.2 Partitioning of cache area for each SVC level

In LBCA cache control, the chunk datum is the smallest data unit. Therefore, the available cache capacity of each device is partitioned into SVC quality levels.

Let C be the total cache capacity available for the entire D2D cache network. The cache capacity C_q allocated to level q in the D2D cache network can be calculated as follows:

$$C_q = r_q C, \quad (2)$$

where $r_q (0 \leq r_q \leq 1)$ is the ratio of the cache capacity allocated to level q of SVC data.

We extended the TEA to make efficient use of the cache space reserved for each level. The TEA determines the number of devices that should have each chunk using a suboptimal calculation, based on the parameters of total available cache capacity and request bias. In LBCA, C_q is the total cache capacity, and a color tag is assigned to each level $q \in \mathcal{Q}$.

3.2 Load-based cache capacity control

3.2.1 Cache capacity allocation based on receive opportunities

To efficiently utilize the limited cache capacity of devices, it is necessary to adjust the cache capacity C_q for each level, based on mobile network conditions. For instance, cache space for the extension layer is wasted in the access concentration scenario because the data chunks in the base layer can

be transferred. In such cases, more cache space must be used on the mobile device for the base layer rather than for the extended layer. Therefore, LBCA determines the cache capacity based on the opportunities received at each SVC level.

LBCA determines the cache capacity ratio of each SVC level based on the ratio of traffic on each SVC level to the total traffic between the devices and the BS. The ratio of the cache space allocated for SVC level q to the total device cache, r_q ($0 \leq r_q \leq 1$) is calculated as follows:

$$\begin{aligned} r_q &= \frac{\text{Traffic occurring for level } q \text{ (bps)}}{\text{Total traffic (bps)}} \\ &= \frac{n_q B_q}{\sum_{r=0}^Q n_r B_r}, \end{aligned} \quad (3)$$

where n_q is the number of devices that can view content quality q and B_q is the bitrate of SVC level q . The viewing quality q indicates that the communication speed between the device and BS is sufficient to receive all the SVC data from levels 1 to q . The BS periodically calculates r_q from the value of n_q and assigns color tags depending on changes in the mobile network.

3.3 Estimation of communication speed with radio attenuation characteristics

Determining the cache ratio for each SVC level Equation (3) uses the number of devices n_q , which is the number of devices available to view video quality q . LBCA estimates n_q using the radio attenuation characteristics of wireless networks.

From the Shannon–Hartley theorem, the communication band capacity C of a device can be calculated as follows:

$$C = b \log_2(1 + SNR) \quad (4)$$

where b denotes the bandwidth allocated to the device and SNR is the signal-to-noise ratio. When the bandwidth is allocated equally to all devices connected to the BS, bandwidth b can be replaced with

$$b = \frac{B}{n}, \quad (5)$$

where n denotes the number of concurrent connections to the BS and B is the transmission bandwidth of the BS. SNR can be calculated as follows:

$$SNR = Pt - Lb - Noise, \quad (6)$$

where Pt (dBm) is the transmission power, Lb (dBm) is the transmission loss, and where $Noise$ (dBm) denotes the background noise. In this study, we assume a large space with few obstacles (i.e., a stadium or event venue) and a free-space model [2] is used in a mobile network environment to calculate transmission loss. The transmission loss is calculated using the Frith's transmission formula [2] as follows:

$$Lb = 20 \log\left(\frac{4\pi f}{c}\right) + 10\alpha \log d, \quad (7)$$

where f is the frequency, c is the speed of light, α denotes the spatial transmission coefficient, and d is the distance between the communicating nodes. From (4)–(7), the condition for maximum distance d where a device that can secure communication band capacity C is given by

$$\begin{aligned} d(n, C) &= 10^\beta, \\ \beta &= \frac{1 + Pt - 20 \log\left(\frac{4\pi f}{c}\right) - Noise - 2\frac{nC}{B}}{10\alpha}, \end{aligned} \quad (8)$$

where n denotes the number of connections to the BS. The BS was assumed to deliver content in all directions. The devices are assumed to be uniformly distributed in the delivery area of the BS.

Number of devices ensuring sufficient transmission speed among all the devices to receive videos of quality q can be calculated as follows:

$$n_q = \frac{d(n, B_q)^2}{R^2} N, \quad (9)$$

where R is the delivery area radius of the BS and N is the number of devices. From Equations (3) and (9), the cache capacity of each SVC level is determined by the number of concurrent BS connections n .

4 Multistage Load-based Content Allocation

4.1 Divided cache network

Within a cache network constructed using LBCA The cache capacity ratio for each SVC level was the same for all the devices. Even for devices within the same cache network, The SVC quality level that the devices often receive depends on their locations because the communication speed changes depending on the distance from the BS. More efficient content placement can be achieved by grouping the devices with similar communication speeds, and individual cache controls for each group. We propose the M-LBCA, which divides a cache network based on its distance from the BS constructs concentric clusters of devices and applies LBCA to each cluster.

As shown in Fig. 2, The communication area is divided into multistage clusters around the BS. The area outline R_i of cluster \mathcal{C}_i ($0 < i \leq M$) is defined as follows:

$$\begin{aligned} R_0 &= 0, \\ R_i &= \frac{R_{BS}}{M} i, \end{aligned} \quad (10)$$

where M is the number of M-LBCA stages and R_{BS} is the radius of the distribution area of the BS. The set of devices in cluster \mathcal{C}_i is defined as

$$\mathcal{C}_i = \{d \mid R_{i-1} \leq \text{distance}(d, BS) < R_i\}, \quad (11)$$

where $\text{distance}(d, BS)$ is the linear distance between devices d and the BS. The topology of the M-LBCA for cases $M = 3$ and $R_{BS} = 150$ (m) is shown in Fig. 2.

4.2 Allocation of cache capacity for each cluster

Similar to 3.2.1, the cache capacities are determined based on the received opportunities, for each SVC in each cluster, respectively. In cluster \mathcal{C}_i , the ratio $r_{\mathcal{C}_i, q}$ ($0 \leq r_{\mathcal{C}_i, q} \leq 1$) of the cache space allocated to quality level $q \in \mathcal{Q}$ is calculated as

$$\begin{aligned} r_{\mathcal{C}_i, q} &= \frac{\text{Traffic occurring for level } q \text{ in cluster } c \text{ (bps)}}{\text{Total traffic in cluster } c \text{ (bps)}} \\ &= \frac{n_{\mathcal{C}_i, q} B_q}{\sum_{r=0}^Q n_{\mathcal{C}_i, r} B_r}, \end{aligned} \quad (12)$$

where $n_{\mathcal{C}_i, q}$ is the number of devices that can be viewed, video quality q in the cluster $n_{\mathcal{C}_i, q}$.

As in 3.3, $n_{\mathcal{C}_i, q}$ is calculated as follows: with the radio attenuation characteristics of wireless networks. Let N be the number of devices in the BS distribution area, and assume a case where devices are uniformly distributed in the area. The number of devices $N_{\mathcal{C}_i}$ belonging to the cluster \mathcal{C}_i is calculated as follows:

$$N_{\mathcal{C}_i} = \frac{R_i^2}{R_{BS}^2} N. \quad (13)$$

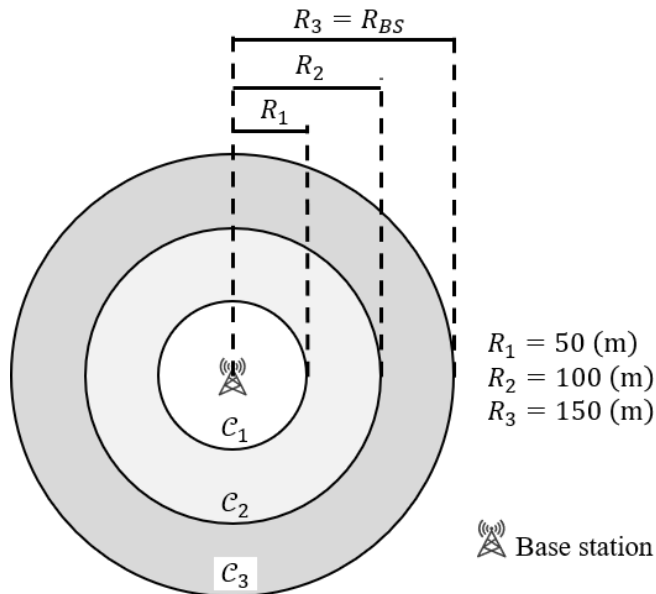


Figure 2: Three-stage M-LBCA topology.

Using Equation (8), the number of devices $n_{C_i,q}$ that can view the quality level q in cluster C_i is calculated as:

$$n_{C_i,q} = \begin{cases} 0, & \text{if } d(n, B_q) < R_{i-1}, \\ N_{C_i}, & \text{if } d(n, B_q) > R_i, \\ \frac{d(n, B_q)^2 - R_{i-1}^2}{R_i^2 - R_{i-1}^2} N_{C_i}, & \text{otherwise.} \end{cases} \quad (14)$$

From Equations (12) and (14), the cache capacity for each SVC level is determined by the number of concurrent BS connections, n .

5 Evaluation

5.1 Simulation parameters

In this study, we demonstrate the usefulness of the proposed LBCA through simulations. The simulation parameters are summarized in Table 1. Parameters related to the BS, such as bandwidth, frequency, and transmission power were determined using [11] [4]. The mobile-network topology is shown in Fig. 3.

The content request models are shown in Figs. 4 and 5. Aggressiveness refers to the probability that each user generates one new request per second. Fig. 4 illustrates a high prevalence of protracted requests. Fig. 5 varies the number of requests in shorter intervals.

Three cases were evaluated and summarized in Table 2. Devices are randomly moved at a speed of 0 to 1 m/s within the BS distribution area. We compared the QoE of users in each of the TEA, LBCA, and M-LBCA. The TEA assigns color tags equally to all SVC levels in a chunk. M-LBCA is evaluated for 2 to 4-stage configurations. The topologies of the M-LBCA are shown in Figs. 6, 2, and 7. The communication sequence used in our simulation is as follows:

1. User's requests are issued by randomly selected devices according to the used request pattern.
2. Selected devices determine request chunks according to the Zipf distribution to simulate assumed access popularity.

Table 1: Evaluation parameters

Number of chunks	50000
Number of users	500
Cache capacity of device	100 chunks
Spectrum length	64
Request bias	Zipf ($s = 0.8$)
Bitrate of low level	1.2 Mbps
Bitrate of middle level	3.6 Mbps
Bitrate of high level	7.2 Mbps
Bandwidth (BS-to-Device)	100 MHz
Bandwidth (Device-to-Device)	50 MHz
Frequency of BS	4.5 GHz
Frequency of devices	28 GHz
Transmission power of BS	23 dBm
Transmission power of Devices	20 dBm
Noise	20 dBm
Spatial transfer coefficient α	2.9

Table 2: Evaluation setting

	Evaluation 1	Evaluation 2	Evaluation 3	Evaluation 4
Request model	Pattern 1	Pattern 1	Pattern 2	Pattern 2
Device movement	Enable	Disable	Enable	Disable

3. Selected device issue content requests for all SVC levels for the chosen chunk.
4. When the requested data chunk is located on a device cache in a D2D communication area, the user receives it via D2D communication from the neighboring device.
5. When the requested data chunk is not found on a device cache, the user receives it via the BS.
6. The BS determines the SVC level for actual service based on the device transmission speed. BS determines the actual SVC level to be transferred based on available transmission bandwidth, which can be calculated according to Equation (5).

SVC data comprise three levels of quality: low, middle, and high as shown in Fig. 1. The quality of the content that users can view is determined by the received level of SVC data. Low-level data are required to view content continuously and devices without the bandwidth required for low-level data are counted as data “failed” users. Temporal buffering was not considered in advance in this study. To view the content received from the device. Based on the simulation results, users were evaluated to determine who could continuously view the requested content as low, medium, or high quality. Fig. 8 shows the flow of chunk data from request generation to fetching.

As shown in Figs. 4 and 5, the duration of the simulation was 1200 s, and the color tags were updated every 60 s based on the access popularity trend of the contents. To vary the access popularity, chunks of new content were added during each color-tag update by replacing 0.05% of the top-ranking chunks. The initial mobile cache warm-up is performed based on the assigned color tags.

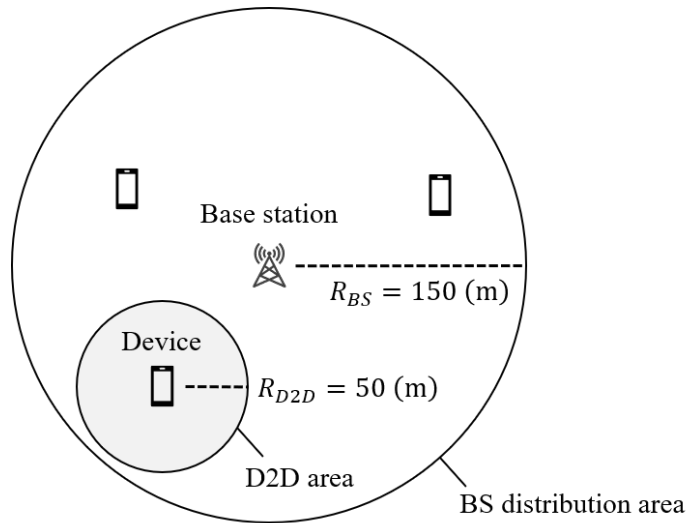


Figure 3: Mobile network topology.

5.2 Evaluation of received video quality

5.2.1 Evaluation 1

Figs. 9–13 show the ratio of the video quality that users can view in Evaluation 1. The legend “failed” in the figures indicates that the ratio of the user whose content view is temporarily stopped because the user device cannot locate the requested chunks in the neighboring mobile cache and the required receiver bandwidth cannot be maintained from the BS, even in the base layer. The numbers of concurrent BS connections are shown in Fig. 14.

For all the methods, the results indicated a significant decrease in the number of users who could view high-quality images from approximately $t = 200$. Fig. 14 shows that the number of concurrent BS connections increases rapidly at $t = 200$. This was caused by an increase in the demand for the bandwidth available at the BS owing to an increase in the number of user requests, thereby resulting in lower communication speeds for each device. In addition, the number of concurrent BS connections is generally higher using the TEA compared to the other methods. This was caused by a lower cache hit ratio via D2D communication owing to the less efficient use of the device cache compared to other methods. Furthermore, the devices in the TEA cache are at all SVC levels regardless of the congestion in the mobile network. The device can deliver only one piece of data at a time. When a device delivers a medium or high level to another device a device that requires a low level may not be able to fetch the device cache, resulting in a viewing failure. Conversely, in LBCA and M-LBCA, a low level is preferentially cached when the mobile network is congested. This increases the probability of receiving a low level from neighboring devices and may reduce the number of concurrent BS connections.

Following $t = 1150$, there is a slight difference in the user’s QoE between LBCA and M-LBCA. In environments where the number of concurrent BS connections remains low, efficient cache allocation is equivalent to using the M-LBCA even when using LBCA, which has a simpler color tag control. Fig. 15 shows the total number of users who could not continuously view requested content (i.e., failed users) in Evaluation 1. Fig. 15 shows that the number of failed users decreased by 76.8% in LBCA, 77.6% in two-stage M-LBCA, and 77.9% in three-stage M-LBCA, and 76.8%, respectively, in the four-stage M-LBCA compared with the TEA. Among the M-LBCAs, the greatest user QoE improvement was observed in the three-stage M-LBCA. There is a limit to the effect of increasing the number of stages on the user’s QoE improvement, and the optimal number of stages may differ depending on the environment. Extending LBCA to M-LBCA reduced the number of “failed” users by up to 4.7%.

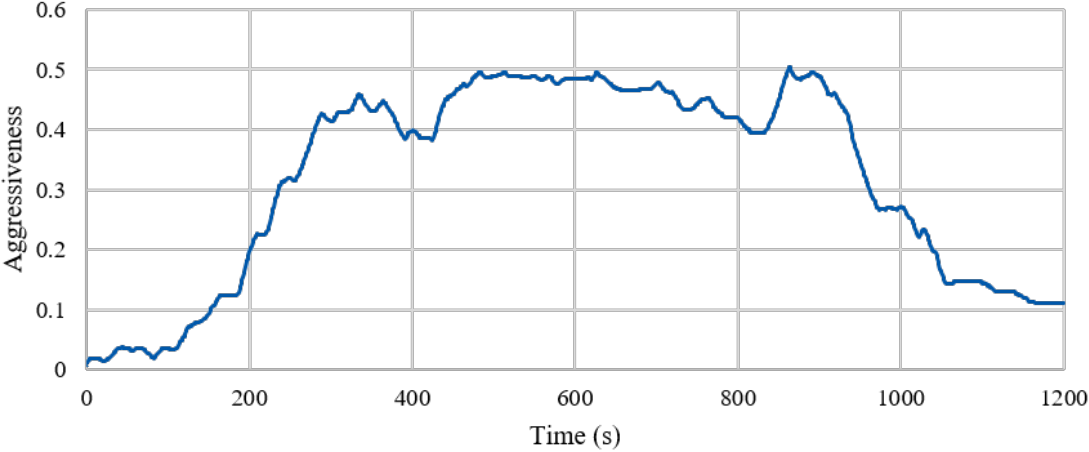


Figure 4: Request pattern 1.

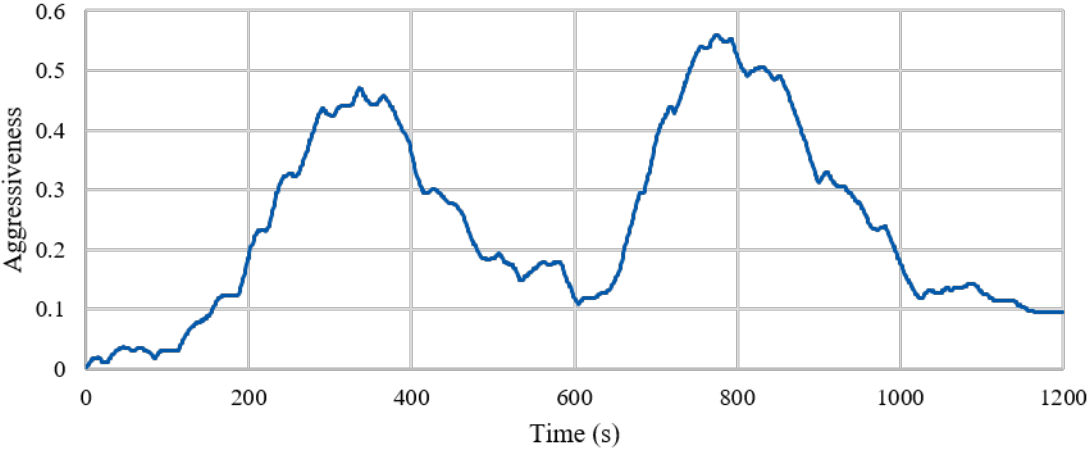


Figure 5: Request pattern 2.

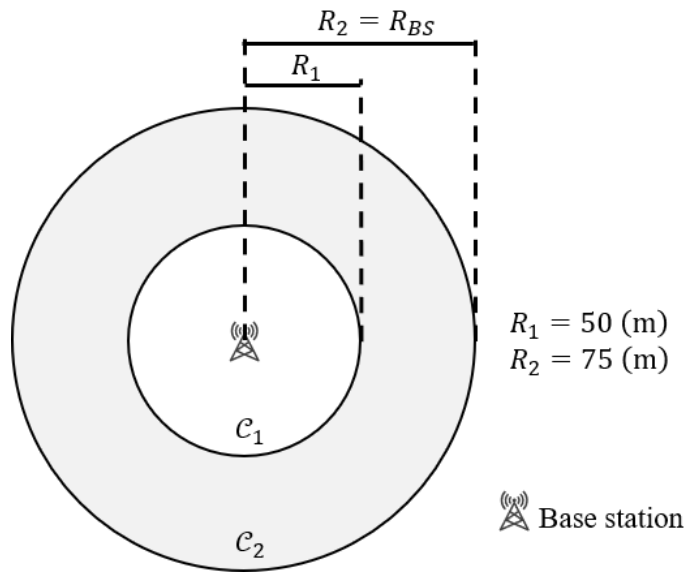


Figure 6: Two-stage M-LBCA topology

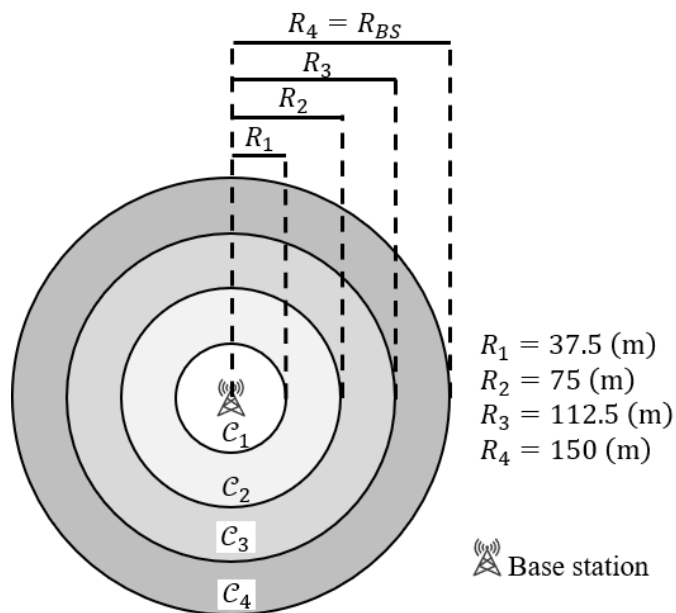


Figure 7: Four-stage M-LBCA topology

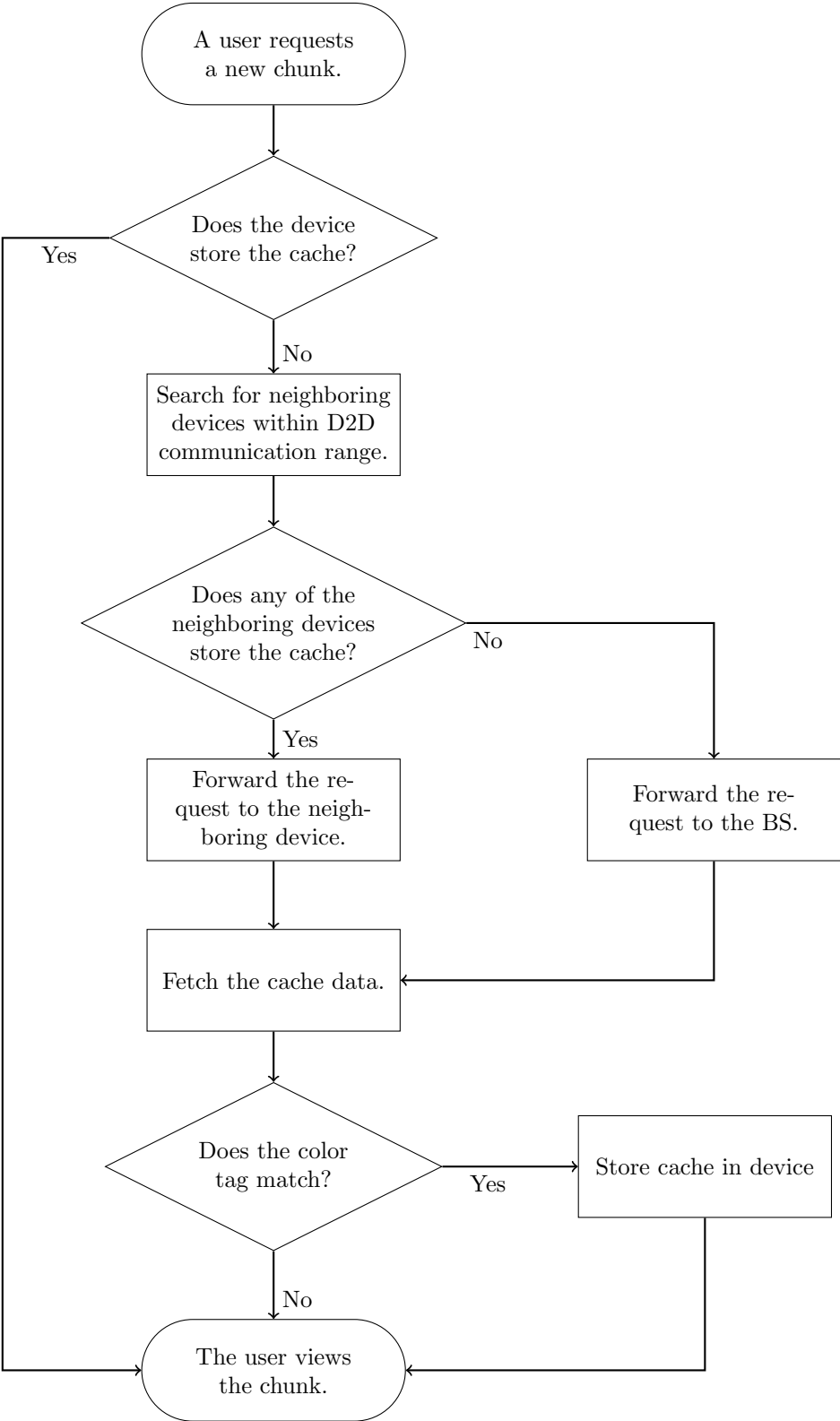


Figure 8: Flow of chunk data from request generation to fetching.

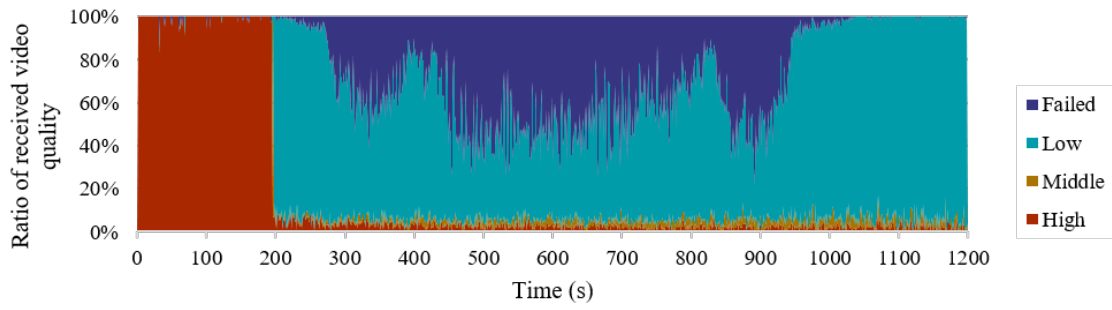


Figure 9: User's QoE with TEA in Evaluation 1.

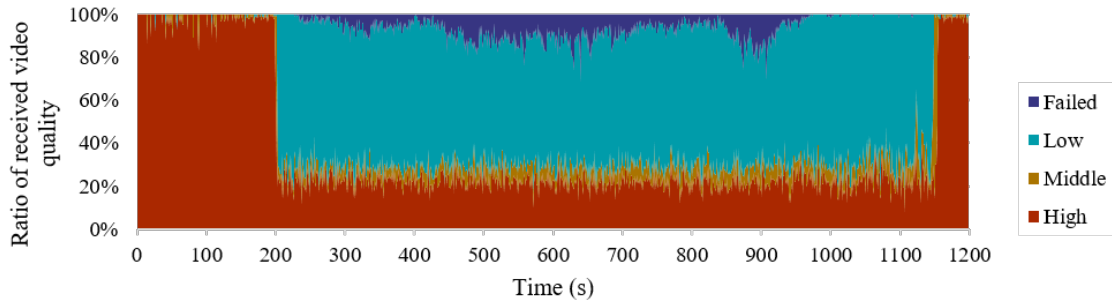


Figure 10: User's QoE with LBCA in Evaluation 1.

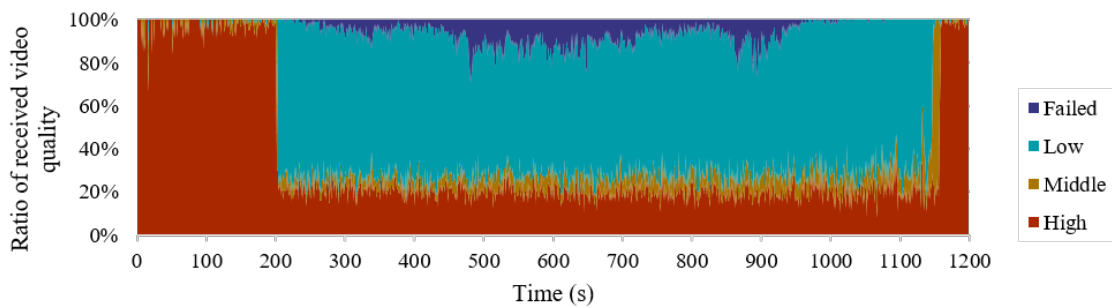


Figure 11: User's QoE with two-stage M-LBCA in Evaluation 1

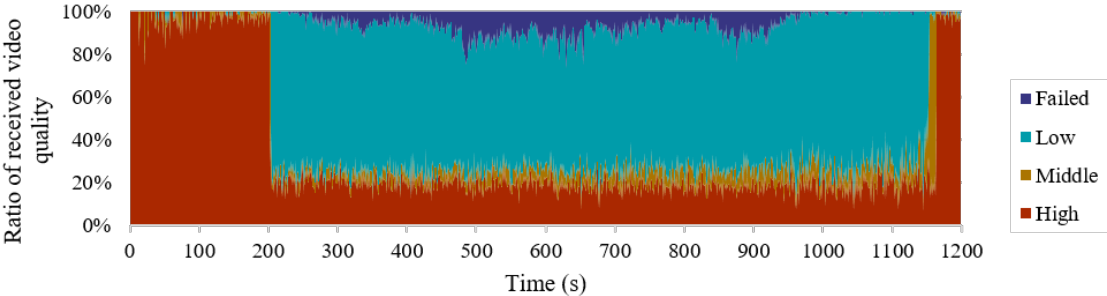


Figure 12: User’s QoE with three-stage M-LBCA in Evaluation 1

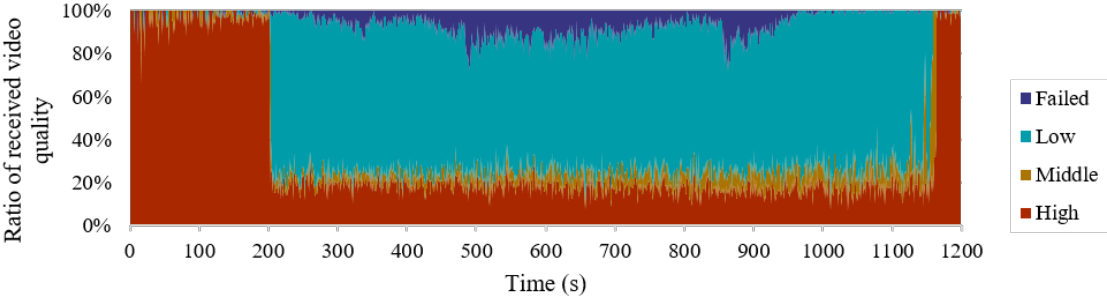


Figure 13: User’s QoE with four-stage M-LBCA in Evaluation 1

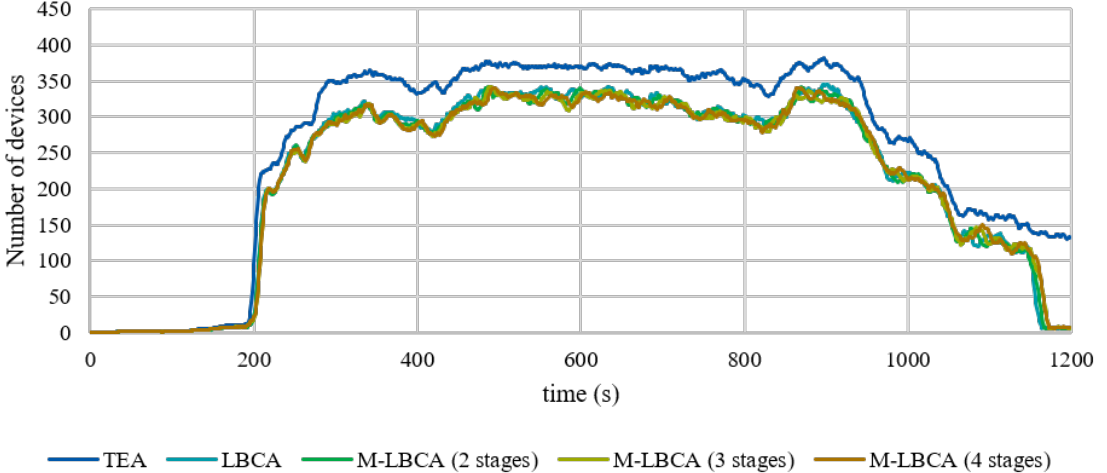


Figure 14: Number of concurrent BS connections in Evaluation 1

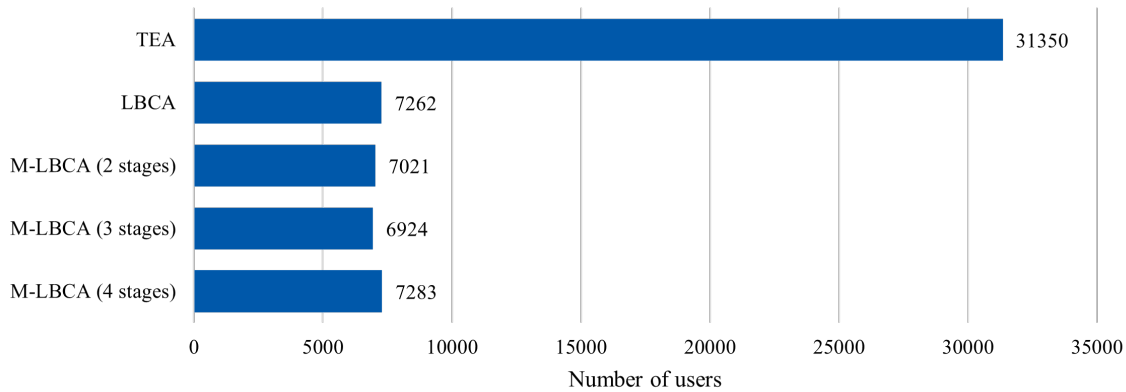


Figure 15: Number of “failed” users in Evaluation 1

5.2.2 Evaluation 2

Figs. 16–20 show the ratio of the video quality that users were able to view in Evaluation 2. The number of concurrent BS connections is shown in Fig. 21.

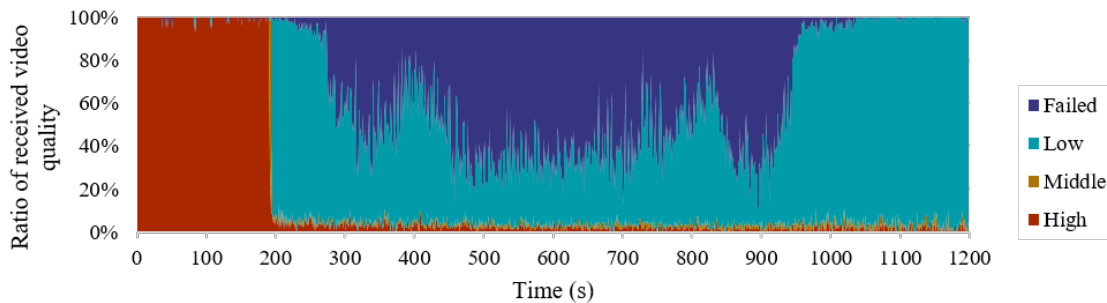


Figure 16: User’s QoE with TEA in Evaluation 2.

As in Evaluation 1, the results indicate a significant decrease in the number of users who could view high quality from approximately $t = 200$. Fig. 21 shows that the number of concurrent BS connections rapidly increases at $t = 200$.

Fig. 22 shows the total number of “failed” users in Evaluation 2. Fig. 22 shows that the number of failed users decreased by 64.0% in LBCA, 65.7% in two-stage M-LBCA, and 68.1% in three-stage M-LBCA, and 67.3% in the four-stage M-LBCA, respectively, compared with the TEA. Among the M-LBCAs, the most significant user QoE improvement was observed in the three-stage M-LBCA. Extending LBCA to M-LBCA reduced the number of “failed” users by at most 11.5%.

5.2.3 Evaluation 3

Figs. 23–27 illustrate the ratio of the video quality that users could view in Evaluation 3. The numbers of concurrent BS connections are shown in Fig. 28.

For all the methods, the results indicated a significant decrease in the number of users who could view high-quality images from approximately $t = 200$. Fig. 28 demonstrates that the number of concurrent BS connections rapidly increases at $t = 200$. In addition, most users could view high-quality videos at approximately $t = 600$ for all the methods except TEA. Fig. 28 shows that the

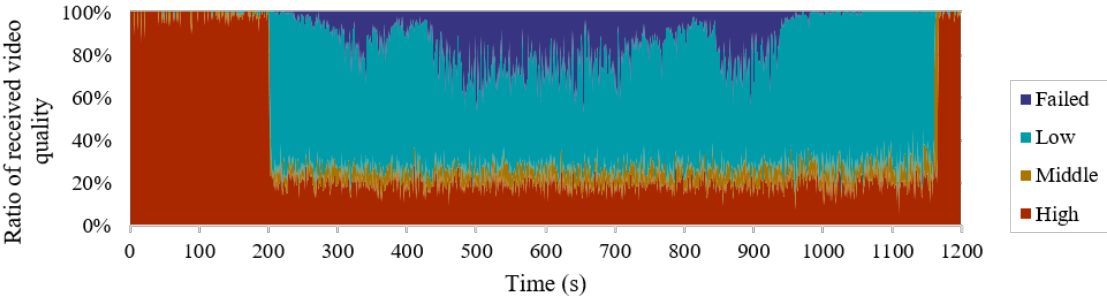


Figure 17: User's QoE with LBCA in Evaluation 2.

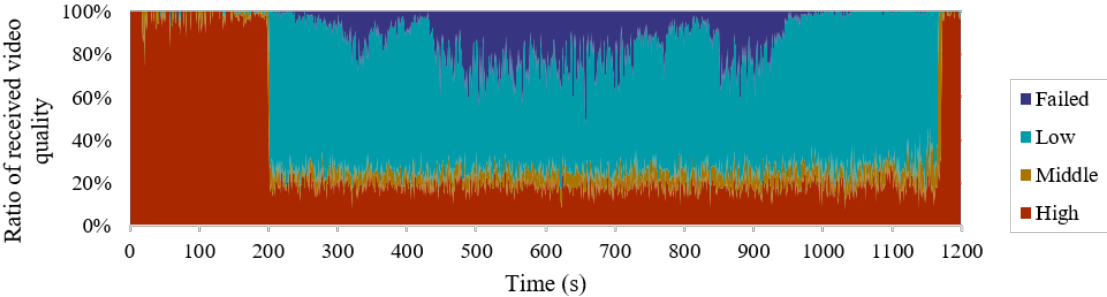


Figure 18: User's QoE with the two-stage M-LBCA in Evaluation 2.

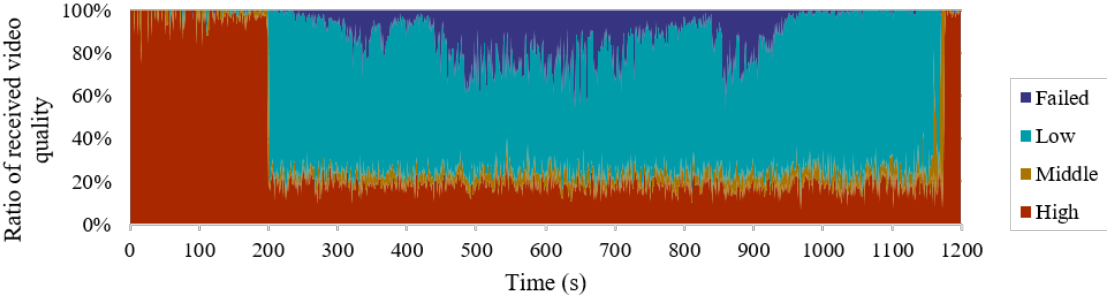


Figure 19: User's QoE with the three-stage M-LBCA in Evaluation 2.

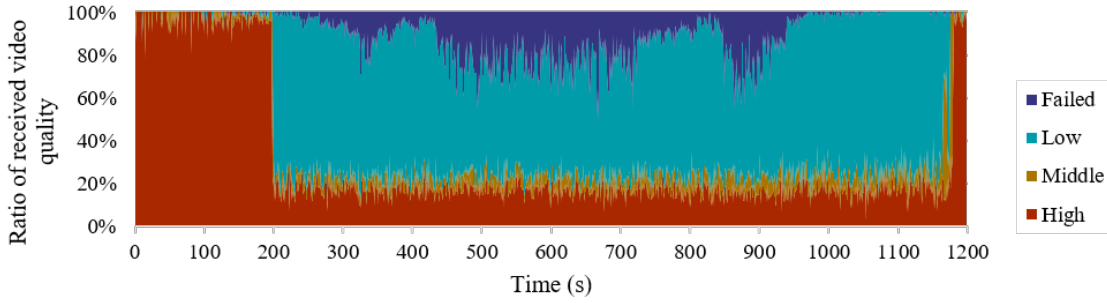


Figure 20: User’s QoE with the four-stage M-LBCA in Evaluation 2.

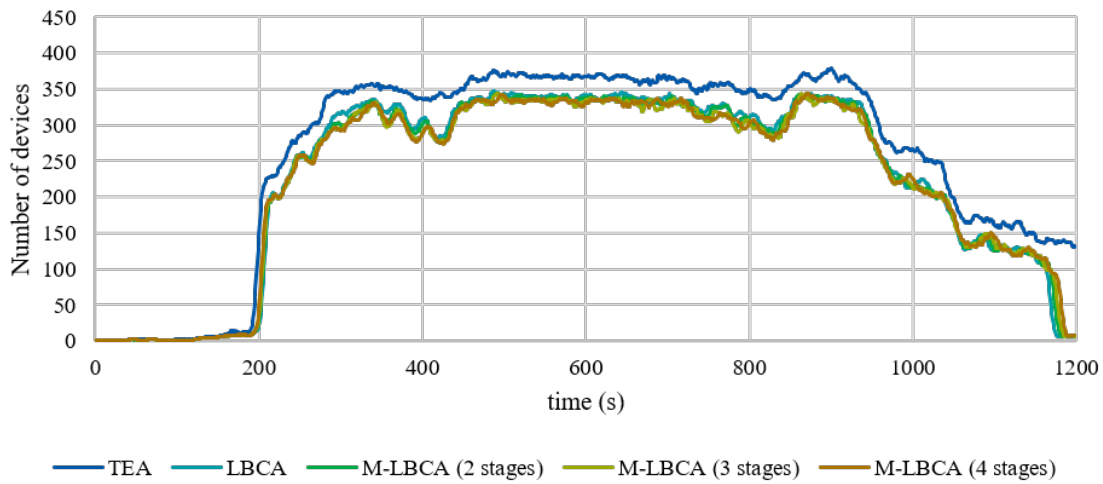


Figure 21: Number of concurrent BS connections in Evaluation 2

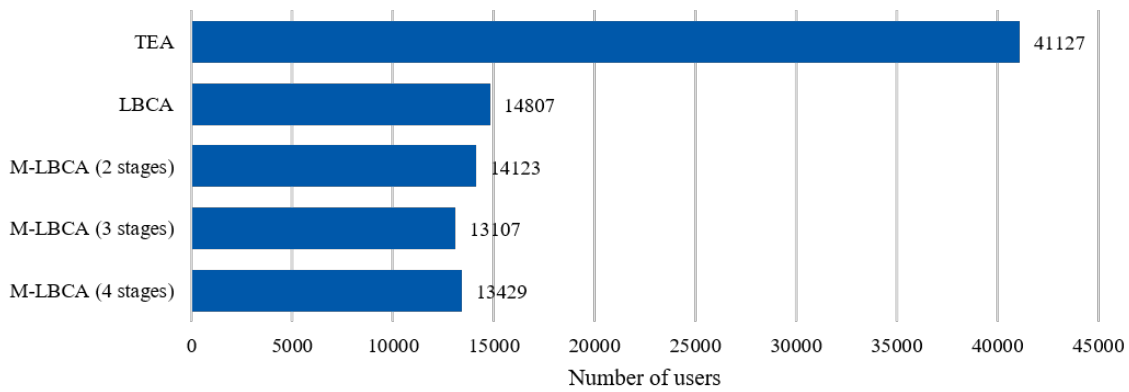


Figure 22: Number of “failed” users in Evaluation 2.

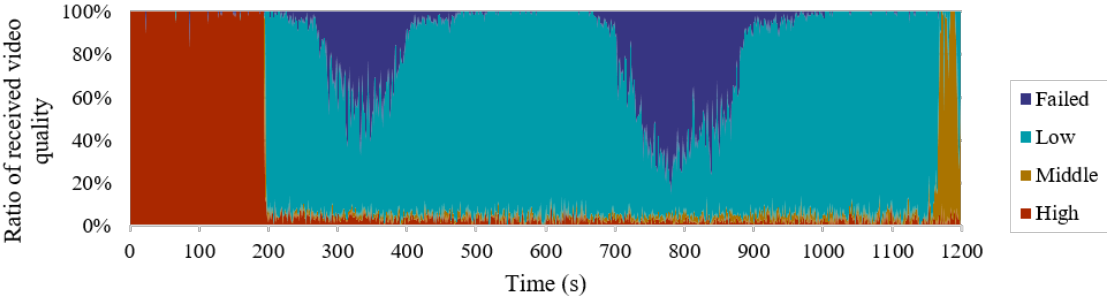


Figure 23: User's QoE with TEA in Evaluation 3

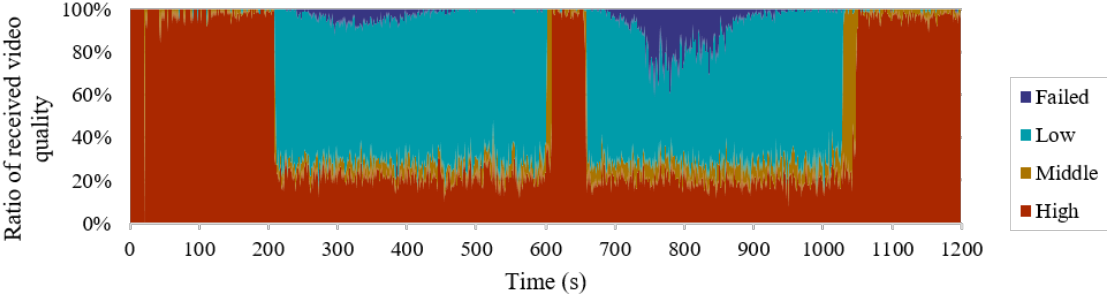


Figure 24: User's QoE with LBCA in Evaluation 3

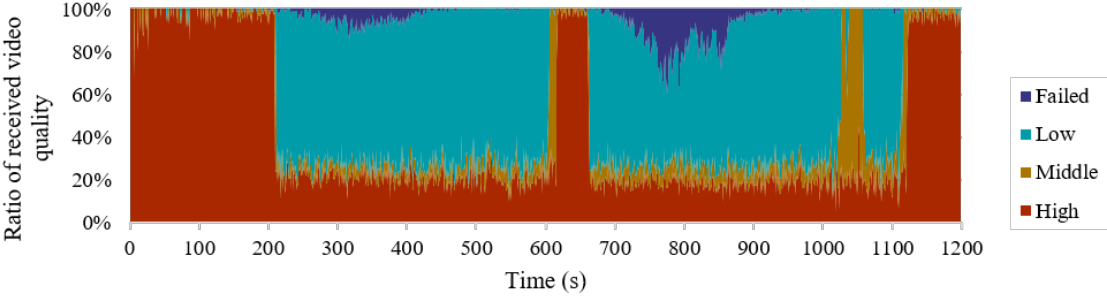


Figure 25: User's QoE with the two-stage M-LBCA in Evaluation 3

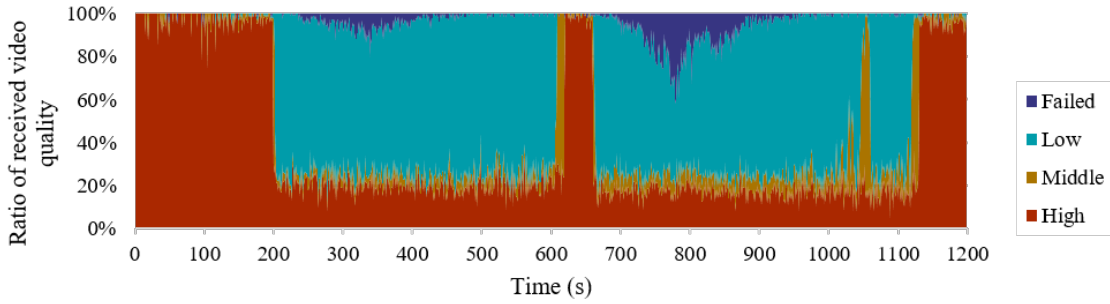


Figure 26: User's QoE with the three-stage M-LBCA in Evaluation 3

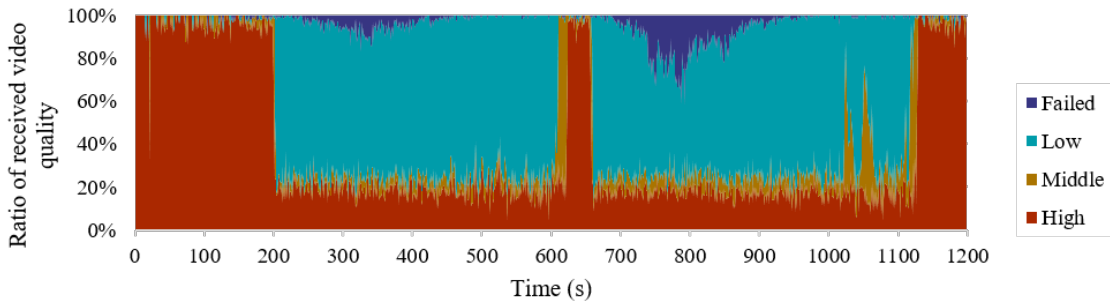


Figure 27: User's QoE with the four-stage M-LBCA in Evaluation 3

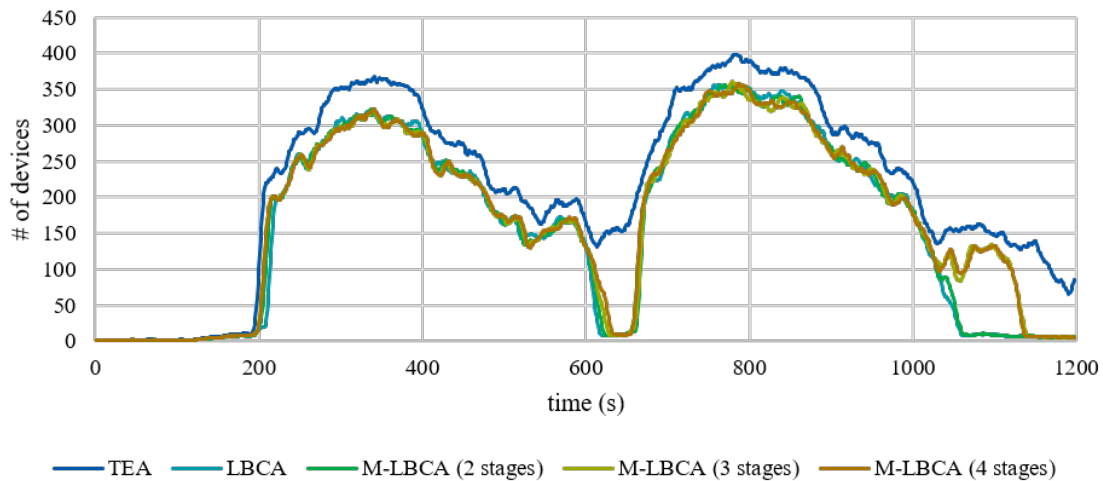


Figure 28: Number of concurrent BS connections in Evaluation 3

number of concurrent BS connections decreased for all the methods except TEA at approximately $t = 600$. This is caused by the efficient cache allocation of LBCA and M-LBCA, which improves the cache hit ratio via D2D communication and minimizes the number of requests to the BS.

Fig. 29 shows the total number of “failed” users in Evaluation 3. Fig. 29 reveals a decrease of

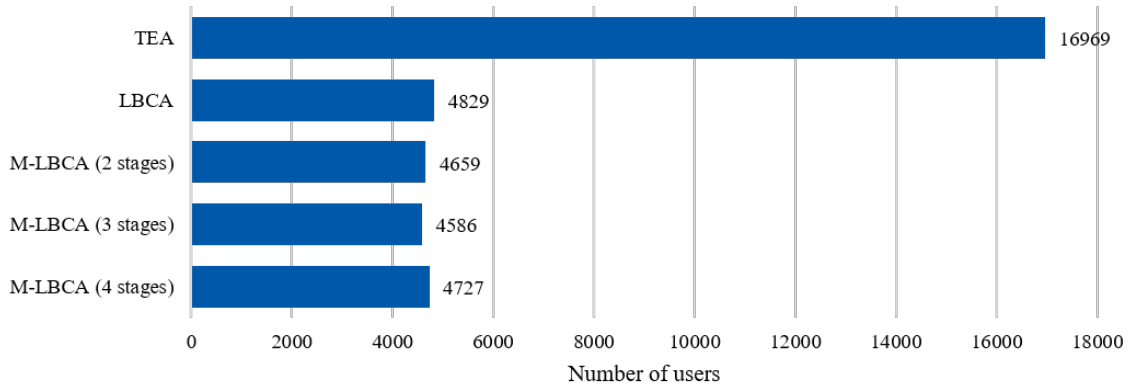


Figure 29: Number of “failed” users in Evaluation 3.

71.5%, 72.5%, 73.0%, and 72.1% in LBCA, two-stage M-LBCA, three-stage M-LBCA, and four-stage M-LBCA, respectively, compared with the TEA in terms of failed users among the M-LBCAs, the most significant user QoE improvement was observed in the three-stage M-LBCA. Extending LBCA to M-LBCA reduced the number of “failed” users by at most 5.0%.

5.2.4 Evaluation 4

Figs. 30–34 show the ratio of the video quality that users could view during Evaluation 4. The numbers of concurrent BS connections are shown in Fig. 35.

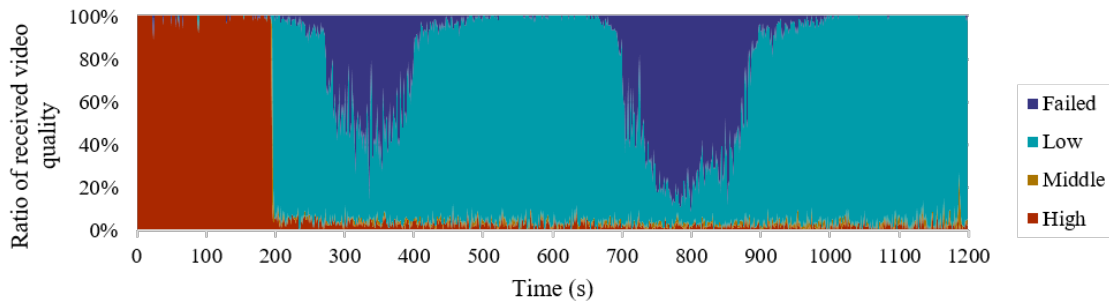


Figure 30: User’s QoE with TEA in Evaluation 4

For all the methods, the results indicated a significant decrease in the number of users who could view high-quality images from approximately $t = 200$. Fig. 35 shows that the number of concurrent BS connections rapidly increases at $t = 200$.

Fig. 36 shows the total number of “failed” users in Evaluation 4. Fig. 29 shows a 71.5%, 72.5%, 73.0%, and 72.1% decrease in LBCA, two-stage M-LBCA, three-stage M-LBCA, and four-stage M-LBCA’s failed users, respectively, compared to TEA. Extending LBCA to M-LBCA reduced the number of “failed” users by up to 7.7%.

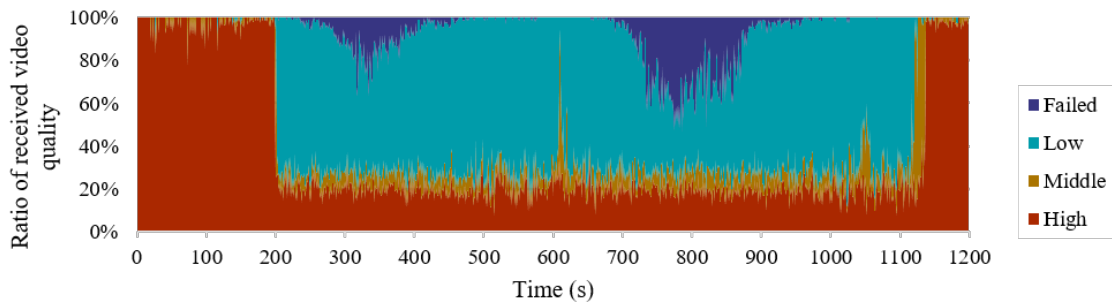


Figure 31: User's QoE with LBCA in Evaluation 4

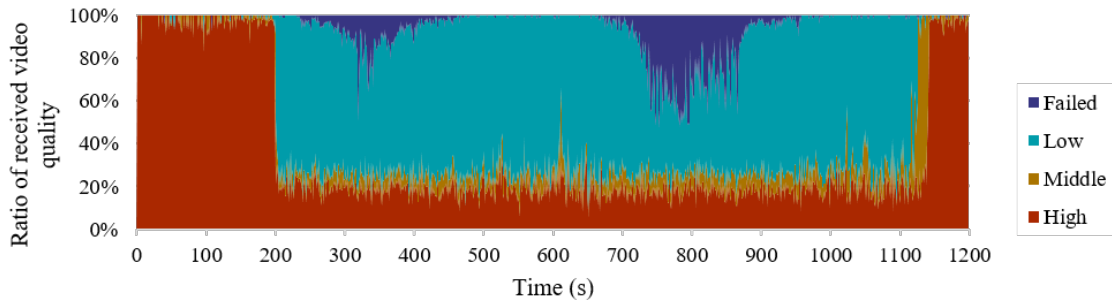


Figure 32: User's QoE with the two-stage M-LBCA in Evaluation 4

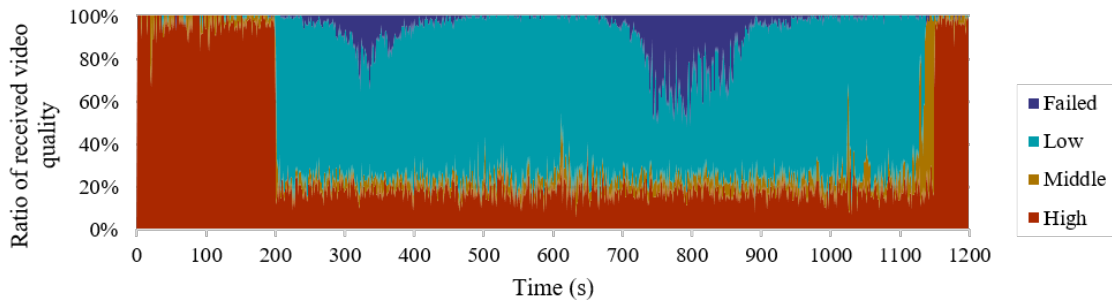


Figure 33: User's QoE with the three-stage M-LBCA in Evaluation 4

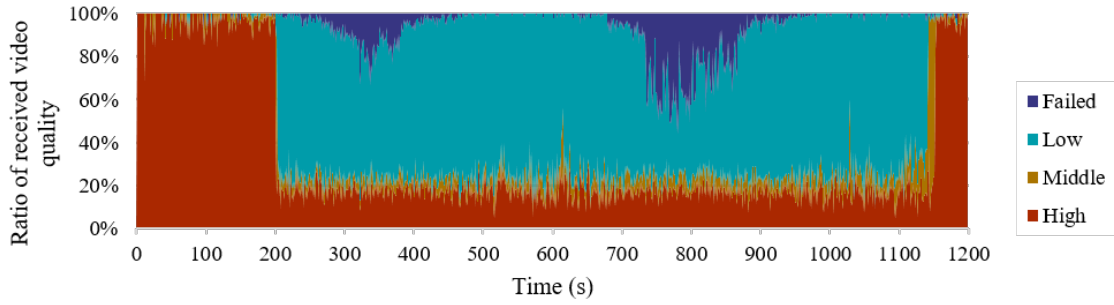


Figure 34: User’s QoE with the four-stage M-LBCA in Evaluation 4

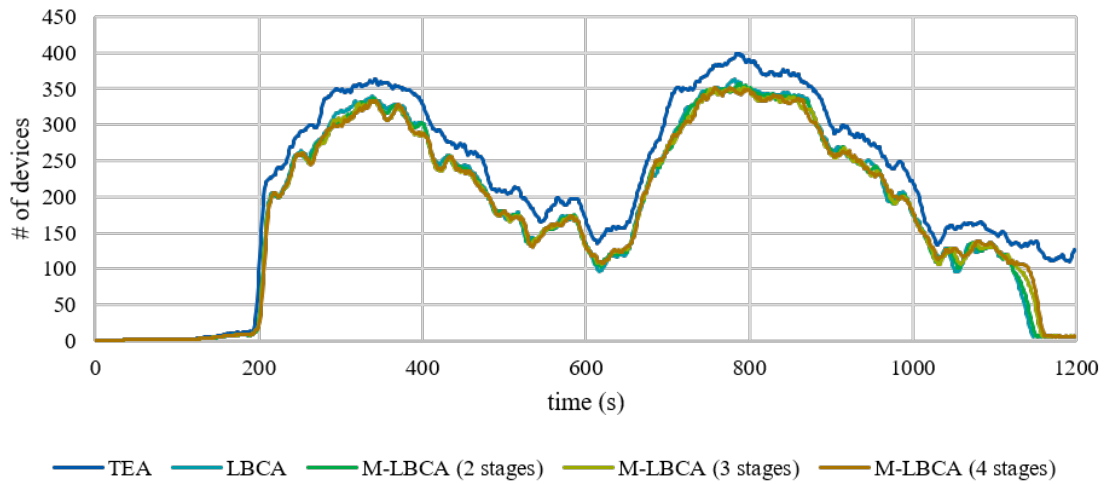


Figure 35: Number of concurrent BS connections in Evaluation 4

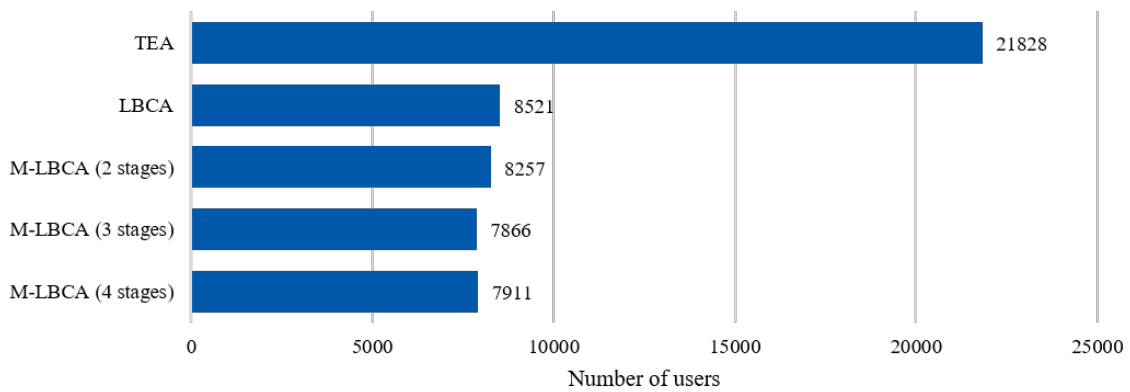


Figure 36: Number of “failed” users in Evaluation 4.

5.3 Comparison of methods

The results indicate that LBCA and M-LBCA can improve the QoE of users in various environments. Compared to the TEA method, the total number of users who temporarily failed to view content was reduced by at most 76.8% for LBCA and 77.9% when extended to M-LBCA.

The effect of extending LBCA to M-LBCA on the user's QoL was significant in Evaluations 2 and 4, where the devices were not moving compared with Evaluations 1 and 3, where the devices were moving. In LBCA, all devices cache each SVC level at the same ratio. Therefore, if the BS continues to be congested, devices far from the BS may not receive the SVC level being cached indefinitely because they cannot ensure communication speed. In an environment in which devices move, there is a possibility of receiving chunks to be cached in the future. However, in a nonmobile-device environment, the user's QoL improvement effect of LBCA was smaller, because there were fewer opportunities to receive caches. By extending this to the M-LBCA, different cache ratios were assigned to the devices close to or far from the BS, thereby allowing for efficient cache placement, even in environments where the devices do not move.

In this study, the largest user's QoL improvement was observed in the four-stage M-LBCA. The user's QoL improvement effect was reduced when using the four-stage M-LBCA. This suggests that the optimal number of stages depends on the environment.

6 Discussion

In this study, we simulated numerous content data requests. For a few requests, the effect of LBCA may be negligible because high-quality content can be received even without LBCA. To achieve a more dynamic change in the number of requests, the update interval of the color tag should be further shortened.

The results indicate that M-LBCA is superior to LBCA in reducing the number of "failed" users. It is necessary to identify a scenario in which the effect of the M-LBCA is maximized compared with that of LBCA. In addition, the computational cost of extending to M-LBCA should also be considered.

Although LBCA and M-LBCA reduced the total number of users who failed to view the requested content continuously, users remain. Because effective cache capacity has an upper limit, caching all the content is infeasible, although LBCA dynamically manages to expand the cache capacity of the data of the base layer. Therefore, cache misses occur when a device requests rare content. This capacity miss on the mobile cooperative cache increases the load of BS, resulting in insufficient bandwidth between the device and BS. The load on the BS in the access concentration can be reduced using multihop D2D communication, which enlarges the D2D communication range. An alternative method is to integrate the licensed and unlicensed spectrums to improve the performance of D2D communication [18].

7 Conclusion and Future Work

In this study, we proposed an LBCA scheme for a mobile cooperative cache. The LBCA scheme extends the colored cache to efficiently treat SVC video chunks in the mobile cooperative cache while carefully considering the radio attenuation characteristics of the wireless networks. Furthermore, we proposed an M-LBCA that divides the distribution area and constructed a multi-stage cache network. Simulations demonstrated that LBCA and M-LBCA reduced the total number of users who temporarily failed to view the content compared to TEA, the traditional method, in the four evaluation patterns. The results indicate that LBCA and M-LBCA improve the user's QoE.

In future work, we aim to evaluate LBCA under practical conditions and compared it with other related studies, such as [10]. We aim to also evaluate the QoE of users from various perspectives.

Acknowledgment

This study was supported by JSPS KAKENHI, Grant Numbers JP20H00592, and JP21K11805. In addition, we were partially supported by TIS Inc. in a collaborative research project for D2D cache network.

References

- [1] Cisco. Cisco annual internet report (2018–2023). <https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.pdf>, Accessed 2022-06-20.
- [2] H.T. Friis. A note on a simple transmission formula. *Proceedings of the IRE*, 34(5):254–256, 1946.
- [3] Dachun Huang, Xiaoxiang Wang, and Dongyu Wang. Caching scheme based on user clustering and user requests prediction in small cells. In *2017 IEEE 17th International Conference on Communication Technology (ICCT)*, pages 970–974, 2017.
- [4] Yusaku Inoue, Yuta Oguma, Satoshi Tarouda, Tetsuya Inagaki, Young-Cheol Yu, and Naoya Sato. Mobile terminals for 5g communications. *NTT DOCOMO Technical Journal*, 22(3), 2021.
- [5] Komal S. Khan, Adeena Naeem, and Abbas Jamalipour. Incentive-based caching and communication in a clustered d2d network. *IEEE Internet of Things Journal*, 9(5):3313–3320, 2022.
- [6] Z. Li and G. Simon. In a Telco-CDN, Pushing Content Makes Sense. *IEEE Transactions on Network and Service Management*, 10(3):300–311, September 2013.
- [7] Junchao Ma, Lingjia Liu, Hao Song, Rubayet Shafin, Bodong Shang, and Pingzhi Fan. Scalable video transmission in cache-aided device-to-device networks. *IEEE Transactions on Wireless Communications*, 19(6):4247–4261, 2020.
- [8] Takuma Nakajima, Masato Yoshimi, Celimuge Wu, and Tsutomu Yoshinaga. Color-Based Cooperative Cache and Its Routing Scheme for Telco-CDNs. *IEICE Transactions on Information and Systems*, E100.D(12):2847–2856, December 2017.
- [9] Hiroki Okada, Takayuki Shiroma, Takuma Nakajima, Celimuge WU, and Tsutomu Yoshinaga. A color-based cooperative cache with chunking contents distribution. volume 117, pages 3–8, November 2017.
- [10] Hiroki Okada, Masato Yoshimi, Celimuge Wu, and Tsutomu Yoshinaga. A failsoft scheme for mobile live streaming by scalable video coding. *IEICE Transactions on Information and Systems*, E104.D(12):2121–2130, 2021.
- [11] Qualcomm. Global update on 5G spectrum. <https://www.qualcomm.com/content/dam/qcom-mm-martech/dm-assets/documents/spectrum-for-4g-and-5g.pdf>, 2019.
- [12] T. Shiroma, C. Wu, and T. Yoshinaga. A template-based sub-optimal content distribution for d2d content sharing networks. In *2018 Sixth International Symposium on Computing and Networking (CANDAR)*, pages 167–173, 2018.
- [13] Takayuki Shiroma, Takuma Nakajima, Masato Yoshimi, Hidetsugu Irie, and Tsutomu Yoshinaga. An implementation of web cache system using access frequency of content pieces. In *IEICE Tech. Rep.*, volume 114, pages 35–40, November 2014.
- [14] M. Vilas, X. G. Paneda, R. Garcia, D. Melendi, and V. G. Garcia. User behavior analysis of a video-on-demand service with a wide variety of subjects and lengths. In *31st EUROMICRO Conference on Software Engineering and Advanced Applications*, pages 330–337, 2005.

- [15] S. Wenger, Y. Wang, T. Schierl, and A. Eleftheriadis. RTP Payload Format for Scalable Video Coding. RFC 6190, May 2011.
- [16] Hongliang Yu, Dongdong Zheng, Ben Y Zhao, and Weimin Zheng. Understanding user behavior in large-scale video-on-demand systems. *ACM SIGOPS Operating Systems Review*, 40(4):333–344, 2006.
- [17] Cheng Zhan and Guo Yao. Svc-based caching and transmission strategy in wireless device-to-device networks. In *2018 16th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt)*, pages 1–6, 2018.
- [18] Hongliang Zhang, Yun Liao, and Lingyang Song. D2D-U: Device-to-device communications in unlicensed bands for 5G system. *IEEE Transactions on Wireless Communications*, 16(6):3507–3519, 2017.